Remittances Review
April 2024,
Volume: 9, No: 2, pp.4995-5004
ISSN: 2059-6588(Print) | ISSN 2059-6596(Online)

Received : 20 Januanry 2024, Accepted: 30 April 2024
DOI: https://doi.org/10.33282/rr.vx9i2.264

Breast Cancer Prediction Making use of Supervised Machine Learning Technique

[1]Maria Malik, [2]Saira Sharif, [3]Farhatul-ain, [4]Haseeb Ur Rehman,

Comsats University Islamabad, Pakistan[1,], University of Management and Technology, Pakistan[2,3,4]

Corresponding Author E.mail id: mariamalikkhan@gmail.com

Abstract:

According to recent data, breast cancer is the most common cancer in the world. Every year it kills almost 900,000 individuals; precise early identification can help minimize breast cancer mortality rates. This work offers a review that illustrates the novel applications of machine learning and deep learning technologies for detecting and classifying breast cancer and provides an overview of progress in this area. It first provides an overview of the many approaches to machine learning, then an overview of the different deep learning algorithms and specialized architectures for detecting and classifying breast cancer. This paper aims to investigate the performance of various algorithms such as Support Vector Machine (SVM), Logistic Regression, Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) in detecting the fatal disease. The proposed model's performance is evaluated using four metrics, i.e., accuracy, precision, recall, and F1-Score. The RNN outperformed the remaining algorithms in terms of accuracy (83%), precision (77%), and F1-Score (68%). However, ANN's recall (66%) was higher than SVM and logistic regression, CNN, RNN, and LSTM.

Keywords: ANN, RNN, SVM, CNN, LSTM, Logistic regression.

1. Introduction:

Over the past few decades, cancer has become more prevalent, with breast cancer being the most common type. It is estimated that around 290 million women are diagnosed with breast cancer every year (WHO). To improve decision-making related to cancer prediction, various machine learning techniques have been utilized to identify patterns in datasets and predict whether a cancer is malignant or benign. One method for detecting breast cancer is through mammography, which involves taking an X-ray image of the breast and extracting features such as cell shape, size homogeneity, and chromatin. Other factors that may be considered in predicting breast cancer include age, family history of breast cancer, previous biopsies, and various measures such as BMI, glucose, HOMA, and leptin. Additionally, machine learning can be used to predict and diagnose breast cancer using non-pathological data, such as obesity, ethnicity, pregnancy history, chemical or radiation exposure, nursing history, sedentary lifestyle, vitamin D deficiency, and irregular menstrual history.

Several researchers used ML techniques for the prediction of breast cancer. For instance, [1] proposed a linked rule base method to figure out the patterns from the breast cancer data. The linked rule has been applied to detect the link among similar features and eliminate the useless or independent features. The Wisconsin breast cancer dataset has been utilized for training and testing the presented technique. The proposed hybridized neural network technique outperforms all other neural networks in the study in terms of efficiency.

[2] employed neural networks to predict the survival rate of breast cancer patients' data. The data consists of 1373 patients, then compared the neural networks technique with the regression model. [3] proposed a linear diagnostic model for the recurring duration and non-recurring cases of disease to forecast malignant risk. The respective model was tested on a dataset of 569 patients by cross-validation approach, resulting in a 97.5% accuracy. [4] modified a model by adding Minimum Description Length (MDL) to the C4.5 decision tree method for diagnosing and predicting breast cancer, achieving an accuracy of 97.74%.

Moreover, big data can also be discussed in the literature. For instance, [5] employed a dataset of 2,00,000 patient records. The respective data is utilized on the C4.5 model with various other models (i.e. neural network and linear regression) for comparison. Thus, the C4.5 model yielded an accuracy of 93.6% and outperformed the other two.

Though much of the work is done on hybrid ML models. [6] introduced a model that combines a fuzzy system with a feature selection algorithm. In the respective model, only a critical feature has been employed for its training dataset of Wine Classification by using Wisconsin's Breast Cancer Classification technique. It was investigated that the model performs better if only relevant features are used rather than all. It was also demonstrated that PCA is best for feature selection and extraction methods to incorporate the model efficiency. [7] estimated the recurrence rate for breast cancer patients by utilizing several derivations of Decision Trees, Hybrid Decision Trees (HDT), and Fuzzy Decision Trees (FDT). The SEER dataset is used to train and tested on the proposed model. Results demonstrated that the Fuzzy Decision Tree model is more robust than other Decision Tree methods.

[8] presented a hybrid method for breast cancer classification by combining a traditional disease detection method with an advanced machine-learning algorithm. It yielded a 98.6% accuracy by acquiring the breast cancer dataset from the UCI repository. [9] presented a training-based method to obtain diagnostic information from the non-invasive procedure for selecting and classifying different texture features. The method achieved an accuracy of 90.7% by employing 128 cases, from which 67 were malignant, and 61 were benign, respectively.

[10] analyzed breast-conserving surgery for breast cancer by combining deep learning technology and ultrasound technology. They derived a deep LDL method by introducing two models, i.e. segmentation model ON and semiautomatic segmentation algorithm RA. A data

total of 102 cases were divided into 3 groups, i.e. W1, W2, and W3, each containing 34 cases based on ultrasound guidance of the DL segmentation model, ultrasound guidance, and palpitation guidance, respectively. The suggested study of the LDL model outperforms in terms of efficiency.

During the 1960s and until now, a significant attempt has been made on computerized medical images. Recently, the computer-aided diagnostic system has given rise to the medical field. These Computer-Aided Detections (CAD) method helps doctors detect and predict breast cancer. However, this automated detection software is not widely used for breast screening. [11] presented a CAD framework for the detection of tumor grades of breast cancer. Clinical data from 44 cases checked the validity of the proposed model. Similarly, [12] demonstrated an SVM classifier and reported an accuracy of 99.02% without utilizing any cross-validation method. [13] introduced Optimization Statistical Model (OSM) for the detection of breast cancer, yielded a 98.71% accuracy. Another innovative technique with a highly reported accuracy of 99.26% was proposed by [14]. This technique combined the Artificial Neural Network and the biological met plasticity property.

2. Proposed Methodology

The training models are created from the collected data using machine learning classification methods and immediately saved in the local server. The information gathered from patients is used for prediction, review analysis, decision-making, and data visualization. This system offers a cutting-edge application model that uses machine learning to improve various specialized health services in a significant way.

The overall functionality of the suggested model is depicted in Figure 1. To anticipate illness, doctors analyze the data they collect. The suggested model classifies the provided data using various machine-learning techniques to discriminate between healthy and ill patients. Machine learning classifiers can predict the disease in a timely and precise manner. Six different machine learning classifiers, including logistic regression, support vector machines (SVM), artificial neural networks (ANN), convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM), have been utilized. There are three steps to the project's execution of the recommended framework.

Data collection is the first stage, followed by pre-processing and computing, and finally, data transparency for clinicians. A range of sources, including the patient's home, the hospital, or the clinic, as well as remote data, are used to gather patient data.

During pre-processing, the collected data are evaluated and checked for errors. The data is transferred to the server for analysis after pre-processing. The data are computed and assessed using six machine learning algorithms: logistic regression, SVM, ANN, CNN, RNN, and LSTM. We split our data into training and testing datasets in the ratio of 80:20 to deploy the learning models.

2.1 Support Vector Machines

The supervised learning method is known as SVM. The statistical learning theory is the foundation of SVM. The SVM method is used for binary classification and multi-class problems. The SVM approach builds a hyperplane with support vectors and optimizes the distance between data points to produce large hyperplanes in a multi-dimensional space.

2.2 Linear Regression

A method of categorization known as logistic regression (34) allocates information to a predetermined particular class. It is a predictive analytic method that is based on probability. Logistic regression may be employed to categorize results based on various data types. It can swiftly pinpoint the variables that are most effective for categorization.

2.3 Artificial Neural Networks (ANN)

Artificial neural networks (ANN), based on feed-forward neural networks, are another widely used machine learning technique. The three layers of an ANN are input, hidden, and output. The input layer applies hidden processing to the input attributes of the Input layer to create output for the output layer. As soon as the intended outcome is attained, the output layer sends the output back to the hidden layer for additional processing. The alteration is carried out during the training process. The output layer reduces output error with the aid of the hidden layer. To offer timely surveillance to patients if breast cancer is discovered, the calculated result is transmitted to the doctor after it has been determined.

2.4 Convolutional neural network (CNN)

Specifically created to handle deep network configurations, CNN (37, 38) is a feedforward neural network. There are three layers in the CNN design: Convolution layer: A convolutional layer offers translation inversion. Since they work on every part of the tensor, convolution kernels search for the same feature across the entire sensor observation tensor. The thinner convolution layer extracts the edge features. On the other hand, the deeper convolution layers retrieve the potential features. Pooling layer: A pooling layer is applied after a convolution layer to downsample the feature maps produced by the convolution layer. A fully connected neural network layer is constructed similarly to a standard neural network. Each input unit is connected to each hidden or output layer neuron (that is, the outputs of the last pooling layer). This layer produces the categorization results.

2.5 Recurrent Neural Networks (RNN)

Multiple layers of the RNN's feedback loops allow it to transmit information from the past to the present. The loops of an RNN enable the information to retain. The RNN's hidden layers are a data storage system like computer memory. RNNs, a form of potent DNN, use loops and internal memory to process sequence data.

2.6 Long Short-Term Memory (LSTM)

A recurrent neural network that can learn long-term dependencies is the LSTM. One input layer, two hidden layers, and one output layer make up an LSTM network's standard four layers. The three gates in this system are the forget gate, input gate, and output gate.

2.7 Evaluation metrics

We used the following four evaluation measures to assess the performance of the six classifiers:

1. **Accuracy:** Accuracy is the proportion of correctly predicted sample points among all the sample points. It is assessed using the formula provided below:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100 \tag{1}$$

2. **Precision:** It's the ratio of genuine positive cases to all illness cases. It is assessed using the formula provided below:

$$Precision = \frac{TP}{(TP + FN)} \times 100 \tag{2}$$

3. **Recall:** It is the ratio of the number of true negative cases (i.e., cases in which the individual does not have the illness and is correctly identified as not having the illness) to the total number of illness cases. This can be calculated using the formula provided in equation 3:

$$Recall = \frac{TN}{(TN + FP)} \times 100 \tag{3}$$

4. **F$_1$-Score:** The F1-score is a metric that combines the precision and recall of a model, calculated using the harmonic mean. It is evaluated using the formula given in equation 4:

$$F_1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{4}$$

TP & TN represent the true positive and negative predictions the healthcare model made, respectively. FP and FN refer to the model's false positive and false negative predictions.

3. Results and Discussion:

This section focuses on the results of classification algorithms, including support vector machines (SVM), logistic regression, artificial neural networks (ANN), convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM). The Wisconsin Breast Cancer dataset from the UCI Machine Learning Repository is used in this study. The dataset includes 699 instances of breast cancer cases, with 458 instances being benign and 241 instances being malignant. There are two classes in the dataset, with 65.5% of the instances classified as malignant and 34.5% classified as benign. The dataset also includes 11 integer-valued attributes.

The accuracy of the six classifiers is as follows: The support vector machine (SVM) classifier achieved an accuracy of 80% on the breast cancer dataset. The logistic regression classifier achieved an accuracy of 78%. The artificial neural network (ANN) classifier achieved an accuracy of 67%. The convolutional neural network (CNN) attained an accuracy of 76%. The recurrent neural network (RNN) achieved an accuracy of 83%. The long short-term memory (LSTM) reached an accuracy of 79%. Among the evaluated classifiers, the RNN classifier achieved the highest accuracy. The accuracy of the classifiers is shown in Figure 2.

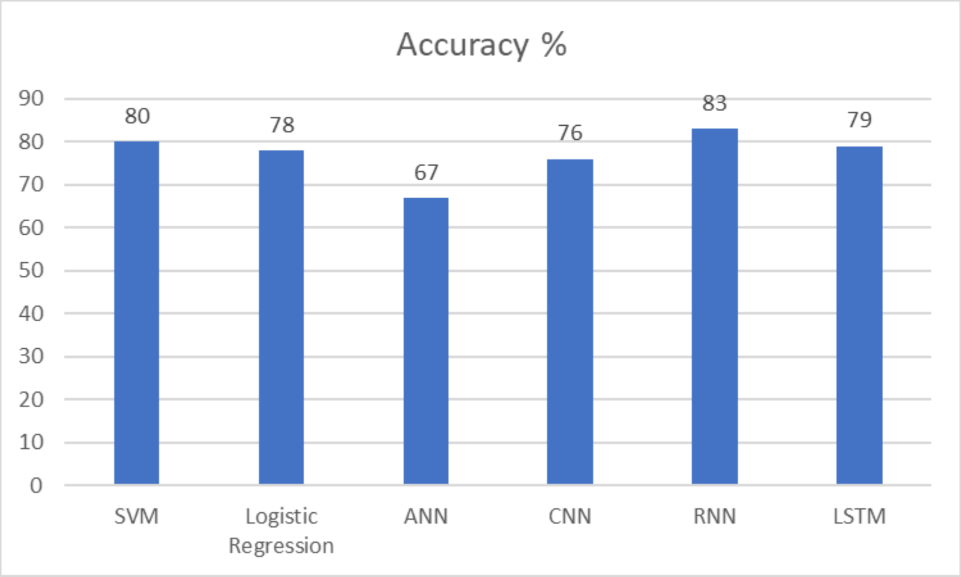

Figure. 2: Accuracy of the classifiers

The precision of the six classifiers is as follows: The support vector machine (SVM) classifier achieved a precision of 73% on the breast cancer dataset. The logistic regression classifier achieved a precision of 69%. The artificial neural network (ANN) classifier achieved a precision of 66%. The convolutional neural network (CNN) achieved a precision of 68%. The recurrent neural network (RNN) achieved a precision of 77%. The long short-term memory (LSTM) achieved a precision of 76%. Among the evaluated classifiers, the RNN classifier achieved the highest precision. The precision of the classifiers is shown in Figure 3.

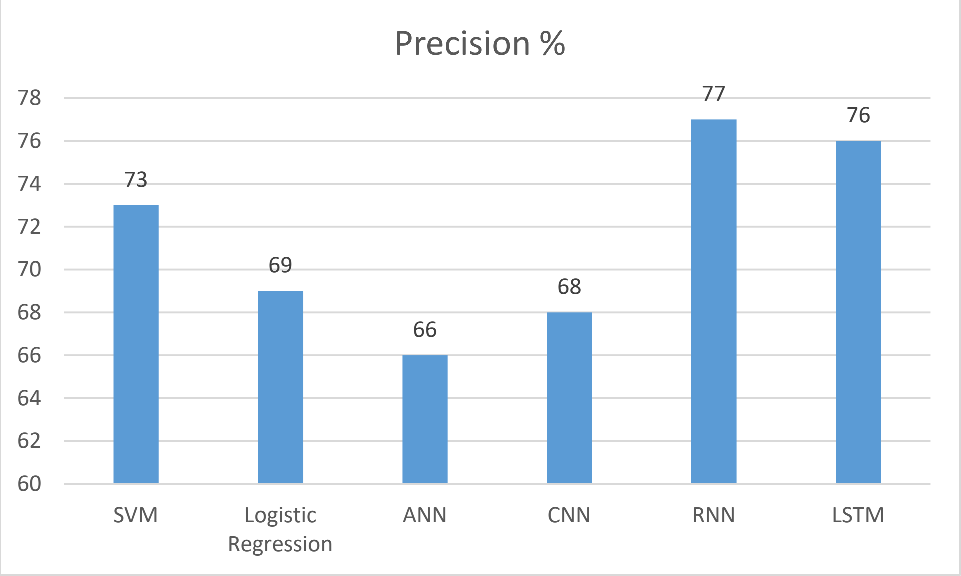

Figure.3: Precision of the Classifiers

The recall of the six classifiers is as follows: The support vector machine (SVM) classifier achieved a recall of 61% on the breast cancer dataset. The logistic regression classifier achieved a recall of 59%. The artificial neural network (ANN) classifier achieved a recall of 66%. The convolutional neural network (CNN) achieved a recall of 65%. The recurrent neural network (RNN) earned a recall of 59%. The long short-term memory (LSTM) achieved a recall of 63%. Among the evaluated classifiers, the ANN classifier achieved the highest recall. Figure 4 shows the recall achieved by all the classifiers.

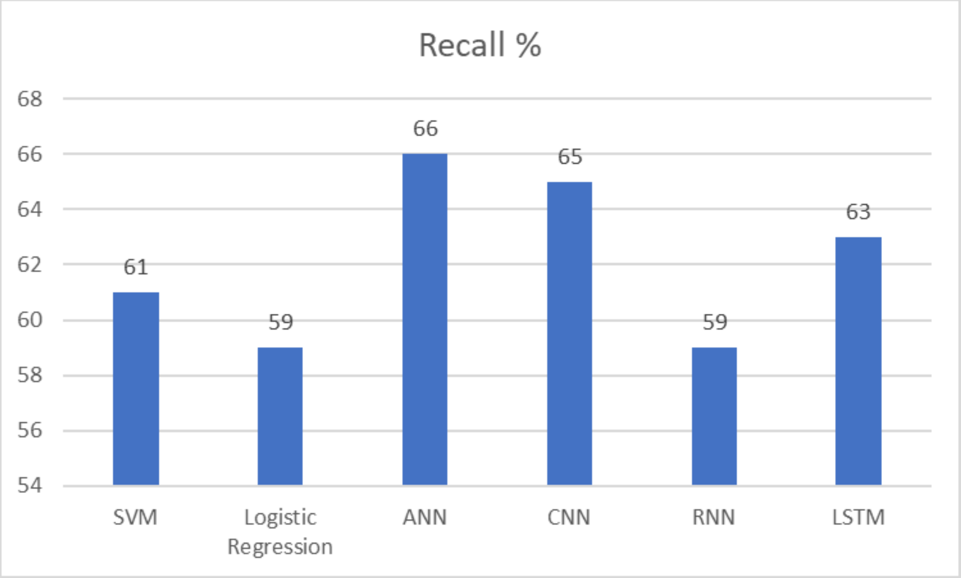

Figure.4: Recall of the Classifiers

The F1-Score of the six classifiers is as follows: the support vector machine (SVM) classifier achieved an F1-Score of 61% on the breast cancer dataset, the logistic regression classifier achieved an F1-Score of 63%, the artificial neural network (ANN) classifier achieved an F1-Score of 60%, the convolutional neural network (CNN) achieved an F1-Score of 64%, the recurrent neural network (RNN) achieved an F1-Score of 68%, and the long short-term memory (LSTM) achieved an F1-Score of 65%. Among the evaluated classifiers, the RNN classifier achieved the highest F1-Score. Figure 5 shows the F1-Score achieved by all of the classifiers.

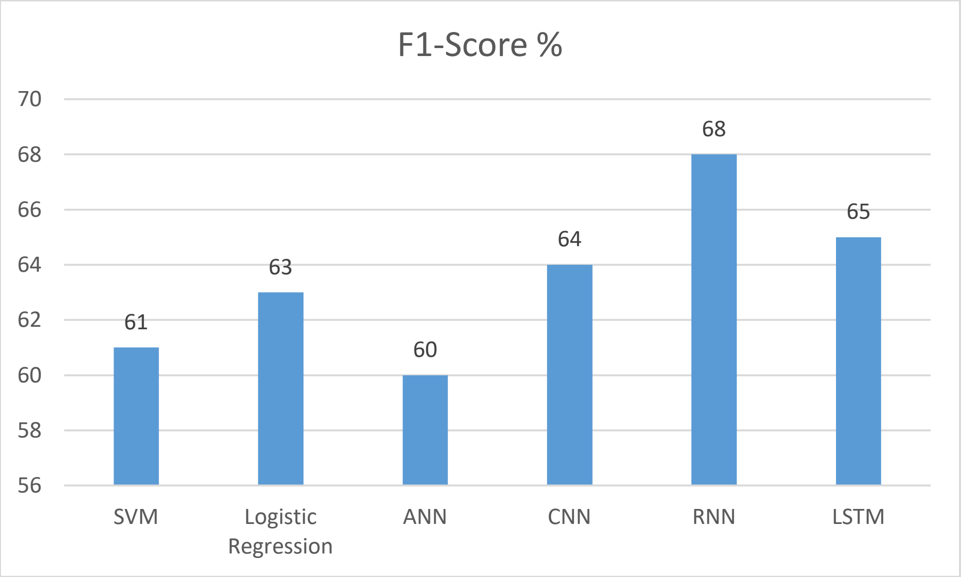

Figure.4: F_1-Scores of the Classifiers

The image part with relationship ID rId17 was not found in the file.

Figure 6 compares the accuracy, precision, recall, and F1-Score for the classification approaches SVM, logistic regression, ANN, CNN, RNN, and LSTM.

4. Conclusion:

This study highlights the significant potential of machine learning and deep learning technologies in the early detection and classification of breast cancer, which is crucial for reducing mortality rates. By evaluating the performance of various algorithms such as Support Vector Machine (SVM), Logistic Regression, Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM), we identified that the RNN model excels in accuracy, precision, and F1-Score, while ANN shows superior recall. These findings emphasize the importance of leveraging advanced computational methods to enhance diagnostic accuracy and provide better patient outcomes. However, this study has certain limitations that need to be acknowledged. Firstly, the dataset size and diversity may limit the generalizability of the results. Future research should aim to include larger and more diverse datasets to validate the findings across different populations. Secondly, while we evaluated multiple machine learning models, there is always scope for exploring additional algorithms and hybrid models to further improve performance. Lastly, the interpretability of complex models like deep learning remains a challenge, and efforts should be made to enhance the transparency and explainability of these models for clinical use. In summary, our research underscores the transformative potential of machine learning in breast cancer detection and calls for continued advancements and evaluations to address existing limitations and optimize these technologies for clinical applications.

References

1. Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert systems with Applications*, *36*(2), 3465-3469.
2. Ravdin, P. M., & Clark, G. M. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast cancer research and treatment*, *22*(3), 285-293.
3. Zahid, U., Ashraf, I., Khan, M. A., Alhaisoni, M., Yahya, K. M., Hussein, H. S., & Alshazly, H. (2022). BrainNet: optimal deep learning feature fusion for brain tumor classification. *Computational Intelligence and Neuroscience*, *2022*.
4. Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. *Journal of artificial intelligence research*, *4*, 77-90.
5. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, *34*(2), 113-127.
6. Khan, M. A., Azhar, M., Ibrar, K., Alqahtani, A., Alsubai, S., Binbusayyis, A., ... & Chang, B. (2022). COVID-19 classification from chest X-ray images: a framework of

deep explainable artificial intelligence. *Computational Intelligence and Neuroscience*, *2022*.

7. Jabeen, K., Khan, M. A., Alhaisoni, M., Tariq, U., Zhang, Y. D., Hamza, A., ... & Damaševičius, R. (2022). Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. *Sensors*, *22*(3), 807.

8. Kaya, Y., & Uyar, M. (2013). A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. *Applied Soft Computing*, *13*(8), 3429-3438.

9. Gong, H., Qian, M., Pan, G., & Hu, B. (2021). Ultrasound image texture features learning-based breast cancer benign and malignant classification. *Computational and Mathematical Methods in Medicine*, *2021*.

10. Zhang, H., Liu, H., Ma, L., Liu, J., & Hu, D. (2021). Ultrasound image features under deep learning in breast conservation surgery for breast cancer. *Journal of Healthcare Engineering*, *2021*.

11. Chen, D. R., Chien, C. L., & Kuo, Y. F. (2015). Computer-aided assessment of tumor grade for breast cancer in ultrasound images. *Computational and mathematical methods in medicine*, *2015*.

12. Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital signal processing*, *17*(4), 694-701.

13. Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert systems with applications*, *36*(2), 3240-3247.

14. Yeh, W. C., Chang, W. W., & Chung, Y. Y. (2009). A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems with Applications, 36*(4), 8204-8211.