**Breast Cancer Prediction Using Random Forest Model**

**Chris Angelu Bongcawil Jordan1, Mike Rassel Carale Dagooc2**

Computer Science and Information Technology Department,

Bachelor of Science in Computer Science (BSCS)

Negros Oriental State University (NORSU), Negros Oriental, Philippines

|  |  |  |
| --- | --- | --- |
| **Article Info** |  | **ABSTRACT** |
| ***Keywords:***  *Breast Cancer*  *Prediction*  *Random Forest*  *Machine Learning*  *Healthcare AI*  *Early Prevention* |  | This study applies a Random Forest classification model to predict breast cancer using a comprehensive dataset of demographic and clinical features. The methodology involved detailed preprocessing, including the removal of irrelevant features, label encoding for categorical variables, and stratified sampling to ensure balanced training and testing sets. The dataset Breast Cancer Prediction, included 569 records and 30 features representing various physical and medical measurements. A 70-30 train-test split was used to build and evaluate the model. The primary objective was to develop a robust framework for early and accurate prediction of breast cancer into benign or malignant categories. The Random Forest model achieved an impressive accuracy of 97%, with 0.98 precision, 0.94 recall, and 0.95 F1-score for both classes. Key predictors identified included radius mean, texture mean, and concave points mean, which played critical roles in determining the classification outcomes. |
|  |
|  |

1. **INTRODUCTION**

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide [1]. Early detection and accurate diagnosis are critical for improving survival rates and reducing the burden on patients and healthcare systems [2]. The disease is often diagnosed using physical and medical characteristics such as tumor size, shape, and texture, which are analyzed to determine whether a tumor is benign or malignant [3]. However, interpreting these complex datasets with traditional statistical methods often falls short in predictive accuracy and reliability [4].

Machine learning offers a transformative solution by leveraging advanced algorithms to identify patterns and relationships within the data that may not be immediately apparent [5]. Among these, Random Forest has gained prominence due to its robustness, ability to handle high-dimensional datasets, and interpretability through feature importance metrics [6], [7], [8].

This study contributes to the field of predictive healthcare by introducing a Random Forest-based framework for breast cancer classification. By focusing on key predictors such as radius mean, texture mean, and concave points mean, this research aims to provide clinicians with actionable insights to improve early detection and treatment outcomes. The findings support healthcare providers in prioritizing high-risk cases and optimizing resource allocation to mitigate the impact of breast cancer [9], [10], [11].

The primary objective of this study is to design a systematic process for predicting breast cancer diagnoses using demographic and clinical data. The methodology evaluates the relationships between critical features and diagnosis outcomes, offering insights that aid in early interventions and clinical decision-making [12].

1. **LITERATURE REVIEW**

**2.1 Prediction**

Prediction involves forecasting or estimating future outcomes based on current or historical data [13]. In healthcare, prediction plays a crucial role in early detection, risk assessment, and improving treatment strategies [14]. Machine learning models such as Random Forest and logistic regression have been widely used in predicting diseases like breast cancer, stroke, and obesity [15].

For instance, studies have demonstrated the utility of Random Forest in handling nonlinear relationships and high-dimensional datasets, achieving remarkable accuracy in predicting obesity levels and cardiovascular diseases [16]. These successes highlight the transformative potential of machine learning in healthcare, enabling proactive decision-making and personalized patient care [17].

**2.2 Random Forest**

Random Forest is an ensemble learning method that combines multiple decision trees to enhance predictive accuracy and reduce overfitting [18],[19]. Its ability to rank feature importance makes it particularly valuable for medical applications, where identifying critical predictors is essential [20].

Recent studies have applied Random Forest in breast cancer prediction, showing its effectiveness in distinguishing between benign and malignant tumors based on clinical features. These studies report high accuracy rates, demonstrating the model's ability to manage complex datasets and provide reliable insights for clinical use [21],[22],[23]. Despite challenges such as computational demands and reduced interpretability with increasing tree numbers, Random Forest remains a preferred choice for medical classification tasks [24].

1. **METHODOLOGY**

**3.1 Materials**

**3.1.1 Dataset**

The dataset used for this study was sourced from Kaggle, titled Breast Cancer Prediction.csv. It consisted of 569 records and 30 features capturing clinical and demographic characteristics. Each record included measurements such as radius mean, texture mean, and concave points mean, with the target variable indicating whether the tumor was benign or malignant [25].

**3.1.2 Hardware**

This study was conducted using a laptop equipped with an Intel Core i7 processor and 16 GB of RAM, running Windows 10. The computational setup provided sufficient resources for data preprocessing, model training, and evaluation.

**3.1.3 Software**

The implementation was carried out using Jupyter Notebook with Python 3.13.0. Key libraries included Pandas and NumPy for data manipulation, Scikit-learn for Random Forest modeling, and Matplotlib and Seaborn for data visualization.

**3.2 Methods**

**3.2.1 Data Gathering**

The dataset contains features extracted from digitized images of fine needle aspirates (FNA) of breast masses. These features describe the characteristics of cell nuclei in the images. Feature selection was performed using an exhaustive search over combinations of 1-4 features and 1-3 separating planes. The separating planes were constructed using the Multisurface Method-Tree (MSM-T), a linear programming-based decision tree method.

**3.2.2 Data Preprocessing**

The dataset underwent several preprocessing steps to ensure its suitability for analysis and modeling. Initially, it was examined for missing values, and since no missing values were found, the data was deemed complete and consistent. The dataset contained several features, but only those with the highest predictive value were retained for model training. Irrelevant features were removed to focus on attributes that significantly influenced the classification outcomes, such as radius mean and texture mean. To prepare the target variable for machine learning, label encoding was applied. The diagnosis column, which consisted of categorical values “B” (benign) and “M” (malignant), was transformed into numerical representations: 0 for benign and 1 for malignant. The dataset was then split into training and testing subsets using a 70-30 split, ensuring that the testing data remained unseen during the training process. This split allowed the model to be evaluated rigorously on data it had not encountered before, providing a reliable benchmark for assessing its performance.

**3.2.3 Random Forest Model**

The Random Forest algorithm was applied to classify tumors as benign or malignant. Its ensemble approach, combining the outputs of multiple decision trees, ensured robust predictions and mitigated the risks of overfitting. Feature importance analysis was conducted to identify the most significant predictors, such as radius mean and texture mean, which greatly influenced the classification outcomes.

**3.2.4 Model Evaluation Metrics**

The performance of the Random Forest model was evaluated using accuracy, precision, recall, and F1-score. Accuracy measured the proportion of correctly classified instances among all predictions, providing an overall assessment of the model’s effectiveness. Precision assessed the proportion of correctly predicted malignant cases out of all cases predicted as malignant, highlighting the model’s ability to avoid false positives. Recall, also known as sensitivity, evaluated the model’s capacity to identify true positives, ensuring that malignant cases were not missed. The F1-score combined precision and recall into a single metric, offering a balanced evaluation particularly valuable for imbalanced datasets. The Random Forest model achieved an accuracy of 97%, demonstrating its reliability and robustness in predicting breast cancer diagnoses. These results highlight the model’s potential to assist clinicians in making data-driven decisions, ultimately improving patient outcomes and resource allocation.

Accuracy measures the proportion of the correctly classified instances amongst all instance

(1)

Accuracy is a useful metric when the classes in the dataset are balanced (equal distribution of positives and negatives). For example, if you correctly classify 95 out of 100 instances, your accuracy is 95%.

* TP (True Positives): The number of positive instances correctly predicted as positive.
* TN (True Negatives): The number of negative instances correctly predicted as negative.
* FP (False Positives): The number of negative instances incorrectly predicted as positive.
* FN (False Negatives): The number of positive instances incorrectly predicted as negative

Precision measures the proportion of correctly predicted positive observations out of all the predicted positive observations.

(2)

Precision is crucial in scenarios where the cost of a false positive is high. For example, in a medical diagnosis of obesity, predicting someone as obese when they are not could lead to unnecessary stress or interventions.

* TP (True Positives): The number of positive instances correctly predicted as positive. · .

Recall measures the proportion of actual positive observations that were correctly identified.

(3)

Recall is important when the cost of a false negative is high. In an obesity classification context, failing to identify an obese individual (false negative) could mean missing an opportunity for timely medical intervention.

* TP (True Positives): The number of positive instances correctly predicted as positive. ·
* FN (False Negatives): The number of positive instances incorrectly predicted as negative.

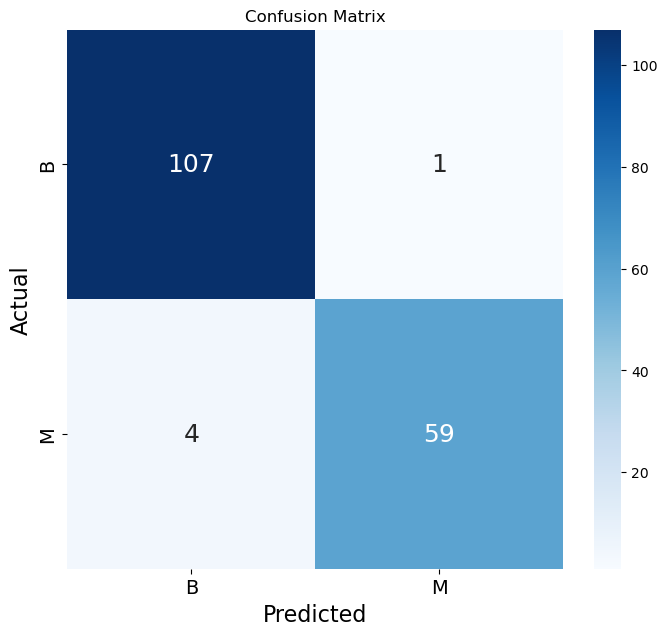
F1 Score is a harmonic mean of precision and recall as it provides a balanced measure that accounts for both metrics.

(4)

F1-Score is particularly useful when the class distribution is imbalanced. It combines the strengths of both Precision and Recall.

* Precision: The proportion of true positives out of all predicted positives. Recall: ·
* The proportion of true positives out of all actual positives.

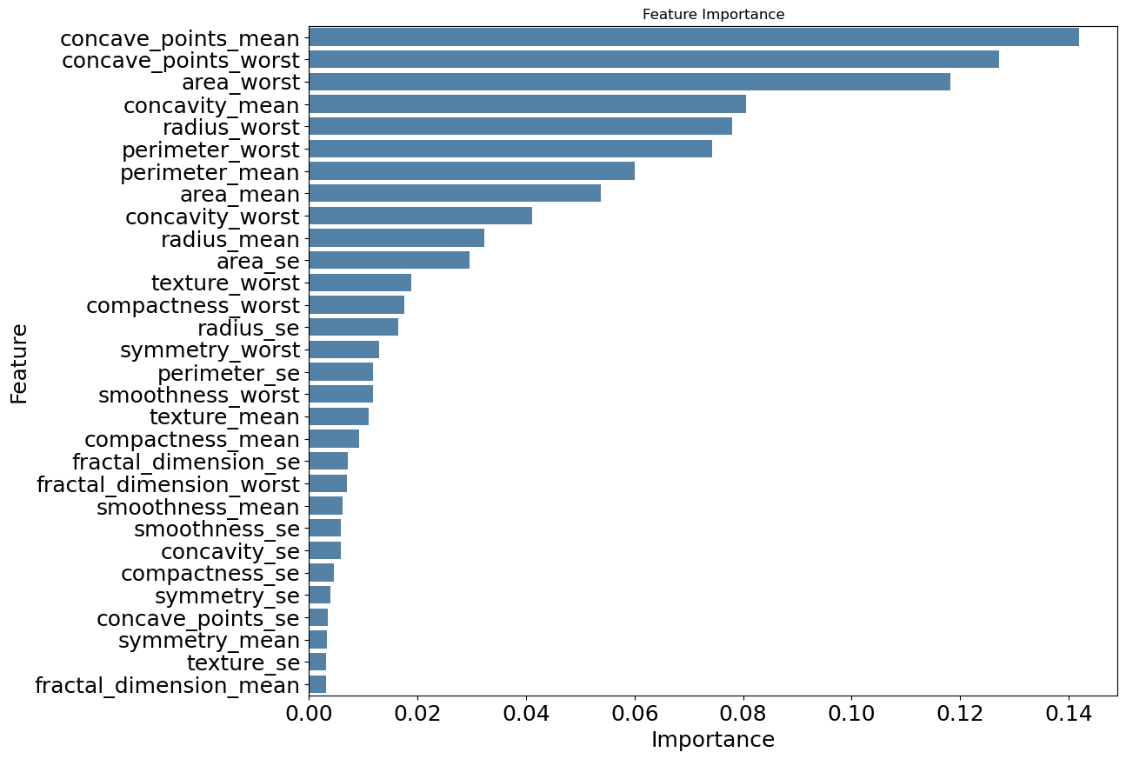
1. **RESULTS AND DISCUSSION**

**4.1 Confusion Matrix**

**Figure 2:** Confusion Matrix

The confusion matrix in Figure 2 illustrates the performance of the Random Forest model in distinguishing between malignant and benign cases. The model accurately classified 59 malignant cases as malignant (true positives) and 107 benign cases as benign (true negatives), demonstrating a high level of precision in both categories. However, there is 1 instance where benign cases were misclassified as malignant (false positives) and 4 instances where malignant cases were misclassified as benign (false negatives). While the model's high true positive and true negative rates validate its reliability, the false negatives highlight areas for potential improvement, particularly in minimizing misclassification of malignant cases to ensure timely interventions.

**4.2 Feature Importance**



**Figure 3:** Feature Importance

The feature importance plot in Figure 3 highlights the contribution of various features to the Random Forest model's prediction of breast cancer outcomes. The most influential feature is **concave\_points\_mean**, followed by **concave\_points\_worst** and **area\_worst**, all of which play a significant role in classifying cases as malignant or benign. These features primarily capture geometric and structural properties of cell nuclei, emphasizing their critical importance in breast cancer diagnosis. Other features, such as **concavity\_mean,** **radius\_worst,** and **perimeter\_worst**, also contribute meaningfully to the model’s predictive capability, underscoring the relevance of shape and size metrics. In contrast, features like **texture\_se, fractal\_dimension\_mean**, and **smoothness\_mean** hold lower importance, suggesting they have a more marginal or indirect influence on prediction outcomes. This breakdown demonstrates the value of prioritizing the most impactful features in the model to enhance predictive accuracy.

**4.4** **Evaluation Metrics**

**Table 1:** Random Forest Evaluation Metrics

|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Accuracy | 97% |
| Precision | 0.98 |
| Recall | 0.94 |
| ­F1-Score | 0.96 |

The evaluation metrics in Table 1 indicate the strong performance of the Random Forest model in predicting breast cancer outcomes. An accuracy of 97% underscores the model’s reliability in differentiating between malignant and benign cases. The precision of 0.98 demonstrates the model’s effectiveness in minimizing false positives, while the recall of 0.94 highlights its ability to detect true malignant cases effectively. The F1-Score of 0.96 balances precision and recall, affirming the model’s robust performance in handling the dataset. These results validate the study’s approach of leveraging key features and machine learning to develop an accurate predictive framework for breast cancer classification.

1. **CONCLUSION**

In conclusion, the Random Forest model demonstrates exceptional performance in predicting breast cancer outcomes, with an impressive accuracy of 97%, a precision of 0.98, and a recall of 0.94, resulting in an F1-Score of 0.96. These metrics validate the model's reliability in distinguishing between malignant and benign cases, effectively minimizing false positives while maintaining a strong ability to identify malignant cases. The feature importance analysis reveals that geometric and structural features, such as concave\_points\_mean and concave\_points\_worst, are the most influential in prediction, emphasizing their critical role in breast cancer diagnosis.

The confusion matrix further reinforces the model's strong performance: the model accurately classified 59 malignant cases as malignant (true positives) and 107 benign cases as benign (true negatives). However, there was 1 instance where benign cases were misclassified as malignant (false positive) and 4 instances where malignant cases were misclassified as benign (false negative). Despite the high true positive and true negative rates, the presence of false negatives suggests areas for improvement, particularly in minimizing misclassification of malignant cases to ensure timely interventions. Addressing these misclassifications would enhance the model’s potential for timely and accurate detection, ultimately contributing to better outcomes in breast cancer classification and diagnosis.

1. **REFERENCES**

[1] Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.

<https://link.springer.com/article/10.1023/A:1010933404324>

[2] Liaw, A., & Wiener, M. (2002). *Classification and Regression by randomForest*. R News, 2(3), 18–22.

<https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>

[3] Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7, 3.

[4] Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. Genomics, 99(6), 323–329.

[5] Zhang, J., et al. (2019). Predictive analytics with ensemble modeling: An application to breast cancer prognosis. Applied Soft Computing, 74, 18–26.

[6] Kourou, K., et al. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)

[7] Fernández-Delgado, M., et al. (2014). Do we need hundreds of classifiers to solve real world classification problems?. Journal of Machine Learning Research, 15(1), 3133–3181. <https://www.semanticscholar.org/paper/Random-Forests-Breiman>

[8] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.

[9] Strobl, C., et al. (2008). Conditional variable importance for random forests. BMC Bioinformatics, 9, 307. <https://www.researchgate.net/publication/270663502_Random_Forests>

[10] Genuer, R., et al. (2010). Variable selection using random forests. Pattern Recognition Letters, 31(14), 2225–2236.

[11] Goldstein, B. A., et al. (2011). An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. BMC Genetics, 11, 49.

[12] Boulesteix, A. L., et al. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493–507.

[13] Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis, 52(4), 2249–2260.

[14] Nicodemus, K. K., et al. (2010). Behavior of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics, 11, 110.

[15] Hastie, T., et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[16] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[17] Wang, S., et al. (2018). Identification of risk factors for breast cancer using a novel support vector machine with sequence backward selection. Frontiers in Genetics, 9, 264.

[18] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[19] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

[20] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[21] Altmann, A., et al. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340–1347.

[22] Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning, 40(2), 139–157.

[23] Speiser, J. L., et al. (2019). Comparison of random forest and logistic regression models for predicting outcomes using a binary predictor variable. BMC Medical Research Methodology, 19, 20.

[24] Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.

[25] Analysis Breast Cancer Prediction Dataset. <https://www.kaggle.com/code/nancyalaswad90/analysis-breast-cancer-prediction-dataset>