

Development and Optimization of Random Forest Algorithm for Breast Cancer Diagnosis

I. Z. Waziri,^{1,2,3} U. N. Muhtar,¹ A. N. Adam,¹ S. Abdullahi,^{4,5} A. M. Gumel^{1*}

¹Department of Microbiology and Biotechnology, Federal University Dutse, P.M.B 7156, Jigawa State-Nigeria

²Department of Biochemistry, Ahmadu Bello University, Zaria, Kaduna State, P.M.B 1045, Nigeria.

³Africa Centre of Excellence for Neglected Tropical Diseases and Forensic Biotechnology, Ahmadu Bello University, Zaria, Kaduna State, P.M.B 1045, Nigeria.

⁴Department of Public Health, College of Health Sciences, Jigawa State Polytechnic, Dutse, P.M.B 7040, Nigeria.

⁵Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, Guangdong, China.



*Corresponding Author:

Prof. Ahmad Mohammed Gumel

Department of Microbiology and Biotechnology,
Federal University Dutse, P.M.B 7156, Jigawa State-
Nigeria.

am.gumel@fud.edu.ng

+2348161101920

ABSTRACT

Breast cancer represents a pervasive global health challenge, often manifesting as small masses or tumors within breast tissue, commonly found in milk ducts and glands. The global incidence of breast cancer, especially among women, has been steadily increasing. Timely detection plays a pivotal role in successful treatment and preventing fatalities. Leveraging advancements in machine learning techniques, various models have been developed for the timely diagnosis of breast cancer. Consequently, the precision and accuracy of these models in diagnosing breast cancer are paramount. This study focuses on the meticulous process of fine-tuning random forest hyperparameters using grid search to create a high-accuracy model for classifying breast cancer into benign and malignant categories. The optimal set of hyperparameters identified through the grid search includes: the number of trees in the forest (150), maximum depth of tree nodes (None), minimum number of samples required to split an internal node (2), minimum number of samples required at a leaf node (1), and the seed for random number generation (123). These refined parameters were utilized to train a random forest model, resulting in an outstanding 99.12% accuracy for breast cancer classification. The hyperparameter-optimized Random Forest classifier, as implemented in our study, showcases exceptional accuracy, surpassing not only previous results but also outperforming established algorithms. This noteworthy achievement underscores the robustness and efficacy of the proposed model in discerning intricate patterns within breast cancer datasets, making a substantial contribution to the ongoing quest for precision and effectiveness in breast cancer classification.

Key words: Breast cancer, machine learning, random forest, precision medicine, diagnosis and healthcare AI.

INTRODUCTION

Breast cancer is a prevalent global health concern, often presenting as small masses or tumors in breast tissue, typically within milk ducts and glands (Bhardwaj et al. 2022). The incidence of breast cancer has been on the rise worldwide, particularly among women (Chen et al. 2023). Timely detection is crucial for successful treatment and preventing fatalities. Breast cancer diagnosis can be achieved through mammographic examinations or physical assessments (Yedjou et al. 2021). Mammography is a pivotal screening method recommended for mass breast cancer diagnosis. Alongside mammography, specialized physician examinations play a vital role in early breast cancer detection. Although, some breast cancers may elude diagnosis through mammography and ultrasound alone, underscoring the importance of physical examinations as a complementary approach (Ozcan, Aydin, and Cetinkaya 2022). Interpreting and diagnosing test results in breast cancer diagnosis require expert human knowledge (Yedjou et al. 2021). Nevertheless, with advancements in machine learning techniques, promising research is being conducted in breast cancer diagnosis, as in other cancer types (Bhardwaj et al. 2022).

Machine learning (ML) is a subset of Artificial Intelligence (AI) that empowers computers to swiftly recognize patterns within intricate and extensive datasets by learning from pre-existing data (Jiang et al. 2017). The adoption of ML in supporting healthcare professionals is rapidly growing (Anon 2023). ML techniques, which are progressively gaining prominence in the diagnosis of various cancer types, are also being applied to detect breast cancer (Uzun Ozsahin et al. 2022). Through advancements in data science, ML has assumed a role in decision-making and refined analytical procedures (Tabari et al. 2022). A crucial component of effective breast cancer treatment involves categorizing tumor types within the breast as benign, malignant, or normal tissue while minimizing misdiagnoses. ML algorithms can facilitate this classification (Rautela, Kumar, and Kumar 2022). Breast cancer ranks as the second most prevalent cancer among all types and can prove fatal without accurate and early diagnosis. Hence, achieving precise and high-performance breast cancer diagnosis is of utmost importance (Chen et al. 2023).

A Random Forest classifier is an ensemble learning method in machine learning that operates based on the principles of decision trees. It combines the outputs of multiple individual decision trees to make more accurate and robust prediction (Biau and Scornet 2016). The "random" aspect comes from two key sources of randomness in the algorithm. Firstly, during the construction of each decision tree, a random subset of features is considered at each split, preventing the dominance of any single feature. Secondly, the algorithm builds each tree on a randomly sampled subset of the training data, a process known as bootstrapping. By aggregating the predictions from numerous diverse and slightly different trees, Random Forest reduces overfitting, enhances generalization, and improves overall predictive performance. The final prediction is typically determined by a majority voting scheme, where the class receiving the most votes from individual trees is selected as the output (Belgiu and Drăguț 2016). This ensemble approach makes Random Forests versatile, robust, and effective for various classification tasks (Assiri, Nazir, and Velastin 2020).

Delen et al., (2005) preclassified 202,932 breast cancer patient records with an accuracy of approximately 93% in predicting survivorship. Aruna et al., (2011) compared Naïve Bayes (NB), C4.5, Support Vector Machine (SVM), and K-nearest neighbor (KNN) classifiers in Wisconsin Breast Cancer Dataset (WBCD), identifying SVM as the most accurate classifier with a precision of 96.99%. (Liu, Wang, and Zhang 2009) utilized decision trees on unbalanced data to predict 5-year survival rates of breast cancer, demonstrating enhanced model performance through a combination of undersampling and the bagging procedure. (Amrane et al. 2018) delved into the Naïve Bayes and KNN classifiers, employing a cross-validation scheme for accuracy assessment. The study found that the KNN algorithm outperformed the NB algorithm with an accuracy of 95.7. In the study of Bhardwaj et al., (2022), various machine learning algorithms, including multilayer perceptron (MLP), K-nearest neighbor (KNN), genetic programming (GP), and random forest (RF), were compared for their efficacy in classifying benign and malignant cases on the WBCD dataset. The findings revealed that RF exhibited an impressive classification accuracy of 96.24%.

Despite the superior performance of random forest compared to other classifiers, there is a potential for further enhancement through hyperparameter tuning. Therefore, this study delves into the process of fine-tuning sets of random forest hyperparameters using grid search to further elevate the accuracy and performance of the random forest model for breast cancer classification.

2. Materials and Methods

2.1. Dataset Acquisition and Exploratory Data Analysis

2.1.1 Dataset Source

The breast cancer dataset utilized in this paper originates from the UCI dataset and was generously provided by Dr. William from the Clinical Medicine Research Institute at the University of Wisconsin. It comprises features computed from a digitized image of a fine needle aspirate (FNA) of a

breast mass, describing the characteristics of cell nuclei within the image. The dataset was obtained from <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>.

This dataset comprises a total of 569 experimental samples, consisting of 357 benign and 212 malignant breast cancer samples with 30 predictive variables and a target variable denoting the diagnosis.

2.1.2 Data Loading, Cleaning and Exploration

The dataset was loaded into a Pandas DataFrame using the `pd.read_csv` function in Python. Initial exploration was conducted to examine the structure and characteristics of the dataset.

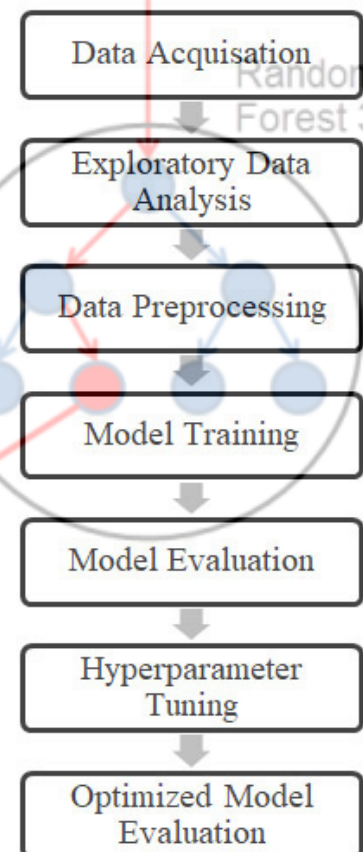


Figure 1 Flowchart of the study approach

2.2. Data Preprocessing

2.2.1 Feature and Target Variable Extraction

The features (X) and the target variable (y) were extracted from the dataset. The target variable, denoting the diagnosis, was separated from the feature set.

2.2.2 Train-Test Split

The dataset was divided into training and testing sets using the `train_test_split` function from scikit-learn, with a test size of 20%.

2.3. Random Forest Model Training

A Random Forest classifier was created with default parameters using the `RandomForestClassifier` from scikit-learn. The key specifications included 100 trees in the forest, a maximum depth of tree nodes set to none, a minimum of 2 samples required to split internal nodes, and a minimum of 1 sample required to be at a leaf node. To ensure reproducibility, a fixed random seed was employed. Subsequently, the model was trained on the training set using the `fit` method.

2.4. Model Evaluation

The trained Random Forest model was employed to make predictions on the test set. Performance metrics, including accuracy, confusion matrix, and classification report, were computed using scikit-learn functions.

2.5. Hyperparameter Tuning using Grid Search

A hyperparameter grid was defined for the Grid Search, encompassing various values for the random forest hyperparameters. These include, number of trees in the forest: [50, 100, 150], maximum depth of the tree nodes: [None, 10, 20, 30], minimum number of samples required to split an internal node: [2, 5, 10], minimum number of samples required to be at a leaf node: [1, 2, 4] and seed for random number generation: [42, 123, 150]. Subsequently, a grid search was executed to identify the optimal combination of hyperparameters for the random forest model. Once the optimal set was determined, the random forest model was retrained using these selected hyperparameters. The retrained random forest model was employed to make predictions on the test set. The performance metrics were computed using scikit-learn functions and compared with those of the initially trained random forest model, which used default parameters.

3. Results and Discussion

3.1. Exploratory Data Analysis

The exploratory data analysis (EDA) conducted on the Breast Cancer Wisconsin dataset, encompassing 569 samples and 30 features, uncovered a well-structured dataset. Subsequently, the unnecessary columns "id" and "Unnamed: 32" were excluded from consideration for modeling. The distribution of the target variable, representing the diagnosis of benign and malignant cases, is visually presented in Figure 2. Specifically, the dataset comprises 357 instances classified as benign and 212 instances classified as malignant.

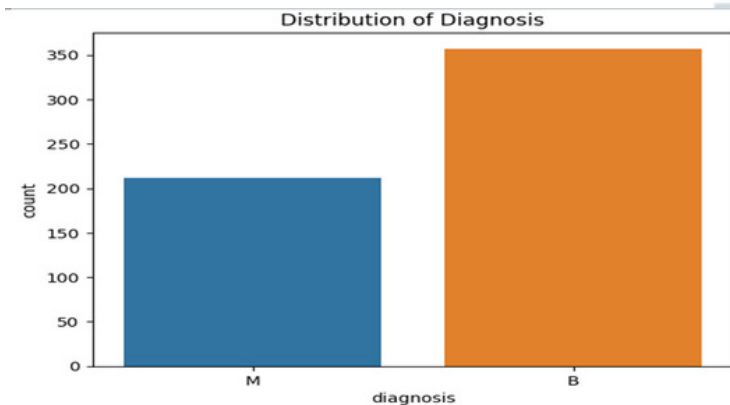


Figure 2 Distribution of Diagnosis (M, Malignant and B, Benign)

3.2. Initial Model Performance

The confusion matrix in figure 3 summarizes the performance of the random forest model that was initially trained on the dataset with default hyperparameters. The rows of the confusion matrix represent the actual labels of the data, and the columns represent the predicted labels of the model.

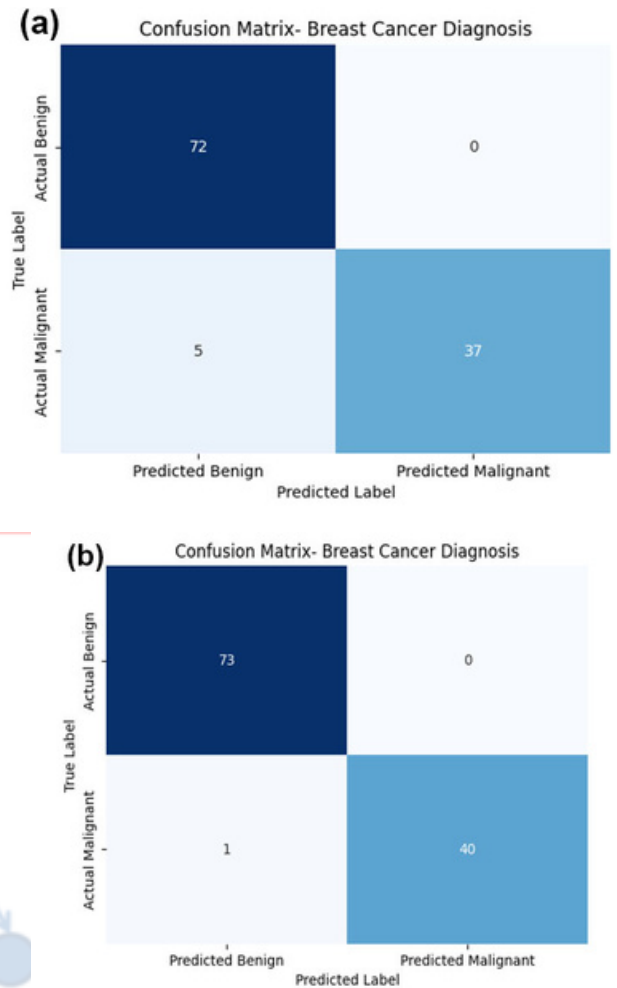


Figure 3: Confusion Matrix plot for (a)Initial model (b) Optimized model

The four cells of the confusion matrix are:

True positives (TP): These are the cases where the model correctly predicted that the data point belonged to a certain class.

False positives (FP): These are the cases where the model incorrectly predicted that the data point belonged to a certain class.

False negatives (FN): These are the cases where the model incorrectly predicted that the data point did not belong to a certain class, when it actually did.

True negatives (TN): These are the cases where the model correctly predicted that the data point did not belong to a certain class.

The confusion matrix shows that there were 71 true positives and 37 true negatives, with 0 false positive and 5 false negatives. This indicated that the model performed well on the test data. This reveals that the initial model correctly identified 71 instances as positive (malignant) and 37 instances as negative (benign), reflecting a good ability to capture true cases. However, the model made 5 false negative errors (misclassifying malignant as benign).

The confusion matrix was used to calculate a number of performance metrics of the models, such as accuracy, precision, recall, and F1 score.

Accuracy: this measure the overall accuracy of the model.

Accuracy: this measure the overall accuracy of the model.

$$Accuracy = \frac{TP + TN + FP + FN}{TP + TN}$$

Precision: Precision measures the accuracy of the positive predictions.

$$Precision = \frac{TP + FP}{TP}$$

Recall (Sensitivity or True Positive Rate): Recall measures the ability of the model to capture all the relevant instances.

$$Recall = \frac{TP + FN}{TP}$$

F1 Score: is the harmonic mean of precision and recall. It provides a balance between precision and recall.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Table 1 Performance report of the initial model and optimized model

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Initial Model	95.61	95.89	95.61	95.54
Optimized Model	99.12	99.13	99.12	99.12

The performance report, summarizing accuracy, precision, recall, and F1-score, are presented in Table 1. These metrics indicate the performance of the initial model. The accuracy of 95.61% reflects the overall correctness of the classifications. Precision, recall, and F1-score, at 95.89%, 95.89% and 95.54% respectively, demonstrate a balanced trade-off between precision (ability to avoid false positives) and recall (ability to capture true positives). While the initial model demonstrates commendable performance there is still evident room for improvement through hyperparameter tuning.

3.3 Hyperparameter Tuning

To optimize the model's performance, a hyperparameter tuning step was conducted using Grid Search. The best set of hyperparameters obtained from the search were; number of trees in the forest: 150, maximum depth of the tree nodes: None, minimum number of samples required to split an internal node: 2, Minimum number of samples required to be at a leaf node: 1 and Seed for random number generation: 123. The model was retrained using these optimal hyperparameters.

3.4. Optimized Model Performance

The updated confusion matrix in figure 4 shows that the optimized model performed excellently on the test. The optimized model demonstrates an improvement in performance with 73 true positive identifications and 40 true negatives. Notably, there are no false positives and only 1 false negative error is observed, signifying a reduction in misclassifying malignant cases as benign. In comparison, the hyperparameter-optimized model exhibits an overall enhancement in classification accuracy and a reduction in false negative errors. The absence of false positives in the optimized model is particularly notable, as it indicates a more conservative and precise classification of benign cases. The decrease in false negatives further underscores the improved sensitivity of the optimized model in correctly identifying malignant cases. The performance report presented in Table 1 show that the hyperparameter-optimized model exhibits a remarkable improvement across all performance metrics. The accuracy of 99.12% signifies an almost flawless classification accuracy, reflecting the model's precision in making correct predictions. Precision, recall, and F1-score, all at 99.13%, 99.12% and 99.12% respectively, indicate a high level of consistency and accuracy in both positive and negative classifications. This improvement suggests that the fine-tuning of hyperparameters has significantly enhanced the model's ability to make accurate predictions, resulting in a more reliable and precise breast cancer classification.

Table 2 Performance Comparison with other work

Author	Year	Method and Classifier	Accuracy (%)
Quinlan	1996	C4.5	94.74
Hamilton et al.	1996	RIAC	95
Nauck and Kruse	1999	Neuro-fuzzy	95.06
Abonyi and Szeifert	2003	Supervised fuzzy clustering	95.57
Lavanya and Rani	2011	Decision tree algorithms	92.27
Xue et al.	2014	Particle swarm optimization	94.74
Rajaguru and S R	2019	PCA with K-NN algorithm as classifier	95.95
Bhardwaj et al.	2022	Random Forest	96.24
Chen et al.	2023	XGBoost	97.4
Our study	2024	Optimized Random Forest	99.12

Table 2 presented the accuracy results from various classifiers across different studies, this provide insightful perspectives on their performance in breast cancer classification. Notably, the hyperparameter optimized random forest classifier applied in our study emerged with the highest accuracy of 99.12 %, surpassing other work and renowned algorithms such as C4.5 (Quinlan 1996), RIAC (Hamilton, Shan, and Cercone 1996), Neuro-fuzzy (Nauck and Kruse 1999), Supervised Fuzzy Clustering (Abonyi and Szeifert 2003) , Decision Tree Algorithms (Doddipalli and Rani 2011), Particle Swarm Optimization (Xue, Zhang, and Browne 2014), PCA with K-NN algorithm as classifier (Rajaguru and S R 2019), Random Forest (Bhardwaj et al. 2022) and XGBoost (Chen et al. 2023). This outcome underscores the robustness of the proposed model in capturing complex patterns within the breast cancer dataset, making it a promising choice for accurate breast cancer diagnosis.

4. Conclusions

The comprehensive evaluation of various classifiers, as detailed in Table 2, sheds light on their respective performances in the context of breast cancer classification. The hyperparameter-optimized Random Forest classifier, as implemented in our study, has demonstrated exceptional accuracy, reaching an impressive 99.12%. This achievement surpasses not only the results from prior works but also renowned algorithms, including C4.5, RIAC, Neuro-fuzzy, Supervised Fuzzy Clustering, Decision Tree Algorithms, Particle Swarm Optimization, Random Forest and XGBoost.

The notable outcome emphasizes the robustness and efficacy of the proposed model in discerning intricate patterns within the breast cancer dataset. Such high accuracy positions the optimized Random Forest classifier as a promising and reliable choice for accurate breast cancer diagnosis or classification into benign and malignant categories. As advancements in machine learning techniques continue to refine the landscape of medical diagnostics, the superior performance of our hyperparameter-optimized Random Forest model signifies a significant contribution to the ongoing pursuit of precision and effectiveness in breast cancer classification. This study underscores the potential impact of thoughtful hyperparameter tuning in enhancing the capabilities of machine learning models for critical medical applications.

The focus of this study is solely on the creation of a highly accurate model for breast cancer classification achieved through the hyperparameter tuning of the Random Forest classifier. However, it is imperative to note that further research is warranted to explore the deployment and practical application of the developed model. Subsequent investigations could delve into the incorporation of images and deep learning methodologies to broaden the scope and comprehensiveness of the study.

5. Data Availability

The data utilized in this study can be obtained by reaching out to the corresponding author upon request.

Reference

- Abonyi, Janos, and Ferenc Szeifert. 2003. "Supervised Fuzzy Clustering for the Identification of Fuzzy Classifiers." *Pattern Recognition Letters* 24(14):2195–2207. doi: 10.1016/S0167-8655(03)00047-3.
- Amrane, Meriem, Saliha Oukid, Ikram Gagaoua, and Tolga Ensari. 2018. "Breast Cancer Classification Using Machine Learning." 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) 1–4. doi: 10.1109/EBBT.2018.8391453.
- Anon. 2023. "Machine Learning | Nature Medicine." Retrieved December 21, 2023 (<https://www.nature.com/subjects/machine-learning/nm>).
- Aruna, S., Rajagopalan S P, and L. Nandakishore. 2011. "Knowledge Based Analysis of Various Statistical Tools in Detecting Breast Cancer." *Computer Science & Information Technology* 2. doi: 10.5121/csit.2011.1205.
- Assiri, Adel S., Saima Nazir, and Sergio A. Velastin. 2020. "Breast Tumor Classification Using an Ensemble Machine Learning Method." *Journal of Imaging* 6(6):39. doi: 10.3390/jimaging6060039.
- Belgiu, Mariana, and Lucian Drăguț. 2016. "Random Forest in Remote Sensing: A Review of Applications and Future Directions." *ISPRS Journal of Photogrammetry and Remote Sensing* 114:24–31. doi: 10.1016/j.isprsjprs.2016.01.011.
- Bhardwaj, Arpit, Harshit Bhardwaj, Aditi Sakalle, Ziya Uddin, Maneesha Sakalle, and Wubshet Ibrahim. 2022. "Tree-Based and Machine Learning Algorithm Analysis for Breast Cancer Classification." *Computational Intelligence and Neuroscience* 2022:6715406. doi: 10.1155/2022/6715406.
- Biau, Gérard, and Erwan Scornet. 2016. "A Random Forest Guided Tour." *TEST* 25(2):197–227. doi: 10.1007/s11749-016-0481-7.
- Chen, Hua, Nan Wang, Xueping Du, Kehui Mei, Yuan Zhou, and Guangxing Cai. 2023. "Classification Prediction of Breast Cancer Based on Machine Learning." *Computational Intelligence and Neuroscience* 2023:6530719. doi: 10.1155/2023/6530719.
- Delen, Dursun, Glenn Walker, and Amit Kadam. 2005. "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods." *Artificial Intelligence in Medicine* 34(2):113–27. doi: 10.1016/j.artmed.2004.07.002.
- Doddipalli, Lavanya, and K. Rani. 2011. "Analysis of Feature Selection with Classification: Breast Cancer Datasets." *Indian Journal of Computer Science and Engineering (IJCSE)* 2:756–63.
- Hamilton, Howard, Ning Shan, and Nick Cercone. 1996. "RIAC: A Rule Induction Algorithm Based on Approximate Classification." *Tech. Report 96-6*, Department of Computer Science, University of Regina.
- Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. 2017. "Artificial Intelligence in Healthcare: Past, Present and Future." *Stroke and Vascular Neurology* 2(4):230–43. doi: 10.1136/svn-2017-000101.

Liu, Ya-Qin, Cheng Wang, and Lu Zhang. 2009. "Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data." Pp. 1–4 in 2009 3rd International Conference on Bioinformatics and Biomedical Engineering.

Nauck, D., and R. Kruse. 1999. "Obtaining Interpretable Fuzzy Classification Rules from Medical Data." *Artificial Intelligence in Medicine* 16(2):149–69. doi: 10.1016/s0933-3657(98)00070-0.

Ozcan, Irem, Hakan Aydin, and Ali Cetinkaya. 2022. "Comparison of Classification Success Rates of Different Machine Learning Algorithms in the Diagnosis of Breast Cancer." *Asian Pacific Journal of Cancer Prevention: APJCP* 23(10):3287–97. doi: 10.31557/APJCP.2022.23.10.3287.

Quinlan, J. Ross. 1996. "Improved Use of Continuous Attributes in C4.5." *J. Artif. Intell. Res.* 4:77–90. doi: 10.1613/JAIR.279.

Rajaguru, Harikumar, and Sannasi Chakravarthy S R. 2019. "Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer." *Asian Pacific Journal of Cancer Prevention: APJCP* 20(12):3777–81. doi: 10.31557/APJCP.2019.20.12.3777.

Rautela, Kamakshi, Dinesh Kumar, and Vijay Kumar. 2022. "A Systematic Review on Breast Cancer Detection Using Deep Learning Techniques." *Archives of Computational Methods in Engineering* 29(7):4599–4629. doi: 10.1007/s11831-022-09744-5.

Tabari, Azadeh, Shin Mei Chan, Omar Mustafa Fathy Omar, Shams I. Iqbal, Michael S. Gee, and Dania Daye. 2022. "Role of Machine Learning in Precision Oncology: Applications in Gastrointestinal Cancers." *Cancers* 15(1):63. doi: 10.3390/cancers15010063.

Uzun Ozsahin, Dilber, Declan Ikechukwu Emegano, Berna Uzun, and Ilker Ozsahin. 2022. "The Systematic Review of Artificial Intelligence Applications in Breast Cancer Diagnosis." *Diagnostics (Basel, Switzerland)* 13(1):45. doi: 10.3390/diagnostics13010045.

Xue, Bing, Mengjie Zhang, and Will N. Browne. 2014. "Particle Swarm Optimisation for Feature Selection in Classification: Novel Initialisation and Updating Mechanisms." *Applied Soft Computing* 18:261–76. doi: 10.1016/j.asoc.2013.09.018.

Yedjou, Clement G., Solange S. Tchounwou, Richard A. Aló, Rashid Elhag, BereKet Mochona, and Lekan Latinwo. 2021. "Application of Machine Learning Algorithms in Breast Cancer Diagnosis and Classification." *International Journal of Science Academic Research* 2(1):3081–86.

