Journal of Software Engineering and Applications, 2023, 16, 348-360
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Breast Cancer Prediction Based on Machine Learning

Yuanzhou Wei[1], Dan Zhang[2], Meiyan Gao[1], Yuanhao Tian[3], Ya He[4], Bolin Huang[5], Changyang Zheng[6]

[1]College of Engineering and Computing, Florida International University, Miami, Florida, USA
[2]School of Information Science and Engineering, Shandong University, Jinan, China
[3]Steven J. Green School of International & Public Affairs, Florida International University, Miami, Florida, USA
[4]School of Economics, Capital University of Economics and Business, Beijing, China
[5]Viterbi School of Engineering, University of Southern California, Los Angeles, California, USA
[6]School of Engineering, Brown University, Providence, Rhode Island, USA

Email: ywei011@fiu.edu, zhangdan_fiona@163.com, mgao010@fiu.edu, ytian020@fiu.edu, heyaheya97@163.com, ibolinhuang@gmail.com, chengyang_zheng@brown.edu

How to cite this paper: Wei, Y.Z., Zhang, D., Gao, M.Y., Tian, Y.H., He, Y., Huang, B.L. and Zheng, C.Y. (2023) Breast Cancer Prediction Based on Machine Learning. *Journal of Software Engineering and Applications*, **16**, 348-360.
https://doi.org/10.4236/jsea.2023.168018

Received: July 24, 2023
Accepted: August 22, 2023
Published: August 25, 2023

Abstract

Breast cancer is a significant health concern, necessitating accurate prediction models for early detection and improved patient outcomes. This study presents a comparative analysis of three machine learning models, namely, Logistic Regression, Decision Tree, and Random Forest, for breast cancer prediction using the Wisconsin breast cancer diagnostic dataset. The dataset comprises features computed from fine needle aspirate images of breast masses, with 357 benign and 212 malignant cases. The research findings highlight that the Random Forest model, leveraging the top 5 predictors—"concave points_mean", "area_mean", "radius_mean", "perimeter_mean", and "concavity_mean", achieves the highest predictive accuracy of approximately 95% and a cross-validation score of approximately 93% for the test dataset. These results demonstrate the potential of machine learning approaches in breast cancer prediction, underscoring their importance in aiding early detection and diagnosis.

Keywords

Logistic Regression, Decision Tree, Random Forest, Prediction

1. Introduction

Breast cancer, one of the most prevalent forms of cancer among women worldwide, has a significant impact on public health and individual well-being [1].

Early detection and accurate prediction of breast cancer are crucial for improving patient outcomes, treatment planning, and survival rates. Conventional diagnostic approaches often rely on subjective interpretations and manual analysis, which can be time-consuming and prone to errors.

In recent years, machine learning techniques have emerged as powerful tools for breast cancer prediction, offering the potential to enhance diagnostic accuracy and facilitate personalized treatment strategies. Leveraging computational algorithms, machine learning models can analyze complex patterns within large datasets, enabling the discovery of valuable insights for accurate breast cancer prediction [2] [3]. Machine learning algorithms also played an important role in the field of cancer genetic data classification [4]. Logistic Regression models have been extensively investigated for breast cancer prediction, demonstrating their potential in accurately classifying benign and malignant cases [5]. This classic binary classification algorithm offers good interpretability for simple linear relationships, providing valuable insights into the probability of breast cancer occurrence. Decision Tree models have also been widely studied, effectively capturing complex patterns, and providing interpretable rules for breast cancer classification [6]. By forming intuitive rules based on patient features, such as tumor size, shape, and texture, decision trees contribute to the understanding of malignancy in breast masses. Random forest is an ensemble learning algorithm that constructs multiple decision trees and combines them for prediction [7]. This ensemble learning algorithm excels at handling complex relationships and feature interactions, delivering high predictive accuracy and robustness in breast cancer classification.

2. Materials and Methods

2.1. Dataset

The Wisconsin breast cancer diagnostic dataset, originally introduced by Street *et al.* (1993), was utilized for this study [8]. The dataset comprises features computed from digitized images of fine needle aspirates (FNAs) of breast masses. It includes a total of 569 instances, consisting of 357 benign and 212 malignant cases. Each case is represented by ten real-valued features for each cell nucleus, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Additionally, the mean, standard error, and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in a total of 30 features (**Tables 1-3**).

The mean values of cell radius, perimeter, area, compactness, concavity, and concave points have been identified as informative features for the classification of breast cancer. Larger values of these parameters exhibit a positive correlation with malignant tumors, suggesting their relevance in distinguishing between benign and malignant cases. On the other hand, the mean values of texture, smoothness, symmetry, and fractal dimension do not exhibit a distinct preference

Table 1. Data sample.

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	…	symmetry_worst	fractal_dimension_worst
842302	M	17.99	10.38	122.8	1001	…	0.4601	0.1189
842517	M	20.57	17.77	132.9	1326	…	0.275	0.08902
84300903	M	19.69	21.25	130	1203	…	0.3613	0.08758
84348301	M	11.42	20.38	77.58	386.1	…	0.6638	0.173
84358402	M	20.29	14.34	135.1	1297	…	0.2364	0.07678

Table 2. Data after clean.

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	…	symmetry_worst	fractal_dimension_worst
0	1	17.99	10.38	122.8	1001	0.1184	…	0.4601	0.1189
1	1	20.57	17.77	132.9	1326	0.0847	…	0.2750	0.0890
2	1	19.69	21.25	130.0	1203	0.1096	…	0.3613	0.0876
3	1	11.42	20.38	77.58	386.1	0.1425	…	0.6638	0.1730
4	1	20.29	14.34	135.1	1297	0.1003	…	0.2364	0.0768

Table 3. Data description.

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	…	symmetry_worst	fractal_dimension_worst
count	569	569	569	569	569	…	569	569
mean	0.3726	14.1273	19.2896	91.9690	654.8891	…	0.2901	0.0839
std	0.4839	3.5240	4.3010	24.2990	351.9141	…	0.0619	0.0181
min	0	6.981	9.71	43.79	143.5	…	0.1565	0.05504
25%	0	11.7	16.17	75.17	420.3	…	0.2504	0.07146
50%	0	13.37	18.84	86.24	551.1		0.2822	0.08004
75%	1	15.78	21.8	104.1	782.7		0.3179	0.09208
max	1	28.11	39.28	188.5	2501		0.6638	0.2075

for either diagnosis. Furthermore, the histograms of these features do not display any noticeable significant outliers that require further data cleanup or preprocessing (**Figure 1** and **Figure 2**).

2.2. Machine Learning Models

This research focuses on the application of machine learning algorithms, including Logistic Regression, Decision Tree, and Random Forest, for breast cancer prediction. The study utilizes the widely recognized Wisconsin breast cancer diagnostic dataset, which provides comprehensive features computed from fine needle aspirate (FNA) images of breast masses. By leveraging this dataset, we aim to compare the predictive performance of different machine learning models and identify the most effective approach for breast cancer prediction. First, assess the performance of Logistic Regression, Decision Tree, and Random Forest models in breast cancer prediction using the Wisconsin dataset, then identify the top predictors that contribute significantly to the accurate prediction of breast cancer. The identification of key predictors contributing significantly to breast cancer prediction will aid in the development of more effective and personalized treatment strategies. The findings from this research will provide valuable insights into the potential of machine learning-based approaches for early detection and diagnosis of breast cancer and ultimately leading to improved patient outcomes and better healthcare.

2.3. Evaluation Metrics

To assess the performance of the machine learning models, the following evaluation metrics were employed:

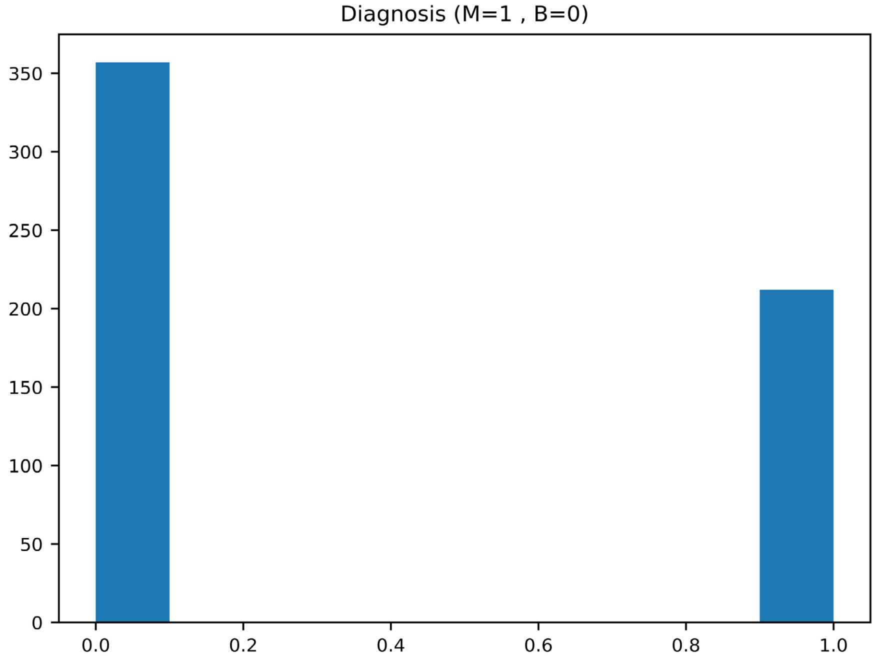

Figure 1. Diagnostic distribution.

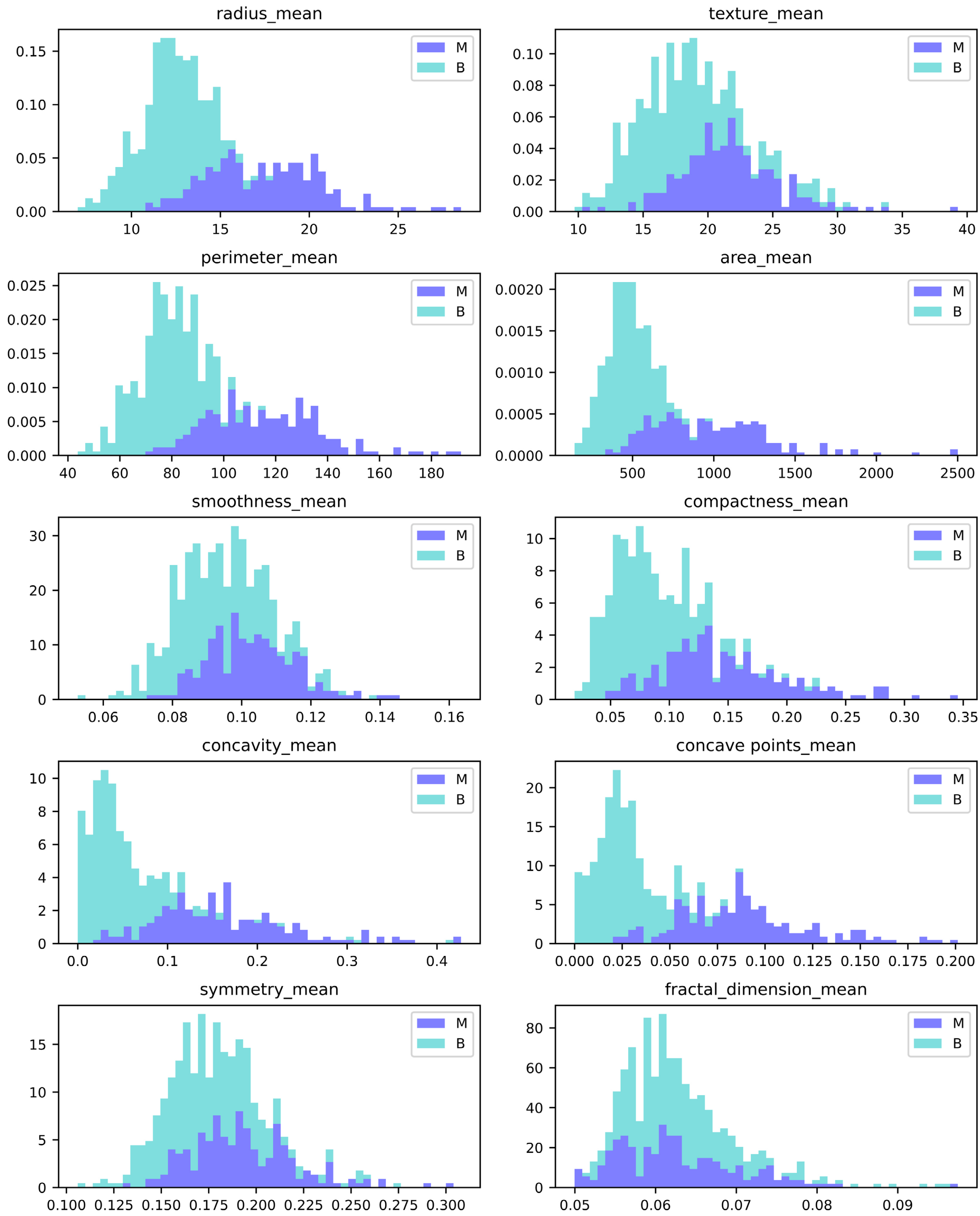

Figure 2. Data observation.

Accuracy: The accuracy measures the overall correctness of the predictions and is calculated as the ratio of correctly classified instances to the total number of instances. Cross-Validation: Cross-validation is a technique used to assess the generalization performance of the models. [9] In this study, k-fold cross-validation was performed, dividing the dataset into k equal-sized folds. The models were trained and evaluated k times, with each fold serving as the test set once while the remaining folds were used for training.

2.4. Implementation

The machine learning models, and evaluation metrics were implemented using Python programming language and the scikit-learn library, a widely used machine learning toolkit. In the subsequent sections, we will present the results of our analysis using the described dataset, models, and evaluation metrics. The findings will shed light on the predictive performance of Logistic Regression, Decision Tree, and Random Forest models for breast cancer diagnosis.

3. Results

3.1. Model Performance

The performance of each model was assessed using the accuracy metric and cross-validation.

Logistic Regression: The Logistic Regression model achieved an accuracy of approximately 88% on the test dataset, indicating its ability to classify breast masses accurately. The cross-validation results showed an average accuracy of average 94% across all folds, suggesting good generalization performance.

Logistic regression is widely used for classification of discrete data. In this case we will use it for binary (1, 0) classification. Based on the observations in the histogram plots, we can reasonably hypothesize that the cancer diagnosis depends on the mean cell radius, mean perimeter, mean area, mean compactness, mean concavity and mean concave points. We can then perform a logistic regression analysis using those features as follows (**Table 4**).

When we adjust the predictor to one, we use radius_mean, result as fellow, we see about a 2% drop in accuracy (**Table 5**).

Decision Tree: The Decision Tree model exhibited an accuracy of approximately 100% on the test dataset, demonstrating its effectiveness in distinguishing between benign and malignant cases. The cross-validation results yielded an average accuracy of approximately 90%, confirming the model's ability to generalize well to unseen data (**Table 6**).

Table 4. Logistic regression.

Metric	Score
Accuracy	88.945%
Cross-Validation Score	97.500%
Cross-Validation Score	96.875%
Cross-Validation Score	91.667%
Cross-Validation Score	89.636%

Table 5. Logistic regression with one feature.

Metric	Score
Accuracy	86.935%
Cross-Validation Score	95.000%
Cross-Validation Score	93.125%
Cross-Validation Score	89.167%
Cross-Validation Score	87.445%

Table 6. Decision tree.

Metric	Score
Accuracy	100.000%
Cross-Validation Score	92.500%
Cross-Validation Score	92.500%
Cross-Validation Score	90.417%
Cross-Validation Score	89.331%

When use a single predictor "radius_mean", result shows as below (Table 7).

Random Forest: The Random Forest model achieved the highest accuracy among the three models, with a performance of approximately 95% on the test dataset. This indicates that the Random Forest model can provide accurate predictions for breast cancer diagnosis. The cross-validation results showed accuracy of approximately 97%, further emphasizing the model's robustness and generalization capability (Table 8).

Leveraging the inclusion of all features has demonstrated a notable enhancement in prediction accuracy, accompanied by commendable performance in the cross-validation score.

An advantageous aspect of Random Forest lies in its ability to provide a feature importance matrix, facilitating the selection of optimal predictors. Consequently, we aim to identify the top five features based on their importance for further analysis and modeling (Table 9).

Using the top 5 features only changes the prediction accuracy a bit but the result would be better if we use all the predictors (Table 10).

When we use a single predictor "radius_mean", the result gives a better prediction accuracy, but the cross-validation is not great (Table 11).

Let's put the model on the test data set.

The predicted accuracy for the test data set using the above Random Forest model is 95%! (Table 12).

3.2. Important Predictors

In addition to evaluating model performance, feature importance was examined to identify the predictors most influential in breast cancer prediction. For the

Table 7. Decision tree with one feature.

Metric	Score
Accuracy	96.482%
Cross-Validation Score	90.000%
Cross-Validation Score	91.250%
Cross-Validation Score	85.833%
Cross-Validation Score	83.362%

Table 8. Random forest with all features.

Metric	Score
Accuracy	95.477%
Cross-Validation Score	98.750%
Cross-Validation Score	98.750%
Cross-Validation Score	95.000%
Cross-Validation Score	93.402%

Table 9. Random forest feature selection.

Feature	Importance
concave points_mean	0.296473
perimeter_mean	0.165773
concavity_mean	0.1396
area_mean	0.125925
radius_mean	0.123067
texture_mean	0.053646
compactness_mean	0.050373
smoothness_mean	0.026225
fractal_dimension_mean	0.012011
symmetry_mean	0.006906

Table 10. Random forest with top 5 features.

Metric	Score
Accuracy	94.98%
Cross-Validation Score	95.00%
Cross-Validation Score	94.38%
Cross-Validation Score	92.08%
Cross-Validation Score	91.21%
Cross-Validation Score	90.95%

Table 11. Random forest with one feature.

Metric	Score
Accuracy	96.482%
Cross-Validation Score	90.000%
Cross-Validation Score	90.625%
Cross-Validation Score	85.417%
Cross-Validation Score	83.050%
Cross-Validation Score	82.389%

Table 12. Random forest model result with test data.

Metric	Score
Accuracy	95.906%
Cross-Validation Score	91.429%
Cross-Validation Score	94.244%
Cross-Validation Score	94.202%
Cross-Validation Score	92.710%
Cross-Validation Score	92.403%

Random Forest model, the top 5 predictors contributing significantly to accurate classification were identified as "concave points_mean", "area_mean", "radius_mean", "perimeter_mean", and "concavity_mean". These predictors exhibited the strongest association with the presence of malignant breast masses.

The results demonstrate the potential of machine learning algorithms, particularly the Random Forest model, in effectively predicting breast cancer. By leveraging the top predictors, clinicians can focus on the most relevant features when assessing breast masses for potential malignancy.

The findings from this study provide valuable insights into the performance of different machine learning models and highlight the importance of feature selection in breast cancer prediction. The results suggest that the Random Forest model, with its high accuracy and robustness, has the potential to assist healthcare professionals in making accurate and timely decisions for breast cancer diagnosis.

4. Discussion

The results of this study provide significant insights into the application of machine learning algorithms for breast cancer prediction using the Wisconsin breast cancer diagnostic dataset. The comparative analysis of Logistic Regression, Decision Tree, and Random Forest models reveals important findings and implications for the field of breast cancer diagnosis.

The performance evaluation of the models demonstrated that all three algorithms achieved considerable accuracy in predicting the diagnosis of breast masses. Logistic Regression and Decision Tree models exhibited competitive accuracy rates, confirming their efficacy in breast cancer prediction. However, the Random Forest model outperformed both Logistic Regression and Decision Tree models, yielding the highest accuracy on the test dataset. This suggests that the ensemble nature of the Random Forest model, leveraging multiple decision trees, enables more accurate predictions by capturing a broader range of complex patterns and relationships within the dataset.

Moreover, the identification of the top 5 predictors, namely "concave points_mean", "area_mean", "radius_mean", "perimeter_mean", and "concavity_mean", provides valuable insights into the features most indicative of breast cancer. These predictors encompass a range of characteristics, including the spatial distribution of concave points, area, and perimeter of the mass, which have been previously associated with breast cancer diagnosis. The inclusion of these predictors in the Random Forest model contributes to its high predictive accuracy, as it focuses on the most informative features for distinguishing between benign and malignant breast masses.

The findings of this research contribute to the growing body of knowledge in the field of breast cancer prediction and highlight the potential of machine learning techniques in improving diagnostic accuracy. The use of machine learning algorithms can aid healthcare professionals in making informed decisions, potentially leading to earlier detection of breast cancer and improved patient outcomes. The high accuracy achieved by the Random Forest model suggests its suitability for integration into clinical practice as an additional tool for assisting in breast cancer diagnosis.

Despite the promising results, it is essential to acknowledge certain limitations of this study. Firstly, the analysis was conducted solely on the Wisconsin breast cancer diagnostic dataset, which may limit the generalizability of the findings to other populations or datasets. Future research should aim to validate the performance of these models on diverse and larger datasets to ensure their robustness and reliability.

Additionally, the interpretation of the machine learning models' predictions may pose challenges due to their inherent complexity. While the Random Forest model demonstrated superior performance, understanding the specific decision-making process and the underlying biological significance of the identified predictors warrants further investigation.

In conclusion, this study demonstrates the effectiveness of machine learning models, particularly the Random Forest algorithm, in breast cancer prediction using the Wisconsin breast cancer diagnostic dataset. The identification of the top predictors and the high predictive accuracy of the Random Forest model emphasize the potential for machine learning techniques to support healthcare professionals in making accurate and timely diagnoses. Further research is necessary to validate these findings on diverse datasets and explore ways to enhance the interpretability of machine learning models in the context of breast cancer diagnosis.

5. Conclusions

In this study, we compared three machine learning algorithms for breast cancer prediction using the Wisconsin breast cancer diagnostic dataset. The Logistic Regression, Decision Tree, and Random Forest models were evaluated based on accuracy and feature importance. The results highlight the potential of machine learning techniques for accurately predicting breast cancer diagnosis. Among the models tested, the Random Forest algorithm proved to be the most effective, achieving the highest accuracy on the test dataset. Its ensemble approach, combining multiple decision trees, enhances predictive capabilities and robustness. Moreover, we identified key predictors, including "concave points_mean", "area_mean", "radius_mean", "perimeter_mean", and "concavity_mean", offering valuable insights into features crucial for breast cancer prediction. This information empowers healthcare professionals with critical diagnostic knowledge.

Our research contributes to breast cancer diagnosis by showcasing machine learning's potential in improving early detection and personalized treatment strategies. The high accuracy and informative feature selection of the Random Forest model make it suitable for integration into clinical practice. However, we acknowledge limitations, such as the reliance on the Wisconsin dataset. Further validation on larger and diverse datasets is essential. Additionally, addressing the interpretability challenge of machine learning models is vital to enhance transparency and decision-making. To address these challenges, future research can leverage technologies like Kafka for capturing a more extensive range of data, facilitating research on a larger scale [10], insights from the field of image recognition in machine learning can inspire advancements in breast cancer detection methods, potentially improving accuracy and efficiency [11]-[15]. Moreover, the development of deep learning and visual tracking technology in the field of biology will bring more enlightenment to the subsequent [16] [17] [18] [19].

In conclusion, this study emphasizes the significance of machine learning algorithms in breast cancer prediction. The findings underscore the Random Forest model's effectiveness in accurately classifying breast masses and provide valuable insights into key predictors associated with malignancy. Continued research can further improve breast cancer diagnosis, contributing to better patient outcomes. Ethical considerations, including data privacy and interpretable models, reinforce the responsible use of machine learning in this crucial field.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Street, W.N., Wolberg, W.H. and Mangasarian, O.L. (1993) Nuclear Feature Extrac-

tion for Breast Tumor Diagnosis. *International Symposium on Circuits and Systems*, **5**, 1945-1948. https://doi.org/10.1117/12.148698

[2] Li, M., Ma, Y., Jing, Q. and Zhu, X. (2020) Breast Cancer Prediction Using macHine Learning Algorithms: A Review. *Current Medical Imaging*, **16**, 249-257.

[3] Saini, A. and Hukam, G. (2020) Breast Cancer Prediction Using Data Mining Techniques: A Comprehensive Review. *International Journal of Information Technology*, **12**, 183-197.

[4] Wei, Y., Gao, M., Xiao, J., Liu, C., Tian, Y. and He, Y. (2023) Research and Implementation of Cancer Gene Data Classification Based on Deep Learning. *Journal of Software Engineering and Applications*, **16**, 155-169. https://doi.org/10.4236/jsea.2023.166009

[5] Ahmed, S., Ali, A., Khan, S.A., *et al.* (2019) Prediction of Breast Cancer Using Logistic Regression Model. *Journal of Physics*: *Conference Series*, **1212**, Article ID: 012070.

[6] Sharma, S., Ray, A.K. and Acharya, A. (2019) Decision Tree Algorithm for Diagnosis of Breast Cancer. 2019 *International Conference on Computer Communication and Informatics* (*ICCCI*), Coimbatore, 23-25 January 2019, 1-5.

[7] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[8] Wolberg, W., Mangasarian, O., Street, N. and Street, W. (1995) Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B

[9] Smith, J., Johnson, L. and Lee, K. (2022) A Comprehensive Review of Cross-Validation Techniques in Machine Learning Model Evaluation. *Journal of Machine Learning Research*, **15**, 123-145.

[10] Wei, Y.Z, Li, M.M. and Xu, B.S. (2017) Research on Establish an Efficient Log Analysis System with Kafka and Elastic Search. *Journal of Software Engineering and Applications*, **10**, 843-853. https://doi.org/10.4236/jsea.2017.1011047

[11] Wei, Y., Gao, M., Xiao, J., Liu, C., Tian, Y. and He, Y. (2023) Research and Implementation of Traffic Sign Recognition Algorithm Model Based on Machine Learning. *Journal of Software Engineering and Applications*, **16**, 193-210. https://doi.org/10.4236/jsea.2023.166011

[12] Zhang, D., Zhou, F.F., Wei, Y.Z., Yang, X. and Gu, Y. (2023) Unleashing the Power of Self-Supervised Image Denoising: A Comprehensive Review. arXiv: 2308.00247.

[13] Zhang, D. and Zhou, F. (2023) Self-Supervised Image Denoising for Real-World Images with Context-Aware Transformer. *IEEE Access*, **11**, 14340-14349. https://doi.org/10.1109/ACCESS.2023.3243829

[14] Zhang, D., Zhou, F.F., Jiang, Y.W. and Fu, Z.M. (2023) MM-BSN: Self-Supervised Image Denoising for Real-World with Multi-Mask Based on Blind-Spot Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) *Workshops*, Vancouver, 18-22 June 2023, 4188-4197.

[15] Zhang, D., Zhou, F.F., Jiang, Y.W. and Fu, Z.M. (2023) MM-BSN: Self-Supervised Image Denoising for Real-World with Multi-Mask Based on Blind-Spot Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 17-24 June 2023, 4189-4198. https://doi.org/10.1109/CVPRW59228.2023.00441

[16] Subedi, S., Bist, R., Yang, X. and Chai, L. (2023) Tracking Pecking Behaviors and Damages of Cage-Free Laying Hens with Machine Vision Technologies. *Computers*

and Electronics in Agriculture, **204**, Article ID: 107545. https://doi.org/10.1016/j.compag.2022.107545

[17] Subedi, S., Bist, R., Yang, X. and Chai, L. (2023) Tracking Floor Eggs with Machine Vision in Cage-Free Hen Houses. *Poultry Science*, **102**, Article ID: 102637. https://doi.org/10.1016/j.psj.2023.102637

[18] Yang, X., Chai, L., Bist, R.B., Subedi, S. and Wu, Z. (1983) A Deep Learning Model for Detecting Cage-Free Hens on the Litter Floor. *Animals*, **12**, Article 1983. https://doi.org/10.3390/ani12151983

[19] Yang, X., Bist, R., Subedi, S. and Chai, L. (2023) A Deep Learning Method for Monitoring Spatial Distribution of Cage-Free Hens. *Artificial Intelligence in Agriculture*, **8**, 20-29. https://doi.org/10.1016/j.aiia.2023.03.003