JOURNAL OF
EMERGING INVESTIGATORS

Article

Applying machine learning to breast cancer diagnosis: A high school student's exploration using R

Mithra Vikram[1], Saigopal Sathyamurthy[2]
[1]Fulton Science Academy, 3035 Fanfare Way, Alpharetta, Georgia
[2]Qure.ai, Bangalore, India

SUMMARY

Early diagnosis of breast cancer is critical for improved prognosis. However, current diagnostic methods, like mammograms, are expensive and not widely available in resource-constrained regions. This study aims to identify alternative diagnostic methods that are more accessible. We hypothesize that features obtained using Fine Needle Aspiration Biopsy (FNAB) can serve as predictive variables in machine learning (ML) algorithms for accurate breast cancer detection. Utilizing the Wisconsin Breast Cancer Dataset (WBCD), we conducted statistical analyses to explore different machine-learning models for classifying tumors as malignant or benign. Initial univariate analysis revealed that certain features were highly correlated with the malignancy of the tumor. We created a second dataset by removing the correlated variables and evaluated various machine learning models using both datasets on their ability to classify tumors, measuring performance by sensitivity, specificity, and accuracy. Among the models tested, logistic regression and random forest classifiers delivered standout results. While the random forest classifier with the full variable dataset and logistic regression with the principal component analysis (PCA) reduced variable dataset achieved the highest accuracy, the overall difference in performance of these two models across the datasets was minimal. These results demonstrate that using a smaller dataset enables models to predict breast cancer with nearly the same accuracy as when using a broader set of variables. The random forest classifier proved highly effective in all scenarios, highlighting the potential for reducing diagnostic complexity without sacrificing accuracy. This finding is promising as it suggests that, with fewer resources, we can still achieve reliable predictive results, potentially improving early detection in resource-constrained regions.

INTRODUCTION

Breast cancer is the most common cancer in females, as reported by the World Health Organization (WHO) (1). The most common screening test for breast cancer is a mammogram. While many developed countries have widespread access to mammogram screening programs, women in lower-income regions often face significant barriers, such as lack of access to healthcare, cost, and limited awareness, preventing them from getting regular mammograms (2). This highlights a significant gap in the field, as there is a need for more cost-effective and accessible diagnostic methods.

Fine Needle Aspiration Biopsy (FNAB) is one of many diagnostic procedures to diagnose breast cancer (3). FNAB is a minimally invasive procedure that requires minimal equipment and can be performed in a wider range of healthcare settings, including those with limited resources (3). It involves using a thin needle with a syringe to extract tissue from suspicious lumps, which is then examined under a microscope (3). This simplicity means that FNAB can be performed by trained healthcare professionals with basic equipment, making it easier to implement in rural or underserved areas. Additionally, FNAB can provide quicker results, allowing for faster diagnosis and treatment decisions, which is particularly important in low-income countries where timely access to healthcare services can be a challenge. By using FNAB, healthcare systems in low-income countries can allocate their limited resources more efficiently, potentially improving early detection rates and outcomes.

Several different statistical methods could be used to detect and diagnose breast cancer using features from FNAB, which include Machine Learning (ML) that involves optimization of statistical data, probabilistic theory, and other analytical methods to recognize patterns and make predictions (4). It can be used to predict both categorical and continuous (numeric) data. ML can be classified into supervised and unsupervised. In supervised learning, a training dataset with features (predictor variables) and labels (the ground truth of the variable to be predicted) is used to build models for making predictions. In unsupervised learning, clusters are identified using the available data (4).

We hypothesize that features obtained using FNAB can serve as predictive variables in machine learning algorithms for the accurate detection and diagnosis of breast cancer. The features measured for the Wisconsin Breast Cancer Dataset (WBCD) study are described in detail elsewhere (5). We started this research with feature selection by removing highly correlated variables. We used the data to create models using various modeling algorithms. Then, we applied several ML algorithms, including logistic regression, quadratic discriminant analysis (QDA), linear discriminant analysis (LDA), k-nearest neighbors (kNN), classification tree, and random forests. We used sensitivity and specificity to determine the accuracy of each model (4). Logistic regression and random forest performed the best among all the models, both showing high values for specificity, sensitivity, and accuracy. Logistic regression showed 99.07% specificity with both datasets, sensitivity of 87.5% and 90.63% and accuracy of 94.77% and 95.93% with the 19 variable and six variable datasets, respectively. Random forest stood out by achieving

https://doi.org/10.59720/24-372

the highest performance across all metrics with specificity of 98.15% and 94.44%, sensitivity of 98.44% and 92.19% and accuracy of 98.26% and 93.6% with the 19 variable and six variable datasets, respectively. To further evaluate the model's robustness, we created another reduced variable dataset using principal component analysis (PCA). We applied both logistic regression and random forest to the new dataset. The results were consistent with the results obtained using the original dataset. Our results support our hypothesis that features obtained using FNAB can serve as predictive variables in machine learning algorithms for the accurate detection and diagnosis of breast cancer.

RESULTS

In this study, we analyzed 397 FNAB test results from the WBCD study, comprising 148 (37.28%) malignant cases and 249 (62.72%) benign cases (5).

Using R, we calculated the mean and standard deviation, and performed statistical analysis (Mann-Whitney U-Test) of the key features in the dataset. Our analysis revealed significant differences in all features except for the mean of the fractal dimension between malignant and benign cases (**Figure 1-2**). In general, features such as radius, texture, perimeter, area, and smoothness had higher values for malignant cases compared to benign cases. For example, the

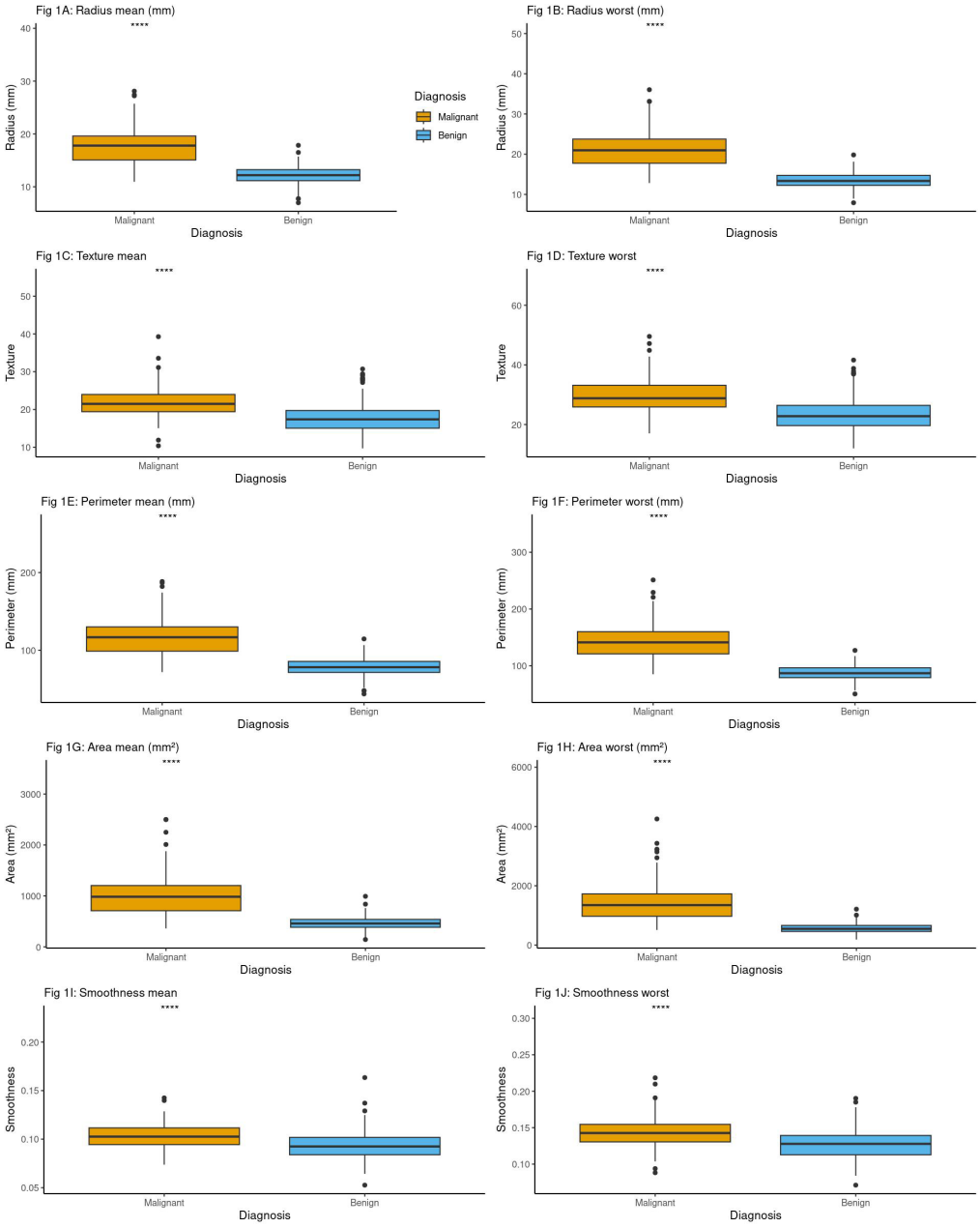

Figure 1: Boxplots of mean and worst of radius, texture, perimeter, area, and smoothness of breast samples in the Wisconsin Breast Cancer Dataset. This boxplot illustrates the distribution of the first five features of a tumor from the Wisconsin Breast Cancer Dataset. The dataset had 10 physical features that were measured using fine needle aspiration biopsy. Each boxplot provides a visual summary of the data's central tendency, variability, and potential outliers for both the mean and worst values of these features. Significance is determined using the Wilcoxon test. ns: not significant; *: p≤0.05, **: p≤0.001, ***: p≤0.001 and ****: p≤0.0001.

https://doi.org/10.59720/24-372

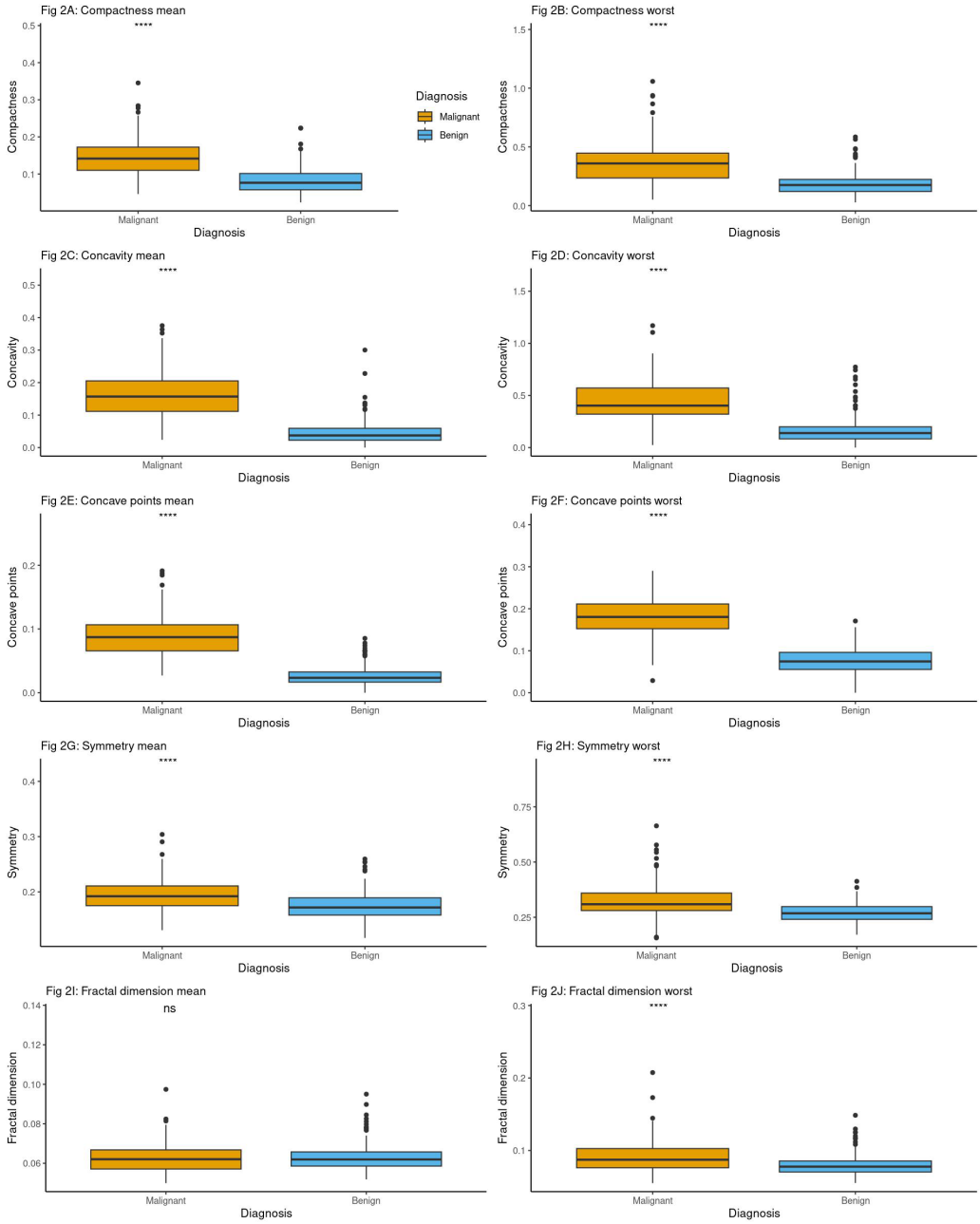

Figure 2: Boxplot of mean and worst values of compactness, concavity, concave points, symmetry, and fractal dimension of breast samples in the Wisconsin Breast Cancer Dataset. This boxplot illustrates the distribution of the last five features of a tumor from the Wisconsin Breast Cancer Dataset. The dataset had 10 physical features that were measured using fine needle aspiration biopsy. Each boxplot provides a visual summary of the data's central tendency, variability, and potential outliers for both the mean and worst values of these features. Significance is determined using the Wilcoxon test. ns: not significant; *: p≤0.05, **: p≤0.001, ***: p≤0.001 and ****: p≤0.0001.

mean radius for malignant tumors was 17.58 mm compared to 12.12 mm for benign tumors ($p < 0.001$). Comparing the most extreme (largest) tumors from each category showed an even bigger difference: the largest malignant tumor measured 21.25 mm in radius, versus 13.37 mm for the largest benign tumor ($p < 0.001$) (**Figure 1**). Similarly, the area of malignant tumors was much larger, with a mean of 993.08 mm² compared to 459.85 mm² for benign cases ($p < 0.001$) (**Figure 1**).

The correlation analysis showed 13 features with very high correlation with each other and a strong correlation with tumor malignancy **(Figure 3).** After removing these 13 features from the original 19 features, the remaining six features (mean radius, mean texture, mean smoothness, mean symmetry, largest symmetry, and largest fractal dimension) were used to create the 6-predictor dataset. The following results demonstrate the performance of the models using the full predictor dataset (19 predictors) and the reduced predictor dataset (six predictors). Results are reported with their 95% confidence interval (CI) in parentheses.

The first model was logistic regression. In this model, the sensitivity (probability of classifying a malignant lesion as malignant) was 95.93% (91.84–98.01%) and was the same

https://doi.org/10.59720/24-372

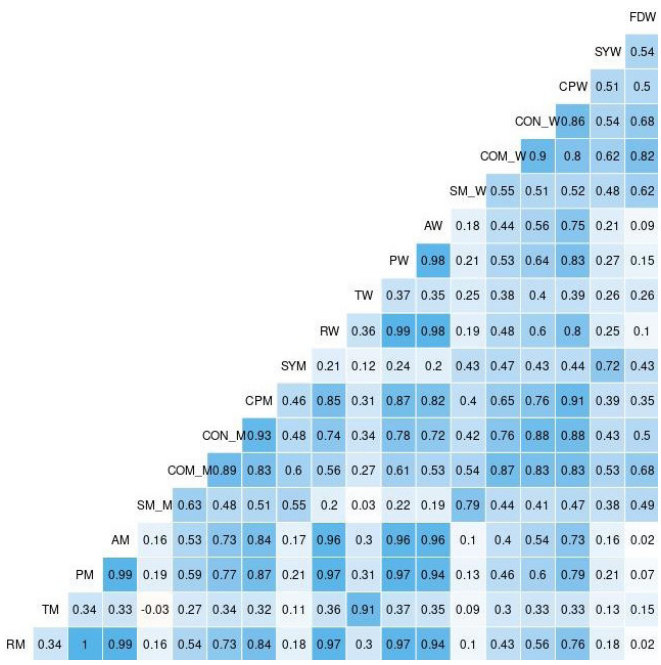

Figure 3: Correlation between predictors. This figure presents the results of the correlation analysis conducted on the predictors. The color gradient, ranging from light to dark blue, indicates the strength of the correlation, with dark blue representing a very high correlation between the predictors. The following abbreviations were used in the figure, R = radius, T = texture, P = perimeter, A = area, SM = smoothness, COM = compactness, CON = concavity, CP = concave points, SY = symmetry, FD = fractal dimension. M represents the mean, and W the worst.

for both the full and reduced predictor model datasets. The specificity (probability of classifying benign as benign) was 95.37% (89.62–98.01%) and 96.30% (90.86–98.55%) for the models with 19 and six predictors, respectively. The overall accuracies of the two models were 95.93% (91.84–98.02%) and 96.51% (92.60–98.39%), respectively. Considering the promising results, we wanted to study the impact of reducing the dataset further using PCA. This PCA-reduced dataset achieved an accuracy of 97.67% (94.1-99.0%). In logistic regression, the conditional probabilities are estimated directly by the Maximum Likelihood Estimator (MLE). No assumptions are made about the distribution of the predictor variables.

With QDA, the sensitivities of the models with 19 and six predictors were 95.93% (91.84–98.01%) and 93.75% (85.00–97.54%), respectively. The specificities were 90.74% (83.79–94.89%) and 94.44% (88.41–97.43%). The overall accuracies were 93.02% (88.20–95.96%) and 94.19% (89.63–96.81%) for the models with 19 and six predictors, respectively.

LDA achieved a very high specificity, with both models having 99.07% (94.94–99.84%). The sensitivity was lower in both models. The model with 19 predictors had a sensitivity of 87.50% (77.23–93.53%), and the model with six had a sensitivity of 90.63% (81.02–95.63%). The accuracies were 94.77% (90.36–97.22%) and 95.93% (91.84–98.01%), respectively.

With kNN, the sensitivities were 93.75% (85.00–97.54%) and 89.06% (79.10–94.60%). The specificities were 97.22% (92.15–99.05%). Accuracies of the model with 19 and six predictors were 95.93% (91.84–98.01%) and 92.44% (87.50–95.53%), respectively. The train function of the *caret* package was to pick the optimal k for this model **(Figure 4)** (6).

In a classification tree, the default minsplit is 20. Minsplit is the number of observations that must exist in a node for a split to be attempted (7). This was changed to five and training was done considering the smaller dataset. The best complexity parameter for 19 predictors was 0.0208 and 0.0166 for the model with six predictors (**Figure 5**). The complexity parameter determines the minimum improvement in the model's fit that is required for a split to be made in the tree (7). The sensitivities were 92.19% (82.98–96.12%) and 82.81% (71.79–0.12%). The specificity was 94.44% (88.41–97.43%). Accuracies of the model with 19 and six predictors were 93.60% (88.91–96.39%) and 89.53% (85.07–93.28%), respectively.

Random forest classifier algorithm resulted in sensitivities of 98.44% (91.66–99.72%) and 92.19% (82.98–96.61%). The specificities were 98.15% (93.5–99.5%) and 94.44% (88.4–97.42%). Accuracies of the model with 19 and six predictors were 98.26% (94.9–99.4%) and 93.6% (88.9–96.39%), respectively. The PCA-reduced feature dataset achieved an accuracy of 94.76% (90.35%-97.22%) and a sensitivity and specificity of 92.3% and 96.26%.

DISCUSSION

This research aims to identify the key physical characteristics of a breast lump collected using FNAB and investigate whether the application of ML on these features would accurately predict breast cancer malignancy. This study utilized the WBCD dataset to test our hypothesis (5). A range of machine learning algorithms, including logistic regression, QDA, LDA, kNN, classification trees, and random forest, were evaluated to compare their performance when using the full dataset of 19 predictors and a reduced dataset of

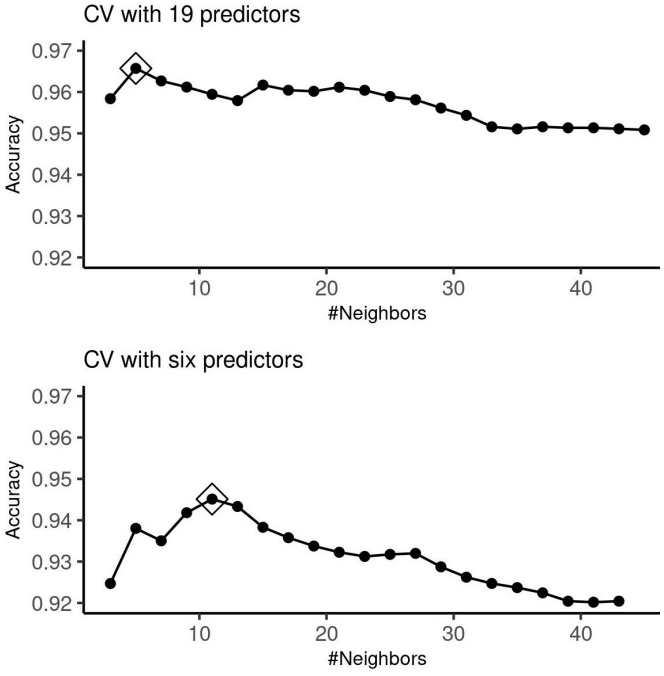

Figure 4: Results of cross validation – K Nearest Neighbor. This shows the results of the cross-validation function in the caret package to determine the value of K for both prediction scenarios.

https://doi.org/10.59720/24-372

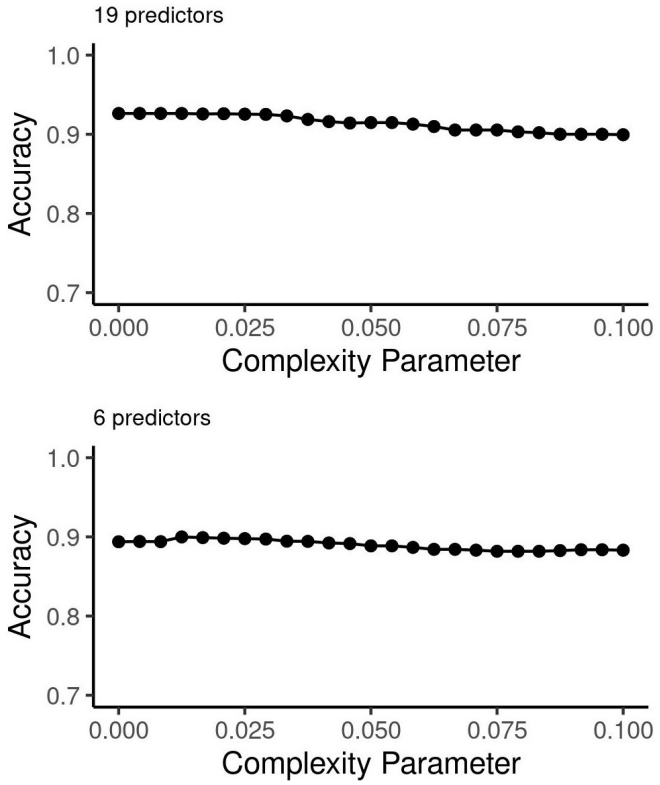

Figure 5: Complexity parameter and accuracy for Classification Tree. Results of the R method to determine the best complexity parameter for the classification tree model. The complexity parameter determines the minimum improvement in the model's fit that is required for a split to be made in the tree.

six key predictors. These predictors were selected based on their high correlation with diagnosis, focusing on size, shape, and texture features of the tumor. Feature reduction was also done using PCA and the two most promising models, logistic regression and random forest, were used against the dataset.

The random forest model demonstrated the highest overall performance with the highest values in sensitivity, specificity, and accuracy in all 3 datasets compared to other models. Logistic regression performed strongly as well. PCA feature reduction further increased the accuracy of the logistic regression model. QDA and LDA followed closely, with LDA standing out for its near-perfect specificity. QDA showed high specificity and accuracy, but its sensitivity reduced with the feature-reduced dataset.

The minimal performance difference between all three kinds of datasets shows promise for using FNAB test results with the appropriate models as a diagnostic tool for breast cancer. This finding, combined with future clinical studies, has the potential to improve breast cancer diagnosis in resource-constrained settings. The success of the random forest classifier, across the 19 variables, six variables, and PCA-reduced datasets, emphasizes the robustness of these physical characteristics as predictors. A follow-up study with a larger dataset with additional variables, along with a comparison of results from mammograms using image processing techniques, could offer more refined and accurate models.

Our findings are consistent with other research that supports the use of ML models for breast cancer diagnosis using FNAB data. For example, Islam et al. evaluated five supervised ML models on a dataset of 500 patients and found that XGBoost achieved the highest accuracy (97%), while decision tree and random forest achieved accuracies of 91% and 96%, respectively (8). Another comprehensive review by Abunasser et al. reviewed 32 studies published between 2002 and 2020 on breast cancer datasets and identified artificial neural networks, support vector machines, and k-nearest neighbors as the most frequently used classifiers, with accuracies ranging from 83% to over 99%, depending on the dataset and feature selection methods (9). Our study used the k-nearest neighbor model, but did not use neural networks or support vector machines. The review by Abunasser et al. showed the accuracy of kNN model ranges from 97.1% to 99.9% while our study gave an accuracy of 95.93% and 92.44% with the 19 and six datasets, respectively. Both studies validate our approach of using ML techniques and appropriate feature selection methods for prediction.

This research underscores the potential for AI-driven healthcare solutions to make diagnostic tools more accessible and cost-effective, especially in resource-limited regions. Future research could explore applying these methods to other cancer types and datasets, as well as investigate advanced techniques such as deep learning to further enhance diagnostic accuracy.

In conclusion, this study supports the hypothesis that data collected from FNAB can serve as predictors of breast cancer and illustrates the promise of ML in enhancing cancer diagnosis.

MATERIALS AND METHODS

For this analysis, we utilized the WBCD dataset, which contains results from FNAB (5). Features from FNAB can serve as predictive variables in machine learning algorithms for the accurate detection and diagnosis of breast cancer. R (version 4.4.3) was used for this study (10).

This dataset had 569 observations, and the predictor variables included the mean, standard error, and the largest observations for radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. In total, 30 features were available for predicting whether the cell's nucleus is suggestive of malignancy (cancer) or benign. The prevalence of malignancy in this dataset was 37.25% (212 tumors). The dataset was randomly split into a training and testing dataset with a 70:30 ratio using the function createDataPartition of the *caret* package (6). This function ensures that the prevalence of the disease is similar in the training and testing datasets.

Though the introductory paper of this dataset, Street, William Nick et al., took the extreme values as its most intuitive, we retained both mean and extreme values to explain the concept of multicollinearity (4). Multicollinearity occurs when two or more variables are highly correlated. This situation complicates the process of identifying the individual impact of each variable on the dependent variable (4).

First, univariate analysis was done to summarize each of the predictor variables and the variable to be predicted in the training dataset. Univariate analysis is a statistical technique that analyzes data with a single variable, or "uni" variable, to summarize and describe the data (4). The statistics we calculated were the mean, standard error, and the largest observation for each measurement described above. As all

https://doi.org/10.59720/24-372

the predictor variables were numeric mean and standard deviation were tabulated and since the dataset is skewed, the Mann-Whitney U-Test was used to test if there was any significant difference in the features among those with malignancy and benign disease. A p-value less than 0.05 was considered statistically significant.

We then performed correlation analysis to identify relationships between predictor variables using the functions ggcorr of the *GGally* package and findCorrelation of the *caret* package (6,11). A cutoff value of 0.75 was used, leading to the exclusion of 13 highly correlated features, resulting in a dataset with six features.

To explore the effect of this correlation on the performance of different algorithms, we created two training datasets – one with all 19 predictors (all features found in the dataset) and one with six predictors (correlation analysis using ggcorr), removing the 13 highly correlated features. We also did feature reduction of the 19 features using the PCA function from FactoMineR package, which resulted in 10 features (12). The most promising algorithms, logistic regression and random forest, were used in the reduced-dimension dataset.

Sensitivity, specificity, and accuracy were computed by creating a confusion matrix using the wilcox.test method from the stats package. Stats package comes with the basic installation of R (4). The 95% CI for each of these metrics was constructed using the modified Wilson's score (4).

All the models used in this research were built using the caret package (6). The first model was logistic regression. Here, the conditional probabilities are estimated directly by MLE. No assumptions are made about the distribution of the predictor variables. In QDA, we assumed that all the predictors were multivariate normal. The mean, standard deviation of each predictor, and the correlation between all the predictors were factored in this model. Next, we chose LDA. This is like QDA, except it assumes that the correlation structure for all the variables is the same, thereby reducing the number of parameters to be computed (4). We then use kNN (4). Unlike the previous algorithms, kNN has a tuning parameter. The number of nearest neighbors (k) must be decided before we use the algorithm on the testing dataset. By default, the train function of the *caret* package does cross-validation and picks the best k**.** A 10-fold repeated cross-validation was chosen to tune the model. Classification Tree is an algorithm that employs recursive (repeated) partitioning to fit data (4). This algorithm has three tuning parameters: 'cp' (complexity parameter – proportion reduction in gini index or entropy), 'minsplit' (minimum number of observation in a node before splitting) and 'minbucket' (minimum number of observations in each node) which is minsplit/3. The default minsplit is 20. This was changed to five and training was done to pick the best complexity parameter. The R script used for the analysis can be found in the GitHub repository: https://github.com/mithravikram09/breastcancer.

ACKNOWLEDGMENTS

We would like to thank all the contributors to the Wisconsin breast cancer dataset.

Received: December 9, 2024
Accepted: March 10, 2025
Published: August 20, 2025

REFERENCES

1. «Breast Cancer: Prevention and Control.» World Health Organization, https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/.
2. Stout, Natasha K et al. "Retrospective cost-effectiveness analysis of screening mammography." Journal of the National Cancer Institute vol. 98, Nov. 2006, https://doi.org/10.1093/jnci/djj210.
3. Amedee, R G, et al. "Fine-needle aspiration biopsy." The Laryngoscope vol. 111, Sep. 2001, https://doi.org/10.1097/00005537-200109000-00011.
4. Irizarry, Rafael A. «Introduction to Data Science: Data Analysis and Prediction Algorithms with R.» 1st edition, Chapman and Hall/CRC, 2019, https://doi.org/10.1201/9780429341830.
5. Wolberg, William, et al. «Breast Cancer Wisconsin (Diagnostic).» UCI Machine Learning Repository, 1993, https://doi.org/10.24432/C5DW2B.
6. Kuhn, Max. «Building Predictive Models in R Using the caret Package.» ***Journal of** Statistical Software*, vol. 28, no. 5, 2008, pp. 1–26, https://doi.org/10.18637/jss.v028.i05.
7. Lucas B.V. de Amorim, et al. «The Choice of Scaling Technique Matters for Classification Performance.» Applied Soft Computing, vol. 133, 2023, 109924, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2022.109924.
8. Islam, Taminul et al. "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI." Scientific reports vol. 14,1 8487. 11 Apr. 2024, https://doi.org/10.1038/s41598-024-57740-5
9. Abunasser, Y., et al. "Machine Learning Approaches for Breast Cancer Diagnosis: A Review of Literature" 2023, https://doi.org/10.1063/5.0133688.
10. R Core Team. «R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.» 2021, https://www.R-project.org/.
11. Schloerke, B., et al. "Extension to Ggplot2" 2010, 2.2.1, https://doi.org/10.32614/CRAN.package.GGally.
12. Lê, Sébastien, et al. "FactoMineR: An R Package for Multivariate Analysis". Journal of Statistical Software, vol. 25, no. 1, Mar. 2008, pp. 1-18, https://doi.org/10.18637/jss.v025.i01.