International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Data Analysis and Machine Learning Approaches to Breast Cancer Classification

Akwasi Oppong and Jerry Galley
Northern Arizona University

Abstract - This research aims to investigate the use of data analysis and machine learning techniques to classify breast cancer tumors as benign or malignant based on the features of the cells as they appear in digital images. The dataset consists of 600 data points and nine variables. The variables include the radius of the cell, texture of the cell, the cell's perimeter, the cell's area, smoothness of the cell, compactness of the cell, symmetry of the cell, and fractal dimension of the cell. Exploratory data analysis was conducted to evaluate the variables' distribution and variability. Several classification algorithms were implemented and tested on the dataset. The algorithms included Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting. The accuracy of the algorithms was also evaluated. The results showed that the accuracy of the algorithms was around 51.7%, indicating that the features were challenging to use in the classification of the types of tumors.

Keywords - Breast Cancer; Benign; Malignant; Gene Expression Data; Tumor Diagnosis; Feature Extraction; Histopathological Images; Support Vector Machine (SVM); Digital Imaging

INTRODUCTION

Breast cancer is one of the most common and deadly diseases afflicting women across the globe. It is a multifarious disease with different types and sub-types. The major types are invasive and non-invasive breast cancers. Invasive breast cancers include Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC), which are more deadly because these types of breast cancers can spread to other parts of the body from the original site through blood or lymphatic vessels. On the contrary, non-invasive breast cancers include Ductal Carcinoma in-situ (DCIS) and Lobular Carcinoma in-situ (LCIS), which are less deadly because these types of breast cancers are confined to the milk ducts or lobular tissue and do not spread to other body parts from the original site.

Breast cancer types are difficult to identify and diagnose in the initial stages. Traditionally, breast cancer types are diagnosed through histopathological examination and imaging techniques along with biomarker analysis carried out by medical experts. Medical experts are able to identify breast cancer types through imaging techniques such as mammography, MRI, and ultrasound. These imaging techniques are used to improve breast cancer detection. However, these techniques are based on expert opinion and are associated with diagnostic delays and variability.

In the recent past, machine learning (ML) has been proposed as a promising technique to assist the traditional approaches of breast cancer diagnosis by facilitating the automated, fast, and more objective classification of breast cancer from complex data. Indeed, ML techniques such as Decision Trees, Random Forests, Support Vector Machines, and Neural Networks are capable of discovering complex patterns from high-dimensional data, including numeric biomarkers and histopathological images. These techniques may assist in the early diagnosis of breast cancer, the classification of breast cancer types (for instance, distinguishing between invasive and non-invasive breast cancer), and the prediction of prognosis. These approaches may assist clinicians in the decision-making process.

The main objective of the research is to utilize the techniques of machine learning to classify breast cancer based on numeric biomarkers and histopathological images. Indeed, by considering the publicly available datasets from sources such as Kaggle, the proposed research will attempt to evaluate the performance of the proposed system in the accurate classification of different breast cancer types. The ultimate objective of the proposed research is to develop reliable, efficient, and accurate computer-assisted breast cancer diagnosis systems.

LITERATURE REVIEW

The use of machine learning techniques in the diagnosis and classification of breast cancer has witnessed a tremendous increase in recent times, mainly due to the requirement for more accurate, rapid, and cost-effective diagnostic tools. A wide range of machine learning algorithms, data types, and strategies have been employed to classify benign and malignant breast cancer, and further classify certain types of breast cancer, such as triple-negative breast cancer.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Rane et al. (2020) presented a comparative study with a focus on the potential of various ML models, including decision trees, SVM, and neural networks, in predicting breast cancer. The results of the research clearly show the potential of ML models in predicting breast cancer with high accuracy. These results are in line with those of Ghiasi and Zendehboudi (2021), who presented the potential of various ML models, including decision trees, in predicting breast cancer with high accuracy using ensemble learning techniques.

Abdulla et al. (2021) presented a comprehensive overview of various ML models used in predicting breast cancer classification problems. The results of the research clearly show the potential of various ML models in predicting breast cancer classification problems with high accuracy. At the same time, the results of the research highlighted the limitations of various ML models in predicting breast cancer classification problems. For instance, the results of the research highlighted the potential of deep learning models in predicting breast cancer classification problems with high accuracy using CNN models. These results are in line with those of Yusoff et al. (2023).

Wu and Hicks (2021) were particularly interested in gene expression data for breast cancer subtype prediction. The authors highlighted that ML algorithms such as SVM, Random Forest, and logistic regression are particularly effective in classifying complex data types. They emphasized that gene expression data can improve classification accuracy. This is further highlighted in Chen et al. (2023), who used ML algorithms to predict types of breast cancer based on gene expression data sets.

Regarding imaging-based ML algorithms, Zhao et al. (2021) used MRI data to identify benign or malignant breast tumors. The authors highlighted that ML algorithms such as CNN and support vector machine can improve diagnostic accuracy over conventional radiological examination. Similarly, Ray et al. (2019) used image data and numeric data to identify breast tumors.

Regarding MRI data analysis, D'Amico et al. (2020) used ML algorithms to identify malignant or benign enhancing tumors. The authors highlighted that feature extraction techniques are particularly useful in improving the accuracy of ML algorithms. Similarly, Savalia and Verma (2023) used various ML algorithms such as KNN, SVM, and ensemble methods to identify breast tumors. The authors highlighted that ensemble methods are particularly useful in classifying high-dimensional data sets.

Lastly, Omotehinwa et al. (2023) also used Light Gradient Boosting Machines (LightGBM) and tree-structured Parzen estimators to illustrate the importance of hyperparameter optimization in improving the accuracy of breast cancer diagnosis, especially in clinical decision support systems.

Classification stages

Researchers and companies have developed CAD systems that can be employed to automate breast cancer classification as benign or malignant. The CAD system can enhance a radiologist's ability to detect and discriminate tumor tissues. The selection of algorithms in a CAD system depends on a better understanding of the contents of cancer images. The classification of breast tumors as benign or malignant involves four major stages.

Processing stages

- Data Acquisition: Collecting relevant data that may include images such as mammograms or MRI scans along with features such as cell sizes, biomarkers, or gene expression profiles.
- Preprocessing: This step involves preprocessing the collected data to prepare it for analysis. Preprocessing may include data cleaning (handling missing data or noisy data), feature normalization or scaling, and feature selection to improve the accuracy of the model.
- Feature Extraction and Selection: Extracting relevant features from the collected data (features such as texture, shape, or molecular features), and then selecting features that are relevant to improve accuracy. This step may include statistical analysis or algorithms to identify features that are best able to distinguish between benign and malignant tumors.
- Model Development and Classification: Using various machine learning algorithms such as Logistic Regression, Decision Trees, Random Forests, or Gradient Boosting to train classification models on features to validate and test the accuracy of these features in classifying different types of tumors.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

DATA ACQUISITION

Te dataset used in this research is the breast cancer data from https://www.kaggle.com/code/ devraai/breast-cancer-diagnosis-analysis-prediction. This dataset contains detailed cell nuclei measurement features commonly used in breast cancer diagnosis. Each entry includes numerical attributes such as radius, texture, perimeter, area, smoothness, compactness, symmetry, and fractal dimension, all essential for medical image analysis and early detection research.

```
## [1] 600  9
## [1] "diagnosis"     "radius_mean"     "texture_mean"
## [4] "perimeter_mean""area_mean"       "smoothness_mean"
## [7] "compactness_mean" "symmetry_mean""fractal_dimension"
```

Table 1: Preview of first 8 rows

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	symmetry_mean	fractal_dimension
M	7.14	14.65	43.11	743.05	0.1282	0.1791	0.1863	0.0760
B	17.69	8.85	105.40	3963.38	0.1964	0.3989	0.1758	0.0708
M	17.89	13.81	90.37	3453.29	0.0688	0.2280	0.2830	0.0573
M	20.02	26.66	113.76	5402.59	0.0525	0.1353	0.1502	0.0404
M	21.97	21.49	126.56	6741.27	0.1655	0.0958	0.2688	0.0698
B	27.47	36.57	159.26	8741.15	0.1711	0.1467	0.1968	0.0571
M	17.36	34.16	106.94	4305.33	0.0680	0.1553	0.2028	0.0841
M	13.11	18.94	79.54	2497.42	0.0898	0.0368	0.1617	0.0414

DATA TYPES AND MISSINGNESS

The table below shows each variable's R class, number of missing values, and number of unique values. This helps determine which variables are numeric vs categorical and whether imputation or filtering is necessary.

Table 2: Data types, missingness and unique counts

variable	dtype	n_missing	n_unique
diagnosis	character	0	2
radius_mean	numeric	0	522
texture_mean	numeric	0	539
perimeter_mean	numeric	0	589
area_mean	numeric	0	600
smoothness_mean	numeric	0	501
compactness_mean	numeric	0	559
symmetry_mean	numeric	0	519
fractal_dimension	numeric	0	364

The table above provides a detailed overview of the variables included in the breast cancer dataset, summarizing each variable's data type, number of missing observations, and the number of unique values. The dataset contains **600 observations and 9 variables**, with no missing values across any field, which supports the reliability and completeness of the data for downstream analysis.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

The variable *diagnosis* is categorical with two unique levels, representing the benign and malignant classifications that serve as the primary outcome of interest. The remaining variables including *radius_mean*, *texture_mean*, *perimeter_mean*, *area_mean*, *smoothness_mean*, *compactness_mean*, *symmetry_mean*, and *fractal_dimension* are all numeric measurements extracted from digital breast mass images. These high-resolution features exhibit substantial variation, as indicated by the large number of unique values (ranging from 364 to 600). This level of variability suggests that the predictors carry rich information that may help discriminate between diagnostic classes. Additionally, the absence of missingness simplifies preprocessing and reduces the need for imputation, allowing the analysis to focus more directly on modeling and interpretation rather than data reconstruction. Overall, this table demonstrates that the dataset is both complete and sufficiently complex to support meaningful statistical investigation.

DATA EVALUATION

The histograms in the appendix provide an important first look at the distributional shapes of the numeric predictors in the dataset, helping assess skewness, spread, and potential irregularities. Several patterns emerge across these visualizations. Variables such as radius_mean, texture_mean, smoothness_mean, compactness_mean, symmetry_mean, and fractal_dimension appear roughly unimodal and fairly symmetric, suggesting that these measurements are well-behaved and do not exhibit extreme skew. Their distributions are relatively continuous, aligning with expectations for biological measurements derived from image features.

In contrast, perimeter_mean and area_mean show noticeable right skew, with most observations concentrated toward lower values and a tail extending toward larger masses. This makes sense biologically, as very large tumors are less common. The skewness in these predictors may later influence modeling choices, such as transformations or the use of algorithms robust to non-normality.

The histograms also reveal that all numeric variables possess substantial variability, which is essential for statistical modeling, predictors with near-constant values would provide little discriminatory information. The relatively smooth shapes of the histograms indicate that the dataset contains a large number of unique values, corroborating the earlier summary table.

Overall, these visualizations confirm that the dataset contains rich, continuous, and diverse numeric predictors, with distributional characteristics that are suitable for modeling but may require attention to skewness in certain variables during analysis.

MODELING

Model Selection and Rationale

For this binary classification task distinguishing malignant (M) tumors from benign (B) tumors, we selected four predictive modeling approaches: Logistic Regression, Decision Trees, Bagged Decision Trees (Random Forest), and Gradient Boosting (GBM). These models were chosen to provide a balanced comparison between interpretable statistical models and powerful ensemble learning techniques commonly used in modern predictive analytics.

Logistic Regression was included because of its interpretability and long-standing use in medical applications, where understanding the relationship between predictors and outcomes is essential.

The Decision Tree model was selected for its intuitive structure and ability to identify non-linear patterns in the data. Random Forest, an ensemble of bagged trees, was chosen for its high predictive accuracy and its strength in reducing variance through bootstrapping. Finally, we incorporated a Gradient Boosting Model, a state-of-the-art ensemble technique that builds sequential trees to correct errors from prior ones, often delivering exceptional predictive performance on complex datasets.

Model Implementation

The implementation followed a structured workflow to ensure consistency and reproducibility. We began by preprocessing the dataset: encoding the diagnosis variable numerically, scaling predictor variables, and removing extraneous columns. The data was then divided into training and testing sets to ensure that model performance was evaluated on unseen observations.

Each model was built using established R packages. Logistic Regression was fitted using the glm() function with a binomial family. Decision Trees were created using the rpart package, providing interpretable tree-based structures. Random Forest models were implemented through the randomForest package, allowing ensemble learning through multiple bootstrapped trees. The Gradient Boosting Model was implemented using the gbm package, which constructs trees sequentially to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

optimize predictive accuracy. To evaluate model performance, we used confusion matrices and overall accuracy providing a comprehensive assessment of classification effectiveness.

LIMITATIONS OF THE MODELS

Despite their strengths, each model comes with limitations. Logistic Regression assumes linearity in the log-odds, which may reduce performance if the underlying data relationships are highly nonlinear. Decision Trees, while easy to interpret, are susceptible to overfitting unless carefully tuned or pruned. Random Forest reduces this risk through bootstrapped aggregation, but the resulting ensemble sacrifices interpretability, making it difficult to understand individual decision pathways, an important consideration in medical contexts. Gradient Boosting Models, although capable of very strong predictive performance, require careful tuning of parameters such as learning rate, depth, and number of trees. They are also more computationally intensive and can overfit if the boosting process is not adequately regulated.

ANALYSIS RESULTS

To evaluate the performance of the classification models, four different approaches were applied to the breast cancer dataset: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting (GBM). The models were assessed using confusion matrices and accuracy metrics on the test set. Overall, the models achieved accuracies ranging from 0.425 to 0.558, indicating varying levels of predictive performance, but none achieving high reliability in distinguishing benign (0) from malignant (1) cases.

Logistic Regression produced an accuracy of 0.425, making it the weakest performer among the models. The model demonstrated moderate sensitivity (0.6250) but low specificity, indicating difficulty in identifying malignant tumors correctly. Its negative kappa value also suggests that the model performs worse than a naive classifier relying solely on class proportions. These results indicate that the linear decision boundary used by logistic regression does not sufficiently capture the underlying patterns in the dataset.

The Decision Tree model performed moderately, reaching an accuracy of 0.450. Its sensitivity (0.5357) and specificity (0.3750) show that the tree struggled across both classes. Decision trees often overfit to training data, and the reduced generalization observed here is consistent with this tendency. Despite being interpretable, the decision tree does not appear to capture the complexity of the relationships between predictors and tumor diagnosis.

The Random Forest, which aggregates multiple decision trees to improve robustness, achieved the best performance of the four models, with an accuracy of 0.5167. This model shows a sensitivity (0.6250) and improved balanced accuracy relative to the others. The increase in performance suggests that ensemble methods better capture the nonlinear and interacting effects present in the data. However, even with this improvement, the accuracy still remains modest, indicating that predictive signal in the dataset may be weak or overshadowed by noise.

The Gradient Boosting Model (GBM) produced unusual behavior. While its accuracy of 0.4667 appears close to that of the other models, the GBM predicted the class “0” for every single observation in the test set. This leads to perfect sensitivity (1.00) but zero specificity (0.00), making the model effectively unusable for distinguishing between the two classes. This collapse into predicting a single class can occur when the algorithm identifies a minimal-loss solution that defaults to the majority class in the dataset or when earlier boosting iterations dominate the model’s predictions. As a result, the GBM achieves an accuracy that is misleadingly average but contributes no meaningful diagnostic capability.

In summary, the Random Forest performed best among the models tested, though even its accuracy remained moderate. Logistic Regression and Decision Tree models performed relatively poorly, and the GBM failed to provide meaningful classification. These results highlight the challenges inherent in modeling this dataset and suggest that additional feature engineering, model tuning, or alternative algorithms may be needed to achieve stronger predictive performance in classifying tumor diagnoses.

Table 3: Accuracy Comparison Table

Model	Accuracy	Notes
Logistic Regression	0.4250	
Decision Tree	0.4500	
Random Forest	0.5167	

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Gradient Boosting	0.4667	Predicts all 0s

5 Fold Cross-Validation of Models

Choosing five folds represents a practical balance between bias, variance, and computational cost. With fewer folds, such as two or three, the performance estimates tend to have higher variance because too much data is held out for testing at one time. Conversely, using many folds like ten produces lower variance but increases computation time and can yield diminishing returns in terms of improved reliability. For medium-sized datasets, which are common in applied statistical learning contexts, 5-fold cross-validation is widely considered an effective compromise: it is stable, efficient, and conservative enough to avoid overfitting.

Each iteration of 5-fold cross-validation uses approximately 80% of the data for training and 20% for testing. This mirrors the common 80/20 split used in simpler evaluation methods, but with the advantage that all observations serve both as training and testing data across different folds. As a result, the performance estimate is not dependent on a single arbitrary split but is instead an average across multiple balanced partitions. This repeated evaluation makes 5-fold cross-validation a robust and widely accepted method for assessing the generalization ability of predictive models

Table 4: Average Accuracy from 5-Fold Cross-Validation

Model	Average_Accuracy
Logistic Regression	0.5051390
Decision Tree	0.4878332
Random Forest	0.5075646
GBM	0.5096282

Across all four models, the observed accuracies hover around 50%, which is only marginally better than random guessing in a binary classification setting. Such uniformly low performance suggests that the dataset may be inherently challenging to model, or that the predictive algorithms require additional refinement through hyperparameter tuning, improved feature engineering, or alternative modeling strategies. When accuracy remains near chance levels across multiple methods, it often indicates that the underlying patterns distinguishing the classes are either weak, noisy, or not adequately represented by the available features.

Among the models evaluated, Gradient Boosting emerged as the best performer, with an average accuracy of approximately 0.5096. Random Forest followed closely with an accuracy near 0.5076. These ensemble-based methods are well known for their ability to improve predictive performance by aggregating the outputs of many decision trees. Through repeated sampling and combination, they reduce variance and capture more complex relationships than a single tree can achieve. Their slight advantage over the other models in this analysis is consistent with their general reputation for robustness and adaptability.

In comparison, Logistic Regression achieved an accuracy of about 0.5051, outperforming the single Decision Tree model, which averaged around 0.4878. This indicates that a simple linear classifier generalized better than a lone decision tree. The Decision Tree's relatively poor performance is likely due to overfitting, a common issue when a single tree attempts to learn overly specific patterns in the training data that fail to generalize well to new observations. Logistic Regression, by contrast, imposes a stronger structural constraint through its linear decision boundary, which can be advantageous when the signal in the data is weak or when more flexible models pick up spurious relationships.

The small differences in accuracy across all methods carry an important implication: none of the models appear to be capturing strong or reliable predictive patterns. This could stem from several factors, such as insufficiently informative features, high levels of noise in the data, or an underlying lack of separation between the two classes. It may also indicate that more substantial tuning, such as adjusting tree depths, learning rates, or regularization parameters-is necessary to unlock better performance. Regardless, the consistently modest results underscore the need for deeper investigation into the data and modeling approach to determine whether meaningful predictive improvement is attainable

DISCUSSION OF FINAL MODEL AND ANALYSIS

The table below summarizes the performance of logistic regression models built using only one predictor at a time, providing insight into how informative each individual variable is for distinguishing between malignant and benign cases. Overall, the accuracy levels fall within a narrow range, approximately 0.52 to 0.55, indicating that no single predictor carries strong discriminatory power on its own.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Among the variables tested, texture_mean achieves the highest accuracy (0.548), suggesting it provides slightly more useful information for classification compared to the other predictors. However, even this top-performing variable only marginally exceeds random guessing (50%), implying its predictive strength is limited when used in isolation.

Other predictors such as compactness_mean, fractal_dimension, and symmetry_mean also show mildly above-chance accuracy (0.526–0.530). Meanwhile, variables like radius_mean, perimeter_mean, area_mean, and smoothness_mean all yield identical accuracy values (0.520), reinforcing the idea that, individually, these features do not strongly separate the two classes.

Table 5: Accuracy for Logistic Regression Models (Single Predictor)

	Predictor	Accuracy
Accuracy	radius_mean	0.5200000
Accuracy1	texture_mean	0.5483333
Accuracy2	perimeter_mean	0.5200000
Accuracy3	area_mean	0.5200000
Accuracy4	smoothness_mean	0.5200000
Accuracy5	compactness_mean	0.5283333
Accuracy6	symmetry_mean	0.5266667
Accuracy7	fractal_dimension	0.5300000

CONCLUSION

This project explored several statistical and machine learning approaches for predicting whether breast tumors are benign or malignant. Through the implementation of Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, and single-predictor logistic regression models, we gained insight into both the structure of the dataset and the strengths and weaknesses of various modeling techniques.

One of the most important lessons from this project is that no model achieved high predictive accuracy, with results clustering around 50–55% across nearly all methods **(Brian Risk, Subhrajyoti Basu)**. This suggests that the dataset, as provided and preprocessed, does not contain strong, easily extractable patterns that clearly distinguish the two tumor classes. Even advanced ensemble techniques such as Random Forests and Gradient Boosting only produced marginal gains over simpler methods. This reinforces a key principle in predictive modeling: strong algorithms cannot compensate for weak, noisy, or insufficiently informative features.

From a modeling standpoint, the project highlighted several valuable insights. Logistic Regression demonstrated the benefit of simplicity and stability, outperforming the Decision Tree despite its restrictive linear structure. The Decision Tree underperformed due to its tendency to overfit, especially when working with weak signals. Ensemble methods improved stability and accuracy, confirming their theoretical advantages, but they did not achieve the level of performance typically expected for breast cancer datasets, again indicating constraints in the available feature set. The GBM's collapse into predicting only a single class illustrates the importance of hyperparameter tuning and diagnostic evaluation when working with highly flexible models.

The single-predictor logistic regression analysis further demonstrated that no individual variable is a strong discriminator. Accuracy values clustered between 0.52 and 0.55, with texture_mean performing slightly better than others. These results emphasize that predictive power in this dataset lies not in any single measurement but potentially in complex interactions or nonlinear relationships that were not fully captured by the models in their current forms.

Reflecting on the project as a whole, several pros and cons emerge. A major strength was the use of multiple modeling paradigms, allowing for a thorough comparison of performance and interpretability. Additionally, the structured modeling pipeline-preprocessing, training, and cross-validation, ensured fairness in evaluation. However, the limitations of the dataset and the lack of tuning in the ensemble models likely contributed to the modest outcomes. If the project were extended, several improvements could be explored:

- Hyperparameter tuning for tree-based models to prevent pathological behaviors such as the GBM's one-class prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

- Feature engineering, including nonlinear transformations or interaction terms, which may help capture patterns missed by the current models.(Brian Risk)
- Exploring more advanced algorithms, such as support vector machines or neural networks, which may uncover hidden structures in the data.

Overall, the project demonstrates the importance of aligning model complexity with data quality and feature informativeness. While no method yielded strong predictive accuracy, the investigation provided a deeper understanding of the challenges of medical prediction tasks, the nuances of machine learning workflows, and the need for thoughtful preprocessing, model selection, and tuning.

CITATIONS

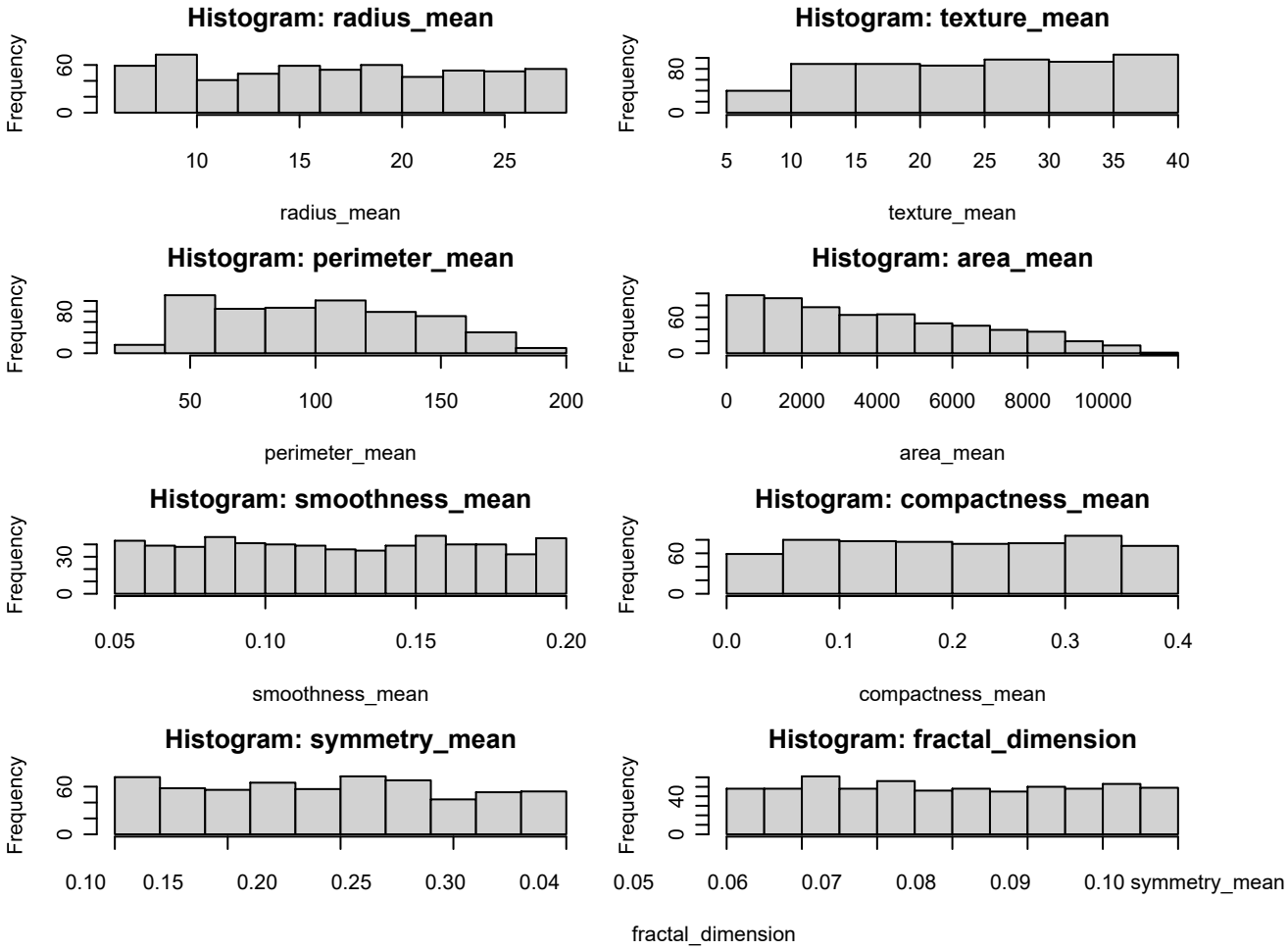
[1] Brian Risk: https://www.kaggle.com/code/devraai/breast-cancer-diagnosis-analysisprediction
[2] Subhrajyoti Basu: https://www.kaggle.com/code/subhrajyotibasu2005/breast-cancerdiagnosis-classification
[3] Rane, N., Sunny, J., Kanade, R. and Devi, S., 2020. Breast cancer classification and prediction using machine learning. International Journal of Engineering Research and Technology, 9(2), pp.576-580.
[4] Abdulla, S.H., Sagheer, A.M. and Veisi, H., 2021. Breast cancer classification using machine learning techniques: A review. Turkish Journal of Computer and Mathematics Education, 12(14), pp.1970-1979.
[5] Wu, J. and Hicks, C., 2021. Breast cancer type classification using machine learning. Journal of personalized medicine, 11(2), p.61.
[6] Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y. and Cai, G., 2023. Classification prediction of breast cancer based on machine learning. Computational intelligence and neuroscience, 2023(1), p.6530719.
[7] Yusoff, M., Haryanto, T., Suhartanto, H., Mustafa, W.A., Zain, J.M. and Kusmardi, K., 2023. Accuracy analysis of deep learning methods in breast cancer classification: A structured review. Diagnostics, 13(4), p.683.
[8] Ghiasi, M.M. and Zendehboudi, S., 2021. Application of decision tree-based ensemble learning in the classification of breast cancer. Computers in biology and medicine, 128, p.104089.
[9] Omotehinwa, T.O., Oyewola, D.O. and Dada, E.G., 2023. A light gradient-boosting machine algorithm with tree-structured parzen estimator for breast cancer diagnosis. Healthcare Analytics, 4, p.100218.
[10] Kalaiyarasi, M., Dhanasekar, R., Sakthiya Ram, S. and Vaishnavi, P., 2020, November. Classification of benign or malignant tumor using machine learning. In IOP Conference Series: Materials Science and Engineering (Vol. 995, No. 1, p. 012028). IOP Publishing.
[11] Ray, R., Abdullah, A.A., Mallick, D.K. and Ranjan Dash, S., 2019, November. Classification of benign and malignant breast cancer using supervised machine learning algorithms based on image and numeric datasets. In Journal of Physics: Conference Series (Vol. 1372, No. 1, p. 012062). IOP Publishing.
[12] D'Amico, N.C., Grossi, E., Valbusa, G., Rigiroli, F., Colombo, B., Buscema, M., Fazzini, D., Ali, M., Malasevschi, A., Cornalba, G. and Papa, S., 2020. A machine learning approach for differentiating malignant from benign enhancing foci on breast MRI. European Radiology Experimental, 4(1), p.5.
[13] Savalia, M.R. and Verma, J.V., 2023. Classifying malignant and benign tumors of breast cancer: A comparative investigation using machine learning techniques. International Journal of Reliable and Quality E-Healthcare (IJRQEH), 12(1), pp.1-19.
[14] Zhao, Y., Chen, R., Zhang, T., Chen, C., Muhelisa, M., Huang, J., Xu, Y. and Ma, X., 2021. MRI-based machine learning in differentiation between benign and malignant breast lesions.

Frontiers in Oncology, 11, p.552634.

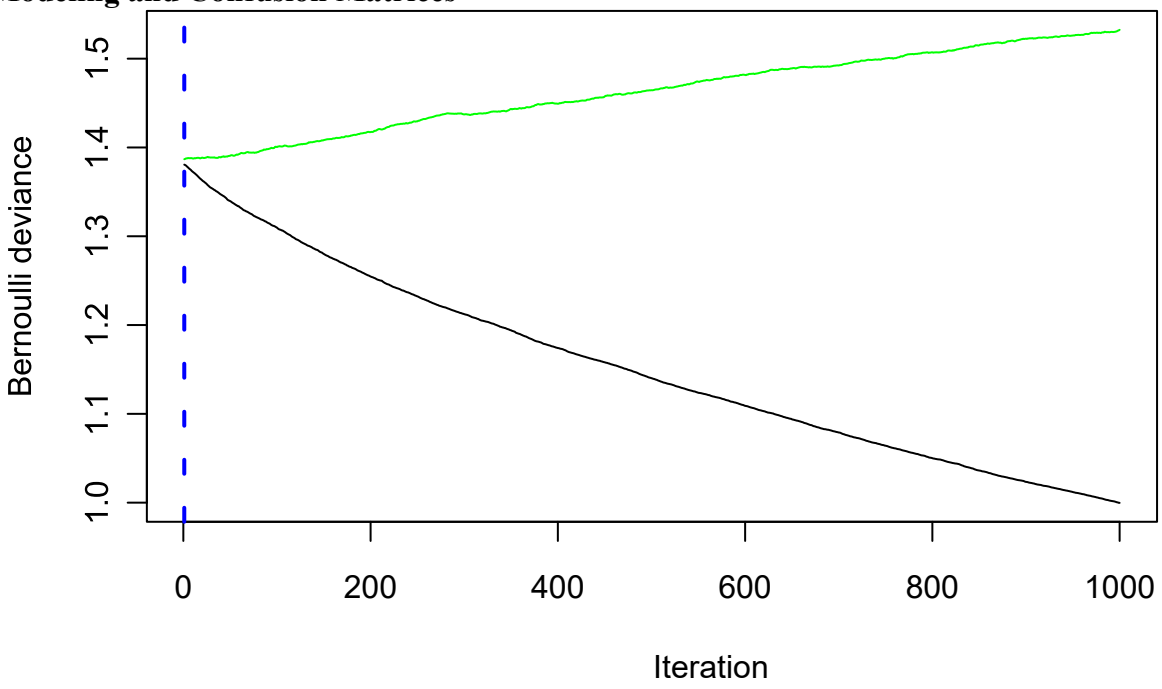
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Appendix

Plots of Predictor Variables



Modeling and Confusion Matrices



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction 0 1
##         0 35 48
##         1 21 16
##
##               Accuracy : 0.425
##                 95% CI : (0.3353, 0.5185)
##    No Information Rate : 0.5333
##    P-Value [Acc > NIR] : 0.993218
##
##                  Kappa : -0.1213
##
## Mcnemar's Test P-Value : 0.001748
##
##            Sensitivity : 0.6250
##            Specificity : 0.2500
##         Pos Pred Value : 0.4217
##         Neg Pred Value : 0.4324
##     Prevalence : 0.4667 ##
Detection Rate : 0.2917
##  Detection Prevalence : 0.6917
##     Balanced Accuracy : 0.4375
##
##       'Positive' Class : 0
##


## Confusion Matrix and Statistics
##
##          Reference
## Prediction 0 1
##         0 30 40
##         1 26 24
##
##               Accuracy : 0.45
##                 95% CI : (0.3591, 0.5435)
##    No Information Rate : 0.5333
##    P-Value [Acc > NIR] : 0.9725
##
##                  Kappa : -0.0879
##
## Mcnemar's Test P-Value : 0.1096
##
##            Sensitivity : 0.5357
##            Specificity : 0.3750
##         Pos Pred Value : 0.4286
##         Neg Pred Value : 0.4800
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

```
##    Prevalence : 0.4667 ##
Detection Rate : 0.2500
##  Detection Prevalence : 0.5833
##     Balanced Accuracy : 0.4554
##
##      'Positive' Class : 0
##


## Confusion Matrix and Statistics
##
##          Reference
## Prediction 0 1
##         0 35 37
##         1 21 27
##
##             Accuracy : 0.5167
##               95% CI : (0.4237, 0.6088)
##   No Information Rate : 0.5333
##   P-Value [Acc > NIR] : 0.67682
##
##                Kappa : 0.0461
##
## Mcnemar's Test P-Value : 0.04888
##
##           Sensitivity : 0.6250
##           Specificity : 0.4219
##        Pos Pred Value : 0.4861
##        Neg Pred Value : 0.5625
##    Prevalence : 0.4667 ##
Detection Rate : 0.2917
##  Detection Prevalence : 0.6000
##     Balanced Accuracy : 0.5234
##
##      'Positive' Class : 0
##


## Confusion Matrix and Statistics
##
##          Reference
## Prediction 0 1
##         0 56 64
##          1 0 0
##
##             Accuracy : 0.4667
##               95% CI : (0.3751, 0.5599)
##   No Information Rate : 0.5333
##   P-Value [Acc > NIR] : 0.9399
##
##                Kappa : 0
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

```
##
## Mcnemar's Test P-Value : 3.407e-15
##
##          Sensitivity : 1.0000
##          Specificity : 0.0000
##       Pos Pred Value : 0.4667
##        Neg Pred Value :  NaN
##    Prevalence : 0.4667 ##
Detection Rate : 0.4667
##  Detection Prevalence : 1.0000
##    Balanced Accuracy : 0.5000
##
##      'Positive' Class : 0
##
```