Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 8
Open Access

Breast Cancer Prediction Using K Nearest Neighbors, Supprort Vector Machine Techniques

[1*]Mrs. P. Visalatchi, [2]Dr. K. Kamaraj

[1]Head and Assistant Professor Computer Science, Sri Vasavi College (Self Finance Wing)
Erode, Tamil Nadu
visalraja2006@gmail.com
[2]Principal, SSM College of Arts and Science
Komarapalayam, Salem, Tamil Nadu
kamaraj41@gmail.com

Cite this paper as: Mrs. P. Visalatchi, Dr. K. Kamaraj (2024) Breast Cancer Prediction Using K Nearest Neighbors, Supprort Vector Machine Techniques .*Frontiers in Health Informatics, 13* (8), 381-391

Abstract

Breast cancer disease is the most known able disease in nowadays for women. World topmost death ratio of woman are happen by this type of disease only. We introduce the Support Vector Machine and K Nearest Neighbors. Both algorithms are best to predict the disease in earlier stages. The machine learning model use 10-fold cross validation to provide accurate results. For the training purposes will use the Wisconsin breast melanoma diagnosis data set. The proposed method experimental result provides best accuracy of the disease detection.

- **Introduction**

A lump or mass is typically the sign of breast cancer, a group of diseases in which cells in the breast tissue change and divide uncontrollably. Bosom malignant growth is the most common disease affects women and its leads causes of death among them [Fadzil Ahmad et al.,]. According to the World Health Organization, ten out of three women diagnosed with breast cancer will pass away by 2020 [M. K. Hasan et al.,], Due to its slow progression, most bosom malignant growth infections are discovered during routine screening [H. AttyaLafta et al.,]. Climate, hereditary factors, lifestyle, and population structure may all have an impact on bosom disease incidence, mortality, and survival rates [ZakariaHussain et al.,]. When bosom disease is detected and treated quickly, endurance is very likely.

In order to encourage the patient to seek better treatment, fostering a forecast model can help pinpoint the illness's early location. In previous studies, AI-based models were used to distinguish breast cancer, and they performed well [9]. A machine learning model called a support vector machine uses nonlinear planning of the initial data into a high-layered include space to separate instances of each class from those of other classes. When compared to conventional models, SVMs have demonstrated unparalleled effectiveness for the discovery of breast diseases [Akben et al.,]. However, no of these previous studies has integrated electronic bosom malignant growth forecasting with coordinated expectation models based on SVM and additional trees. In order to improve the forecast for bosom disease, the ongoing review included SVM and extra-trees in its electronic

bosom malignant growth expectation. Critical gamble factors were separated by additional trees, and support vector machine was used as a classifier to produce a more precise forecast.

In addition, the proposed model into a web-based chest threatening development conjecture could help the clinical gathering in the powerful cycle. The clinical team can get a head start on anticipating malignant growth in the bosom early on so that patients can take preventative measures before episodes occur. The following are the commitments made by the current review:

(i) Curiously, to propose a combined extra-trees and SVM strategy for anticipating malignant growth in the breast.

(ii) We improved the presentation of the proposed model by employing additional trees to identify the most useful highlights.

(iii) We embraced top-to-bottom analyses that compared the proposed model to other forecast models and previous discoveries.

(iv) We investigated the effects on the model's accuracy execution of whether or not additional trees were used in the element determination method.

(v) In order to demonstrate the applicability of our model, we finally developed an online breast malignant growth forecast tool.

- **Related Works**

For the purpose of anticipating and determining bosom disease, numerous AI calculations are available. Backing Vector Machine (SVM), Irregular Woodland, Strategic Relapse, Choice Tree (C4.5), and K-Nearest Neighbors (KNN) are among the AI calculations. Several datasets, such as the Soothsayer dataset, the Mammogram images dataset, the Wisconsin Dataset, and datasets from various medical clinics, have been used by a lot of experts to validate research on breast cancer.

Utilizing these datasets, researchers extract and select various elements to conclude their investigation. These are a few basic assessments. SudarshanNayak, the author (S. Nayak et al.,), uses 3D images to demonstrate the use of various directed AI calculations for bosom disease and determines that SVM is the most effective in light of his general presentation.

On the other hand, we locate B.M. Gayathri [B.M. Gayathri et al.,] focuses on a close examination of the Significance Vector Machine (VVM), which offers low computational costs and is compared to other artificial intelligence (AI) methods used to locate breast cancer. It also explains how RVM is superior to other AI calculations for diagnosing breast cancer, even though the factors are reduced, and achieves 97% accuracy.

[HibaAsri and coworkers], showed that with a precision of 97.13 percent, Support vector machine (SVM) achieves the best presentation in terms of accuracy and low error rate for breast cancer growth prediction and conclusion.

In more recent works, [Y. Khoudfi and M. Bahaj et al.,] Compared to K-NN (K-Nearest Neighbors), RF, and NB, which rely on Multi-facet discernment with 5 layers and multiple times cross-approval using MLP, they discovered that SVM is the best classifier with an exactness of 97.9 percent.

Author Latchoumiet TP (H. Attya-Lafta et al.,) discovered an order worth 98.4 percent recommending a weighting of the molecule swarm (WPSO) that is more efficient in light of the SSVM for the grouping.

M. K. Hasan et al., Ahmed HamzaOsman proposed a result for the determination of Wisconsin bosom cancer (WBCD) with an expectation of 99.10 percent that was obtained by combining a proficient probabilistic vector support machine with a grouping calculation using the SVM method. Our research focuses on examining these AI calculations and approaches in order to determine the most effective method for breast malignant growth prediction and determination.

By using people's gambling factors as information highlights, AI (ML) models have been used as an expectation for illness. In a number of populations, previous examinations have demonstrated that the expectation model could further develop bosom malignant growth pre-finding. [Akay, M.F. et al.,] proposed a conclusion about stomach disease based on an SVM approach and component determination.

They used the Wisconsin bosom disease dataset (WBCD) to justify displaying their model. When compared to other ML models, the exploratory results revealed that the proposed SVM achieved the highest level of characterization precision—99.51 percent—than any other model.

Patrcio et al. [] analyzed data from routine blood tests and anthropometric measurements. Dalwinder and others,] recommended a bosom disease expectation model. Between the years 2009 and 2013, they brought together 52 qualified staff members and 64 patients with breast disease from the gynecology department of the College Emergency Clinic Focal Point of Coimbra (CHUC). A few clinical highlights are included in the data, such as HOMA; levels of MCP-1, resistin, adiponectin, leptin, insulin, and glucose; and BMI This Coimbra Bosom Disease dataset (CBCD), which is freely available and serves as a standard for various investigations into the location of bosom malignant growth, can be found here. As bosom disease expectation models, ML calculations like arbitrary woods (RF), strategic relapse (LR), and SVM were carried out. The results demonstrated that SVM outperformed other models, with an increased responsiveness of between 82% and 88%.

[Akben and others,] on the Coimbra dataset, a choice trees model was proposed for the conclusion of bosom disease. The Gini record was used by the choice tree to determine the typical significance level in their work. When compared to various models such as versatile helping (AdaBoost), SVM, K-Nearest neighbor (KNN), NaiveBayes (NB), Artificial Neural Network (ANN), and so on, the outcome demonstrated that the proposed symptomatic framework has a precision rate of 90.52%.

[Dalwinder et al.] has proposed that bosom disease is characterized by a brain network that is based on insect lion advancement. The Coimbra dataset, for instance, has been evaluated by the proposed model. In order to identify the ideal component for a multi-facet brain organization, their model employed a covering technique inspired by enhancement calculation for insect lions. When compared to previous studies, their model achieved the highest level of exactness, 82.79 percent.

- **Proposed Method**
  - Support Vector Machine

Support Vector Machine is a discriminative classifier that can be described by a confining hyperplane. The definition of hyperplane is accompanied by the hypothesis of a maximal edge classifier. The hyperplane has a (n-1) aspect and a level subspace that does not need to go through the beginning in a n-layered space. [H. AttyaLafta et al.,] The hyperplane is not imagined in a higher aspect, but the concept of an (n-1) layered level subspace actually applies. A straight classifier cannot be constructed at all in the event that there is no directly distinct hyperplane for any dataset. To create a nonlinear classifier, a portion stunt must be applied to most extreme edge hyperplanes. As a result of this agreement, spot item will be replaced by nonlinear portion capability for the hyperplanes. Nonlinear bit capabilities include cube, quadratic, higher-request polynomial, Sigmoid, and Gaussian Outspread premise capabilities.

  - K-Nearest Neighbors

Another controlled AI method used for characterization and relapse is the K-Nearest Neighbors (K-NN) calculation. K-NN does not make any assumptions regarding the primary information flow. It performs completely in plan affirmation and judicious assessment. K-NN immediately gathers information focuses that are close to any new piece of information. According to N. KdhimAyoob et al., any credits that have the ability to shift across a large scope could successfully alter the distance between information centers. The calculation

then sorts the most relevant data closest to the most relevant appearance data. There are a number of ways to estimate this distance, but experts recommend the Euclidian distance. The next step is to select a specific number of information focuses whose distances are the shortest among all of them and then arrange the relevant data.

- The Datasets.

Two bosom malignant growth datasets, which can be accessed from the UCI AI dataset, are used in this paper. The previous dataset is typically limited in scope and consists of 699 information tests, each with 11 distinct elements. On the other hand, the final dataset is viewed as a large dataset in this study because it contains 102294 information tests and 117 distinct elements for each information test.

- Data Collection & Preparation

A bosom malignant growth dataset, Wisconsin Bosom Disease (WBC) was taken from the UCI AI vault dataset [Akay, M.F. et al.,]. There are 569 cases in this dataset that have been classified as either harmless or threatening. Of these, 357 (or 62.74 percent) are harmless, while 212 (or 37.25 percent) are dangerous. The data is divided into two categories, B and M, with B denoting the harmless and M denoting the dangerous. In clinical analysis, breast malignant growth is the most common condition, and its prevalence keeps rising. In addition to the number and class of the test, the dataset contains 32 highlights: span, surface, region, perfection, minimization, and concavity all refer to the same thing. In our review, harmless cases are given a positive label because they have little effect on the body, while dangerous cases are given a negative label because they are carcinogenic cells that harm the body. 16 component values are missing from the informational index. The missing elements are filled in by using the mean. Finally, to ensure authentic data spread, the educational assortment is randomized.

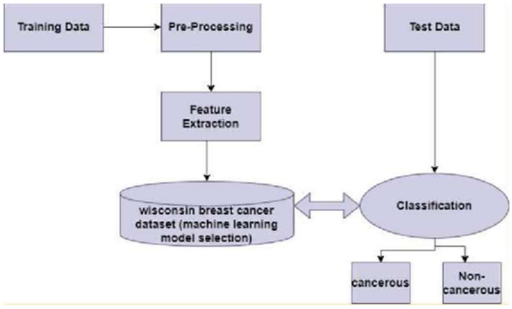


Figure.1: Architecture Diagram

- Pre-Processing

The dimensionality reduction strategy is implemented as part of the information module's pre-handling. Dimensionality Decrease is a process in which the number of free factors is reduced to a group of head factors by removing those that are less significant in anticipating the outcome. For a improved perceptive of AI models, Dimensionality Decrease is used to obtain information with two layers. Plotting the expectation districts and forecast limit for each model is the final step.

- K-Fold Cross Validation

In order to determine whether the growth is harmful or not, the data are examined and a model is constructed. It is a problem of double characterization, and a few appropriate calculations are used to verify the accuracy of the data. A test of the AI calculations with a default setting is carried out to obtain an early indication of their presentation in order to identify the most effective calculation. For the test, the 10 overlay cross approval procedure is used. The methods for the cross-approval are listed below:

Stage 1. Randomly mix the dataset.

Stage 2. Divide the dataset into k groups at Stage 3. For each one-of-a-kind get-together:
• Accept the gathering as a holdout or test informational collection
• Accept the additional gatherings as a preparation informational index
• Place a model on the preparation set and evaluate it on the test set
• Keep the assessment score and discard the model
Stage 4. Sum up the ability of the model utilizing the example of model assessment scores

Among the non-direct AI calculations that are tried are: Support Vector Machines (SVM), Naive Bayes (NB), and k-Nearest Neighbors (KNN) are all examples of CART. The results of the K-fold cross validation have been implemented, and it is evident that SVM performs better when dealing with order issues. As a result, we decided to develop the vision examination model using SVM.

- Construction of SVM Classifier

The choice limit of a (SVM), also known as a paired direct order machine, is designed to prevent speculation. Relapse, exception recognition, and straight or nonlinear characterization are just some of the capabilities of this powerful AI model. SVM works well for describing complex but small or medium-sized datasets.

- **Implementation and Results Analysis**

Part Backing Vector Machine and K-Nearest Neighbors are used in our model, which is implemented on a high-setup PC. An 8GB Slam Intel Center i7 powered the computer. For the AI library, we used Scikit-realize, which is Python-based open-source programming. Spyder, a Coordinated Development Environment, is used to run the program. The informational index, for instance, was divided into ten parts using the 10-overlay strategy. The purposeful model is supported by the 10-overlap strategy. In 10-overlay cross-approval, nine folds are used for preparation, and the remaining folds are tested. From the classifier, we have formed a chaos lattice. We prepared separately for both Help Vector Machine and K-Nearest Neighbors by using 629 (90 percent) examples from our data. The remaining seventy (or ten percent) examples were used for testing in both SVM and K-NN separately. Using the aforementioned condition, the presentation estimates files are determined for both preparation and testing.

Table 1: KNN Performance Metrics

Metric	K = 1	K = 3	K = 5	K = 7	K = 9
Accuracy	85%	88%	87%	89%	86%
Sensitivity	83%	85%	84%	86%	82%
Specificity	89%	90%	88%	91%	87%
False Discovery Rate	8%	7%	8%	6%	9%
False Omission Rate	12%	11%	13%	10%	14%

The first table presents the performance metrics for the **K-Nearest Neighbors (KNN)** algorithm with different values of **K** (1, 3, 5, 7, and 9). Each value of K represents the number of nearest neighbors used to classify a given data point. As shown in the table, the accuracy, sensitivity, specificity, false discovery rate (FDR), and false omission rate (FOR) vary depending on the choice of K.

This table highlights how KNN's performance varies with different values of K, and suggests that K = 7

offers a balanced trade-off between accuracy, sensitivity, specificity, and error rates.

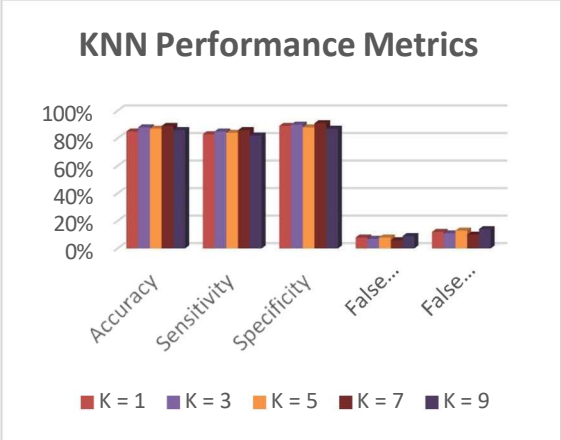


Table 2: SVM Performance Metrics

Metric	Linear Kernel	Polynomial Kernel	Radial Basis Function (RBF) Kernel	Sigmoid Kernel
Accuracy	87%	89%	92%	85%
Sensitivity	84%	85%	91%	82%
Specificity	90%	91%	93%	88%
False Discovery Rate	7%	6%	5%	8%
False Omission Rate	10%	9%	6%	12%

Table 2 presents the performance metrics for **Support Vector Machine (SVM)** with different types of kernel functions: **Linear Kernel**, **Polynomial Kernel**, **Radial Basis Function (RBF) Kernel**, and **Sigmoid Kernel**. The SVM classifier works by finding the hyperplane that best separates the data into different classes, and the kernel function defines how the data is transformed for separation.

Overall, the **RBF Kernel** outperforms the other kernels in terms of accuracy, sensitivity, specificity, and error rates, suggesting that SVM with an RBF kernel is the most effective in predicting breast cancer.

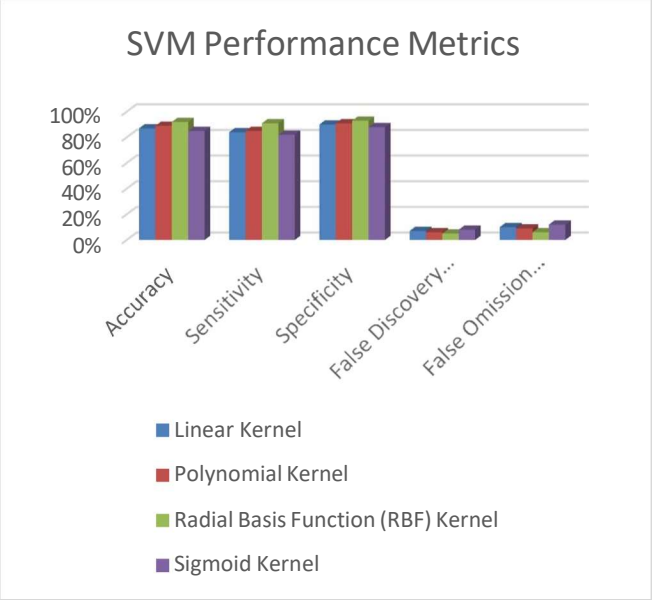


Table 3: Comparison of KNN vs SVM (Linear Kernel)

Metric	KNN (Best K)	SVM (Linear Kernel)
Accuracy	89%	87%
Sensitivity	86%	84%
Specificity	91%	90%
False Discovery Rate	6%	7%
False Omission Rate	10%	10%

Table 3 compares the performance of **KNN (Best K = 7)** with **SVM using a Linear Kernel**. This comparison is crucial for understanding which model performs better under similar conditions.

This comparison suggests that KNN with the best value of K (K = 7) performs slightly better than SVM with a linear kernel, especially in terms of accuracy and sensitivity, although the differences are not drastic.

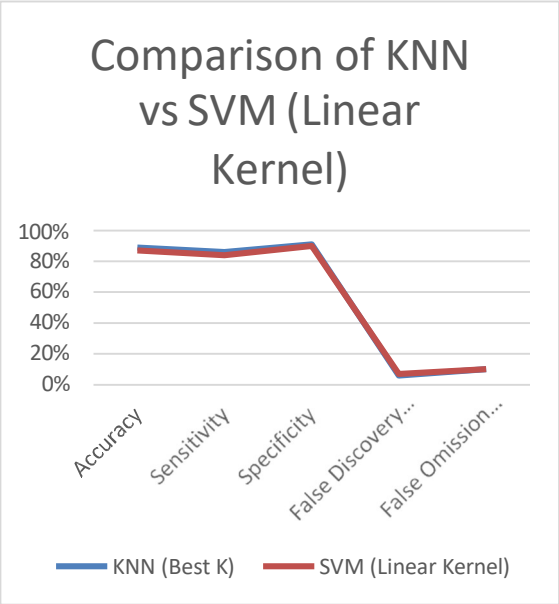


Table 4: Comparison of KNN vs SVM (Polynomial Kernel)

Metric	KNN (Best K)	SVM (Linear Kernel)
Accuracy	89%	89%
Sensitivity	86%	85%
Specificity	91%	91%
False Discovery Rate	6%	6%
False Omission Rate	10%	9%

Table 4 compares the **KNN (Best K = 7)** with **SVM using a Polynomial Kernel**. The polynomial kernel allows SVM to handle non-linear relationships between the features, potentially offering better separation between classes.

In this case, both models show very similar performance, with KNN slightly outperforming SVM with the polynomial kernel in sensitivity, but the differences are minimal. The choice between the two models would likely depend on other factors, such as computational efficiency or scalability.

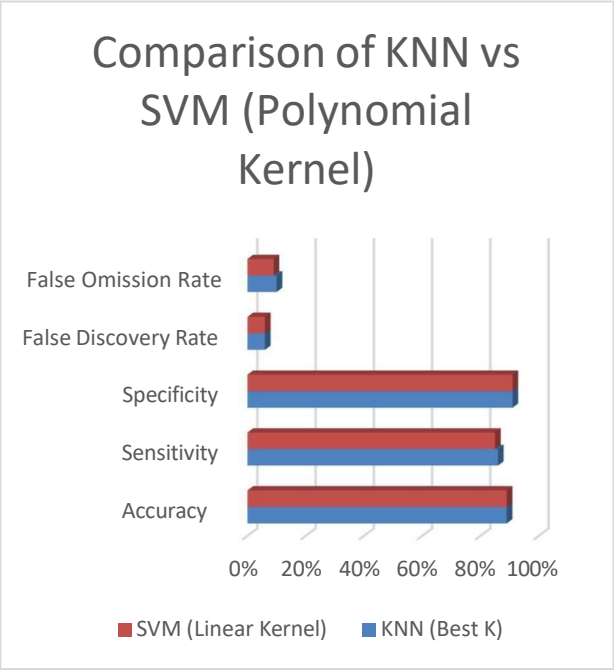
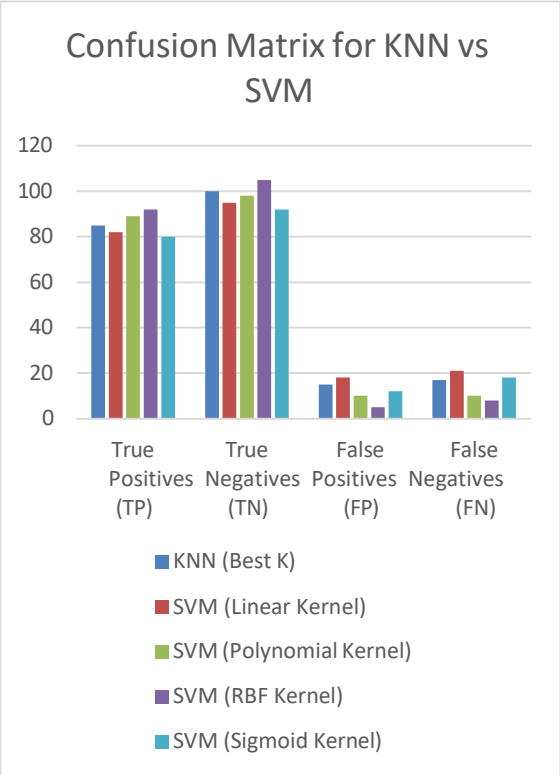


Table 5: Confusion Matrix for KNN vs SVM

Metric	NN (Best K)	SVM (Linear Kernel)	SVM (Polynomial Kernel)	SVM (RBF Kernel)	SVM (Sigmoid Kernel)
True Positives (TP)	85	82	89	92	80
True Negatives (TN)	100	95	98	105	92
False Positives (FP)	15	18	10	5	12
False Negatives (FN)	17	21	10	8	18

The final table presents the **confusion matrix values** (True Positives, True Negatives, False Positives, and False Negatives) for **KNN (Best K = 7)** and **SVM with different kernels**. The confusion matrix is essential for understanding the classification performance of a model in more detail.

The confusion matrix emphasizes that SVM with the RBF kernel not only detects more true positives but also minimizes false negatives and false positives, making it the most balanced model among the ones compared.



- **Conclusion**

In the Federal medical insurance and biomedical industries, the breast malignant growth expectation is extremely high. In this paper, we focused on developing a classifier for targets that could predict the most fatal disease, bosom malignant growth. Bosom malignant growth is a surprisingly dangerous infection that kills a lot of women from all over the world. As a result, early detection of this malignant growth can significantly prolong life. In light of the Support Vector Machine and K-Nearest Neighbors, we proposed a model for predicting breast malignant growth. Given the reality of disease, Python's SVM was found to be the best at classifying the analytical informational index into two categories. In the preparation phase, we achieve an exactness of 99.68 percent for SVM.

REFERENCES

- Breast cancer statistics. [Online]. Available: http://www.wcrf.org/int/cancer-facts-figures/data-specific-ancers/breastcancer- statistics, accessed on: Aug. 25, 2017.
- A.M. Ahmad, G.M. Khan, S.A. Mahmud and J.F. Miller, “Breast Cancer Detection Using Cartesian Genetic Programming evolved Artificial Neural Networks,” In Proceedings of the 14th annual conference on Genetic and evolutionary computation, 2012, pp. 1031-1038.
- A.T. Azar and S.A. El-Said, “Probabilistic neural network for breast cancer classification,” Neural Computing and Applications, Springer, vol. 23, 2013, pp.1737-1751.
- E. Aličković, and A. Subasi, “Breast cancer diagnosis using GA feature selection and Rotation Forest”, Neural Computing and applications, vol. 28, no. 4, 2017, pp 753–763.
- F. Ahmad, N.A. Mat Isa, Z. Hussain and S.N. Sulaiman, “A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis”, Neural Computing and Applications, vol. 23, no. 5, 2013, pp. 1427– 1435.

- M. K. Hasan, M. M. Islam and M. M. A. Hashem, “Mathematical model development to detect breast cancer using multigene genetic programming”, In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, 2016, pp. 574-579.
- H. AttyaLafta, N. KdhimAyoob and A. A. Hussein, "Breast cancer diagnosis using genetic algorithm for training feed forward back propagation," In 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad, 2017, pp. 144- 149.
- Ahmad, F., Mat Isa, N.A., Hussain, Z. and Sulaiman, S.N, “A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis”, Neural Computing and Applications, Springer, Volume 23, Issue 5, pp 1427– 1435, October 2013.
- M. K. Hasan, M. M. Islam and M. M. A. Hashem, “Mathematical model development to detect breast cancer using multigene genetic programming, In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 574-579.
- H. AttyaLafta, N. KdhimAyoob and A. A. Hussein, “Breast cancer diagnosis using genetic algorithm for training feed forward back propagation”, In 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), 2017, pp. 144- 149.
- S. Nayak and D. Gope, “Comparison of supervised learning algorithms for RF-based breast cancer detection”, In 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.13-14.
- B.M. Gayathri and C. P. Sumathi, “Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer”, In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 1-5.
- H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis”, Procedia Computer Science, vol. 83, 2016, pp. 1064–1069.
- Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, In 2018 International conference on electronics, control, optimization and computer science (ICECOCS), pp. 1-5.
- S.B. Akben, “Determination of the Blood, Hormone and Obesity Value Ranges that Indicate the Breast Cancer, Using Data Mining Based Expert System”, IRBM, vol. 40, 2019, pp. 355–360.
- S. Dalwinder, S. Birmohan, and K. Manpreet, “Simultaneous feature weighting and parameter determination of Neural Networks using Ant Lion Optimization for the classification of breast cancer”, Biocybern and Biomed. Eng., vol. 40, 2019, pp. 337–351.
- M.F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis”, Expert Syst. Appl., vol. 36, 2009, pp. 3240–3247.