2026,Volume 3 , Issue 9 , PSIJK (63-68)

Pure Sciences International Journal of Kerbala

Enhancing Cancer Diagnosis through Machine Learning and Swarm Intelligence Techniques

Maha Sabri Altememe [a*]

a College of Computer Science and Information Technology, University of Kerbala, Iraq

P A P E R I N F O

Received: 12.09.2025
Accepted: 15.12.2025
Published: 31.03.2026

***Keywords*:**
classification, breast cancer, medical images, ABC algorithm

A b s t r a c t

Machine learning has achieved wide success in detecting and identifying cancer diseases, especially those based on medical images. To increase the performance of models, machine learning techniques are combined with other techniques in the processing stage. In this research, the ABC algorithm was used in the feature identification process to help the algorithm choose the optimal solution more accurately.The medical images used are CT images of breast cancer collected in the form of a standard dataset of 570 known as WDBC dataset. The proposed model achieved classification accuracy 94.74%.

DOI: 10.53851/psijk.v3.i9.63-68

1. INTRODUCTION

Today, artificial intelligence has become widely important in many applications through the use of its multiple branches. Among these branches are machine learning techniques that are used to classify data and achieve high performance in implementation. The implementation of machine learning techniques requires the use of data for the purpose of training the model and then predicting the results and making decisions on new data (Zhang et al., 2020). However, the effectiveness of the models depends on a number of determinants, including the percentage of data, its quality, the type of algorithm that was applied and its suitability for this data, in addition to the method of extracting features from this data. Therefore, to increase the efficiency of the model, algorithms are used that help extract important features. One of these algorithms is the artificial bee colony algorithm, which works to select the ideal features and can be applied to medical images, especially in medical image data for cancer. It is considered one of the major determinants due to the differences in cancer diseases, so it needs high-quality techniques to infer features(Tharwat et al., 2017). In this research, a model is designed that combines machine learning techniques and ABC algorithm to select optimal features. Contributions to this study: Artificial Bee Colony algorithm-based feature selection for breast cancer detection. In order to improve patient care and diagnosis accuracy in the field of oncology, the project intends to create an efficient .

1. 1. Artificial Bee Colony algorithm

Karaboga presented the Artificial Bee Colony (ABC) algorithm in 2005 (Karaboga, 2005) as a swarm-based technique to solve numerical problems (Rao et al., 2020) This algorithm's concept is derived from the clever ways that colonies of honey bees hunt for food. The work crew is divided into three primary groups. The first group is known as the scout bees, and their primary duty is to randomly hunt for food. The job of the second team, known as the worker bees, is to look for food sources around the ones that have already been found. The job of the third team, known as the observer bees, is to keep an eye on the worker bees and assess the food's worth and quality in order to choose the best (Tharwat et al., 2017). In the ABC theory, the search process extracts the features present in the data in general, which represents finding specific solutions to the problem, and then the optimization stage is chosen by finding the best types of features, and this process is repeated repeatedly until the best features are obtained that help us in the classification stages. To clarify the work of the algorithm that is used to find the optimal solution, as mentioned previously, it consists of three groups: worker bees, spectator bees, and scout bees. The following code illustrates the steps of the ABC algorithm. Pseudo-code of the ABC algorithm (Rao et al., 2020) (Shahini Shamsabadi & B. S., 2009).

1. 2. Machin learning algorithms

Machine learning algorithms are mathematical techniques that help understand patterns, make decisions based on data, and predict results without the need for continuous programming. In the field of artificial intelligence, there are a number of learnable algorithms, including machine learning and deep learning. In these techniques, the model can be trained on a portion of the data and is called the learning phase. Based on this phase, the results are extracted and decisions are made based on the experience gained by the model without human intervention using these algorithms. Machine learning algorithms are divided into two groups, the first is called supervised and the other is unsupervised. To implement these algorithms, a large amount of data is needed (Shahini Shamsabadi & B. S., 2009) (Vetteth & K. W., 2003).

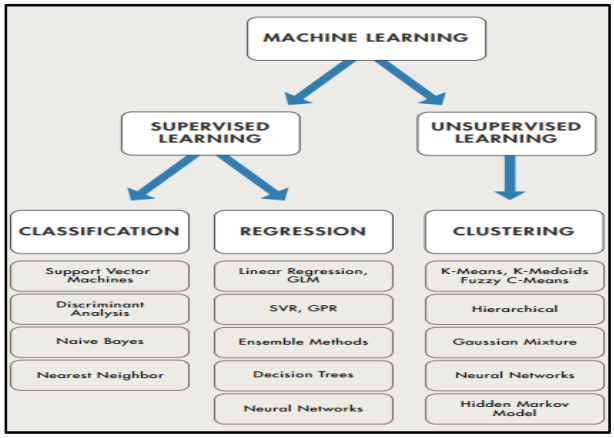


Figure 1. Types Of Machin Learning(Lent & M. L., 2006).

1.2.1 Logistic Regression

A statistical technique called logistic regression is employed to examine the correlation between a set of independent variables and a binary dependent variable.

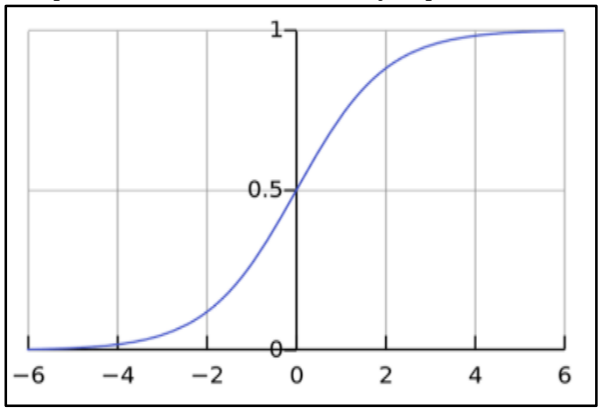


Figure 2. displays the regression function(Karaboga & Ozturk, 2011).

The goal of using logistic regression is to ascertain the odds, or adjusted odds, of a specific event taking place in relation to the independent variables. In order to transform the regression results into a range between 0 and 1, which can be used to produce probability estimates, the logistic function utilized in logistic regression is crucial. A useful tool for data analysis and forecasting decisions is logistic regression. It can be useful in determining the variables that affect an event's likelihood of occurring and in calculating the event's probability based on the available independent variables (Fallah-Mehdipour et al., 2018).

1.2.2 Decision Tree

A decision tree is a binary tree where each node is assigned a letter and each edge represents a possible value for the letter. A path label is produced that leads to each leaf of the tree, and the result is either 0 or 1. The number of nodes in the decision tree determines its size.

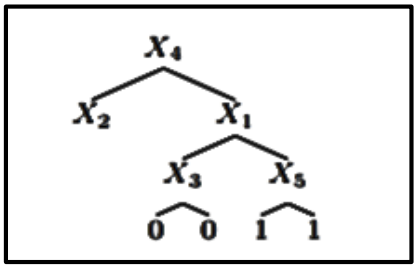
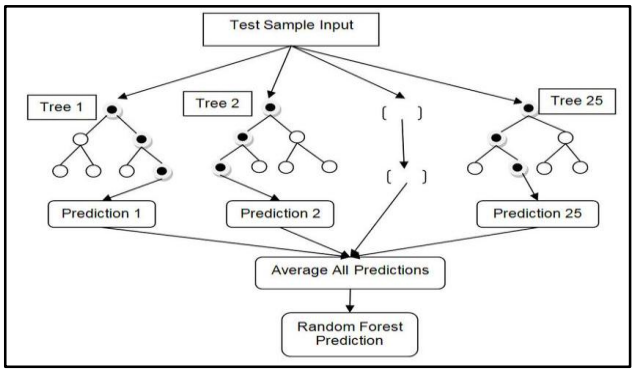


Figure 3: shows an example of a decision tree (Karaboga & Ozturk, 2011).

1.2.3 Random Forest

As the name implies, it is an improvement on decision trees known as the random forest and is made up of numerous independent decision trees that collaborate to produce predictions that are more reliable and accurate. The highest voting result is derived from this group's prediction outcomes, which is better than the outcome of employing the best model alone. In Figure 4 the algorithm is shown (Huang & M. M., 2007).

Figure 4. Random Forest (Harini & Uma Maheswari, 2023).



2. RELATED WORKS

Over the past ten years, artificial intelligence has become widely used for the identification of different tumors. Consequently, it is vital to look for extremely effective

techniques. I'll go over a few of the accomplishments from the previous years, which include the following:

In (Tharwat et al., 2017) this work presents a hybrid technique that selects features using ABC algorithm and classifies them using SVM. The contributions of this work are through the importance of removing irrelevant data features that affect classification performance by applying the SVM algorithm. The established method is typically employed in the diagnosis of diabetes and liver illnesses, which are prevalent and lower quality of life. The UCI database's hepatitis, liver disease, and diabetes datasets were utilized to diagnose these conditions; the suggested method achieved classification accuracy of 94.92%, 74.81%, and 79.29%, in that order.

In (Walus & T. D., 2004) This paper proposes a Deep CNN system, which outperforms the Vision Transformer model and other transfer learning models in terms of accuracy. In this work, the model demonstrated high performance in breast cancer detection and classification by combining the advantages of MobileNet and Xception models. The proposed model has a remarkable accuracy of 87.82% according to our test results.

In (Khurasia, 2006) this paper, used on Self-Care Activities Dataset based on ICF-CY (SCADI) with disabilities containing 206 features was used. Therefore, the ABC algorithm was applied to select the features for classification purpose. Only 7 features were selected by the algorithm and an accuracy rate of 88.5714% and a scale value of 0.871 were obtained, while the percentage was lower, 84.2857%, was obtained by applying Gain Ratio and Chi-Square.

3.ROPOSED METHODOLOGY

In this research, the ABC algorithm was applied to features selection and then we applied three different classification models for breast cancer diagnosis, including Random Forest (RF), Decision Tree (DT) and Logistic Regression (LR). These process stages were applied to breast cancer data by using cross-sectional images and Figure 6 shows the model's working stages.

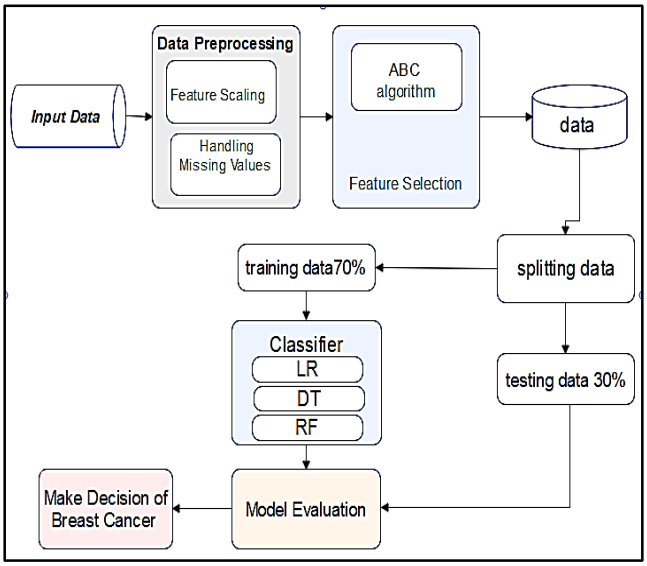


Figure 5. Proposed Methodology Diagram

3.1. Database

The Breast Cancer Wisconsin (Diagnostic) Data Set is a popular dataset for big data and machine learning applications. When dealing with issues related to the categorization of diagnosis for breast cancer, it is really helpful. It possesses traits that can be used to identify malignant or benign masses. These properties are calculated using digital photographs. There are 570 cancer-related photos in the dataset. In table 1. Explain the characteristics of dataset and Figure 6 shows a sample of the data.

Table 1. Properties of the Dataset

Inf	No.
Total number of situations	570
Attributes in data	30
Classification	2(Malignat,Benign)

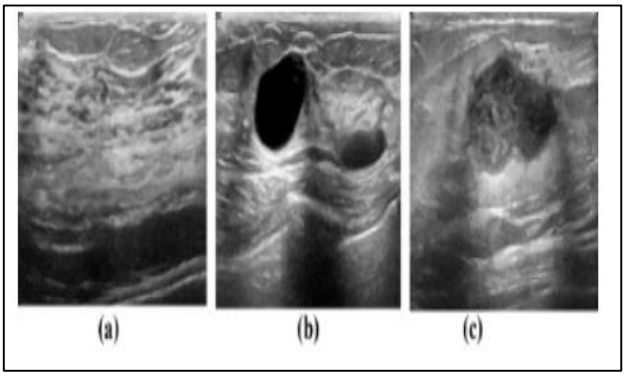


Figure 6.Sample Test

3.2 Data preprocessing

Dataset analysis is the process of looking through, cleaning, transforming, and modeling data to make sense of it, uncover useful information, and help make decisions. The purpose of the data cleaning process is to remove any anomalies, including missing or distorted information, from the data that require attention.

3.3 Feature Selection

This process, which is the feature selection process, is used to obtain the most relevant features in the dataset, which helps improve the performance of machine learning algorithms. The bee algorithm was applied because it is useful and achieves good results with high-dimensional data because it can identify complex relationships between features and is also flexible with different machine learning algorithms. The program looks for the best answer using a colony of bees. Every bee has a fitness value, which indicates the extent to which the features it represents improve the machine learning model's accuracy. Higher fitness bees are more proximate to the ideal outcome. The figure 7 shows the results of feature selection and the figure 8 shows the cross-sectional image after feature selection. According to the data used, a number of features were relied upon, including the following in table 2. In figure 7 shows Illustrative images after extracting features.

Table 2. Properties of the features for Dataset

No	Features	Description
1.	Radius:	Mean of distances from center to points on the perimeter.
2.	Texture:	Standard deviation of gray-scale values.
3.	Perimeter:	Perimeter of the cell nucleus.
4.	Area:	Area of the cell nucleus.
5.	Smoothness:	Variation in radius lengths.
6.	Compactness:	Perimeter^2/Area - 1.0.
7.	Concavity:	Severity of concave portions of the contour.
8.	Concave Points:	Number of concave portions of the contour.
9.	Symmetry:	Symmetry of cell nuclei.

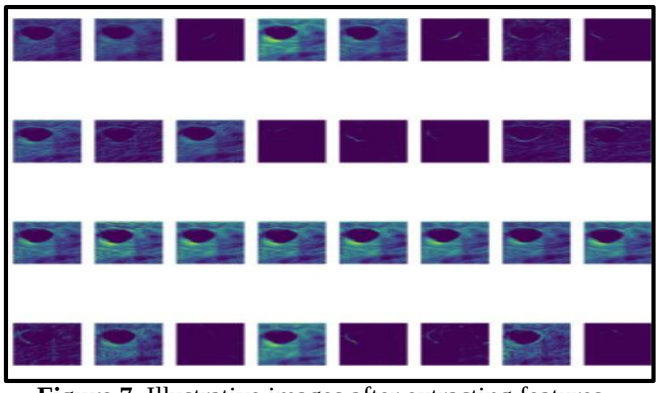


Figure 7. Illustrative images after extracting features

3.4 Implementation Machin learning

The algorithmic performance in terms of accuracy, precision, recall, and F1 score is displayed in the table. Among the algorithms tested, the RF Classifier yielded the lowest accuracy rate. The decision tree approach also produced the best accuracy rate. Additionally, the recall, precision, and F1 scores for each approach are included in this table. When it came to all confusion matrices and other scores, the decision tree strategy performed admirably overall. In table 3 shows the outcomes for all algorithms.

TABLE 3. Outcomes For Implementation

ML Algorithms	**Dataset**	**Processing**	Result of Model
			Accuracy
LR	Collection Dataset	Training	94.95%
		Testing	92.98%
RF	Collection Dataset	Training	1.00%
		Testing	94.74%
DT	Collection Dataset	Training	52.98%
		Testing	61.40%

- Logistic Regression: For both the training and test datasets, the accuracy score is calculated. This shows how effectively the model predicts the categories of the training set of data. With a score of around 94.95%, the model has classified 94.95% of the training dataset's cases correctly. This gauges how well the model adapts to fresh, untested data. A score of roughly 92.98% indicates that roughly 92.98% of the test dataset's cases were properly classified by the model.
- Decision Tree: This algorithm was This algorithm was used to process the ready-to-classify data, which was obtained from images of infected people. We obtained somewhat acceptable results, as the accuracy rate during training reached up to 52% and 61% testing data.
- Random Forest: While implementing the algorithm on the same data that was mentioned previously, it achieved good results, as the accuracy of 1% was achieved during the training phase, while the accuracy rate during the testing phase was 94%. With these results, the RF algorithm achieved the highest results that can be relied upon in discovering and classifying data.

1. CONCLUSION

Breast cancer has become common recently and it causes death for women, so it has become necessary to detect this disease early with wide importance. This helps to reduce the mortality rate because early detection helps in speeding up the treatment of this disease. Early breast cancer tumor detection has grown more precise and effective in recent times thanks to the development of sophisticated machine-learning classifiers. In order to diagnose breast cancer, this work combines feature selection and classifiers. To test the classification accuracy of three different methods, we ran them on the WDBC dataset. With an average accuracy of 94.00%, our study's results demonstrate that the RF model was the most successful. The topic of machine learning for breast cancer diagnosis has a number of opportunities for future research and development. To increase the precision and dependability of breast cancer detection, integrating several machine-learning algorithms is one possible field of study. To help find the most pertinent features for breast cancer diagnosis, more research into feature selection techniques may be conducted. This could ultimately result in more precise and effective diagnoses.

References

Altememe, M. S., & El Abbadi, N. K. (2022). Gesture interpreting of alphabet Arabic sign language based on machine learning algorithms. Iraqi International Conference on Communication & Information Technologies (IICCIT-2022), University of Basrah.

Altememe, M. S., & El Abbadi, N. K. (2023). A hybrid model between one-dimensional convolutional neural network and machine learning algorithms for Arabic sign language word recognition. 4th International Scientific Conference of AlKafeel University (ISCKU 2022).

Fallah-Mehdipour, E., Bozorg-Haddad, O., & Mariño, M. (2018). Prediction and simulation of monthly groundwater levels by genetic programming. Journal of Hydro-environment Research, 7, 253–260. https://doi.org/10.1016/j.jher.2013.03.005

Harini, K., & Uma Maheswari, S. (2023). A novel static and dynamic hand gesture recognition using self-organizing map with deep convolutional neural network. Journal for Control, Measurement, Electronics, Computing and Communications, 64(4).

Huang, J., & others. (2007). Design of sequential circuits by quantum-dot cellular automata. Microelectronics Journal, 38, 525–537.

Islam, M. R., Rahman, M. M., Ali, M. S., Nafi, A. A. N., Alam, M. S., Godder, T. K., Miah, M. S., & Islam, M. K. (2024). Enhancing breast cancer segmentation and classification using ensemble deep convolutional neural network and U-net. Machine Learning with Applications. https://doi.org/10.1016/j.mlwa.2024.100555

Karaboga, D., & Ozturk, C. (2011). A novel clustering approach: Artificial bee colony (ABC) algorithm. Applied Soft Computing, 11(1), 652–657.

Keleş, M. K., & Kılıç, Ü. (2018). Artificial bee colony algorithm for feature selection on SCADI dataset. 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo. https://doi.org/10.1109/UBMK.2018.8566287

Khurasia, P. G. (2006). Quantum cellular automata.

Kim, K., & others. (2006). Quantum-dot cellular automata design guideline. IEICE Transactions on Fundamentals, 6, 1607–1614.

Lent, C. S., & others. (2006). Bennett clocking of quantum-dot cellular automata. Nanotechnology, 17, 4240–4251.

Liu, J., et al. (2019). River level estimation using camera images and machine learning. Water Resources Research, 55(5), 4321–4335.

Mustafa, W. A., & Abdul Kader, M. M. (2017). A review of histogram equalization techniques in image enhancement application. 1st International Conference on Green and Sustainable Computing (ICoGeS). https://doi.org/10.1088/1742-6596/1019/1/012026

Niemir, M. (2004). Designing digital systems in quantum cellular automata (Master's thesis). University of Notre Dame.

Rao, S. S., et al. (2020). Machine learning approaches for water level prediction in rivers: A review. IEEE Journal of Selected Topics in Applied Earth Observations and

Remote Sensing, 13, 5325–5344.
Sara, H., & others. (2012). New robust QCA D flip-flop and memory structures. Microelectronics Journal, 43, 929–940.
Shahini Shamsabadi, A., & others. (2009). Applying inherent capabilities of quantum-dot cellular automata to design: D flip-flop case study. Journal of Systems Architecture, 55, 180–187.
Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. AI Communications, 30, 169–190.
Tougaw, P. D., & Lent, C. S. (1994). Logical devices implemented using quantum cellular automata. Journal of Applied Physics, 75, 1818–1824.
Uzer, M. S., & Yilmaz, N. (2013). Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification. The Scientific World Journal. https://doi.org/10.1155/2013/419187
Vankamamidi, V., & others. (2008). Two-dimensional schemes for clocking/timing of QCA circuits. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 27, 34–44.
Vetteth, A., & others. (2003). RAM design using quantum-dot cellular automata. Nanotechnology Conference and Trade Show, 2, 160–163.
Walus, K., & others. (2004). QCA designer: A rapid design and simulation tool for quantum-dot cellular automata. IEEE Transactions on Nanotechnology, 3, 26–31.
Zhang, R., et al. (2004). A method of majority logic reduction for quantum cellular automata. IEEE Transactions on Nanotechnology, 3(4), 443–450.
Zhang, Y., et al. (2020). River level estimation from river-camera images using deep learning. IEEE Transactions on Geoscience and Remote Sensing, 58(5), 3421–3432.