World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Research Article)

Breast cancer diagnosis using logistic regression on top predictive features

Aabha Parag Tembhurne *

Department of Biotechnology RV College Of Engineering, Bengaluru, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(03), 691-702

Publication history: Received on 27 March 2025; revised on 05 June 2025; accepted on 07 June 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.3.0970

Abstract

Breast cancer is among the most prevalent and life-threatening female cancers globally. Early and precise diagnosis is essential in enhancing patient survival. This research evaluates logistic regression to classify breast tumours as malignant or benign through a publicly available database of 569 cases. We centred on two sets of features: the top 5 and top 10 most predictive features for tumour size and shape irregularity. The logistic regression classifier attained an accuracy of about 94.7% using the top 5 features and 97.4% using the top 10 features, exhibiting sound performance with a smaller set of features. Visualization methods also attested to unique distribution patterns between benign and malignant cases. Our results demonstrate the promise of feature selection and simple but robust models for accurate breast cancer diagnosis. Ensemble methods and validation using an external dataset will be considered in future work to improve generalizability.

Keywords: Breast cancer; Logistic regression; Feature selection; Diagnosis accuracy; Tumor classification; Data visualization

1. Introduction

Breast cancer is a disease where abnormal breast cells grow uncontrollably, forming tumours. If left untreated, these tumours can spread throughout the body, becoming life-threatening. Breast cancer cells originate in the milk ducts and/or milk-producing lobules of the breast. The earliest form (in situ) is not typically life-threatening and can be detected early. However, invasive breast cancer can spread to nearby breast tissue, lymph nodes, or other organs. Treatment for breast cancer often involves surgery, radiation therapy, and medications.

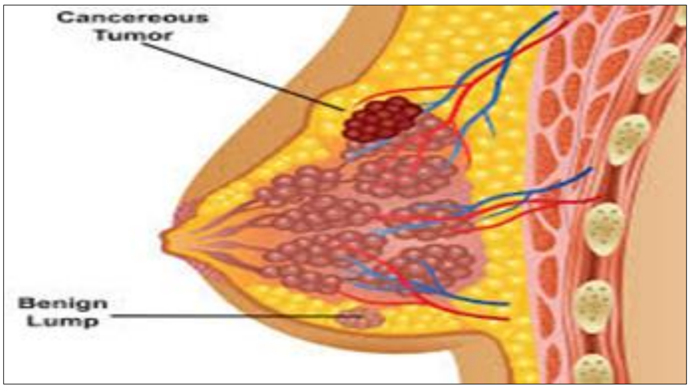


Figure 1 Cancerous lumps in breast

* Corresponding author: Aabha Parag Tembhurne

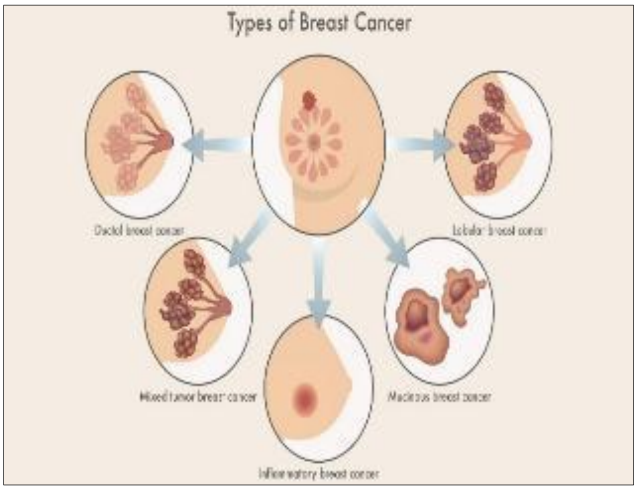


Figure 2 Types of breast cancer

Breast cancer ranks among the most common causes of death related to cancer in women all over the world.

Early diagnosis and proper diagnosis are critical for proper treatment and survival rates. Conventional diagnostic procedures like biopsies and imaging are time-consuming and, at times, subjective. Hence, the application of machine learning algorithms has become increasingly popular in recent years for making quick and trustworthy diagnoses for medical professionals.

Here, we utilize logistic regression, a widely used and easily interpretable machine learning model, to predict whether a breast tumour is benign or malignant. We train the model on a well-documented breast cancer dataset consisting of 569 patient samples with various features summarizing cell nucleus characteristics in biopsied tumors. We emphasize the relevance of feature selection by assessing model performance based on the top 5 and top 10 most strongly related features to cancer diagnosis.

By comparing the predictive ability of these features and the accuracy of the model, we seek to identify the most important factors determining diagnosis and illustrate the real-world applicability of logistic regression in medical decision-making. This work complements the current endeavor to establish effective and affordable diagnostic instruments for breast cancer.

2. Materials and Methods

2.1. Dataset Description

The dataset employed in this research is the Breast Cancer Wisconsin (Diagnostic) Dataset, made available from the UCI Machine Learning Repository. It includes information from 569 patient samples, each characterized by 30 numerical features extracted from digitized images of fine needle aspirate biopsies of breast masses. The features correspond to cell nucleus characteristics, including radius, texture, perimeter, area, smoothness, and concavity. The target variable, diagnosis, indicates whether the tumor is benign (0) or malignant (1).

2.2. Data Preprocessing

The unnecessary columns like the patient ID were first dropped. The diagnosis labels, initially categorical (M for malignant and B for benign), were converted into binary values (1 and 0, respectively). The dataset was scanned for missing values, and there were none. Features were standardized utilizing z-score normalization to make their mean zero and variance one, which aids in model convergence.

2.3. Feature Selection

To achieve optimal model performance and interpretability, feature selection was conducted. The top 5 and top 10 features most strongly correlated with diagnosis were determined using exploratory data analysis and correlation measures. These subsets were utilized to train individual logistic regression models to assess performance.

2.4. Model Details

Supervised classification algorithm logistic regression, being simple and interpretable, as employed for developing the prediction models. It predicts the probability that a particular sample is from the malignant class given the input featuresThe data was randomly divided into 80% training and 20% testing sets to test model generalization.

2.5. Evaluation Metrics

Model performance was evaluated with accuracy, precision, recall, and F1-score. Accuracy estimates how correct the model is overall. Precision shows the percentage of true positives out of the total predicted positives, whereas recall estimates the percentage of true positives found among all actual positives. The F1-score gives a trade-off between precision and recall, particularly helpful when class distributions are uneven.

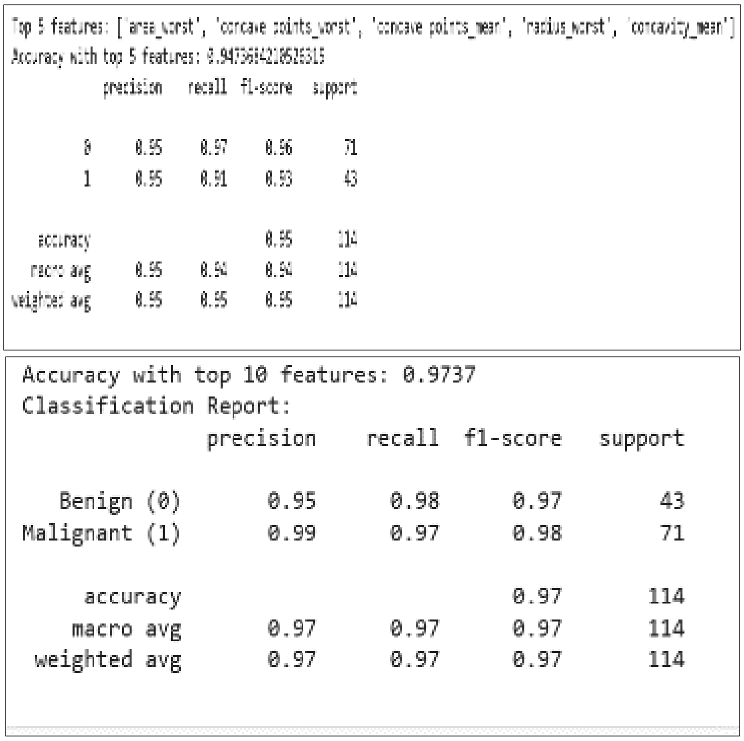
3. Results

3.1. Summary Statistics

The data consisted of 569 samples with 30 numerical features representing tumor cell attributes. There were 357 benign and 212 malignant cases for the diagnosis labels. Simple statistics indicated significant differences in the feature values for the two classes, justifying the likelihood of successful classification.

3.2. Model Performance

Two logistic regression models were tested with the top 5 and top 10 selected features due to their correlation with the diagnosis. The model with the top 5 features resulted in an accuracy of about 94.7%, whereas the model with the top 10 features enhanced accuracy to about 97.4%. This illustrates that including more pertinent features strengthens model prediction capability without overfitting.



```
Top 5 features: ['area_worst', 'concave points_worst', 'concave points_mean', 'radius_worst', 'concavity_mean']
Accuracy with top 5 features: 0.9473684210526315
              precision    recall  f1-score   support


           0       0.95      0.97      0.96        71
           1       0.95      0.91      0.93        43

    accuracy                           0.95       114
   macro avg       0.95      0.94      0.94       114
weighted avg       0.95      0.95      0.95       114
```

```
Accuracy with top 10 features: 0.9737
Classification Report:
                precision    recall  f1-score   support

   Benign (0)       0.95      0.98      0.97        43
Malignant (1)       0.99      0.97      0.98        71

     accuracy                           0.97       114
    macro avg       0.97      0.97      0.97       114
 weighted avg       0.97      0.97      0.97       114
```

Figure 3 Top features accuracy

Heatmap of correlation indicates the strength of relationship between each feature pair (your dataset columns). It is a nice method to identify redundancy or relationships between features.

- What are the numbers and colours indicating?
- Red (near +1): High positive correlation — as one value rises, so does the other.

- Blue (near -1): High negative correlation — as one value rises, the other falls.
- White/light hues (approximately 0): No high correlation — the characteristics aren't closely aligned.
- Diagonal cells (1.0): A characteristic in comparison to itself — always has ideal correlation (1.0).

What can we conclude?

radius mean, perimeter mean, and area mean all have extremely high positive correlations (~0.99). This indicates that if one of these is high, the others probably are high as well — which makes sense for tumor radius.

The _worst versions (such as radius worst, area worst) are also highly correlated with one another.

Attributes such as fractal dimension and smoothness have low correlation with others — they might provide distinct information.

The id column (just a tag) and Unnamed: 32 (presumably empty or not useful) can be dropped prior to training a model.

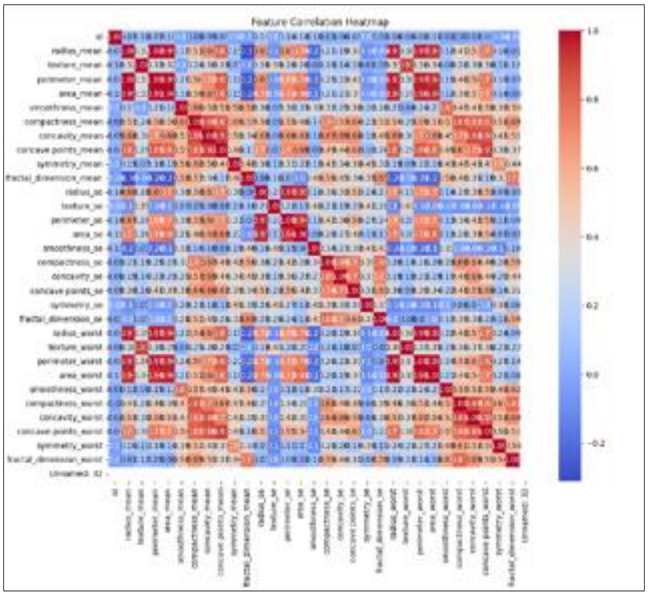


Figure 4 Feature corelation heatmap

This bar chart shows the breakdown of benign (B) cases as well as malignant (M) cases. We have seen more benign cases than malignant cases. This is normal in medical datasets as there are simply fewer people with malignant tumors, though they are dangerous.

This box plot shows a comparison of tumor size (using radius mean) between benign and malignant groups. Malignant tumors generally have larger radii (larger numbers for radius mean) than benign tumors. You can see that the middle line in each box (which shows the median) is clearly larger for malignant tumors. There is a greater spread of values for malignant tumors which indicates more variation in size. There are a few extreme values/outliers (dots) in both groups, but most noticeably in the malignant group indicating that some malignant tumors are quite a bit larger/smaller than the average size.

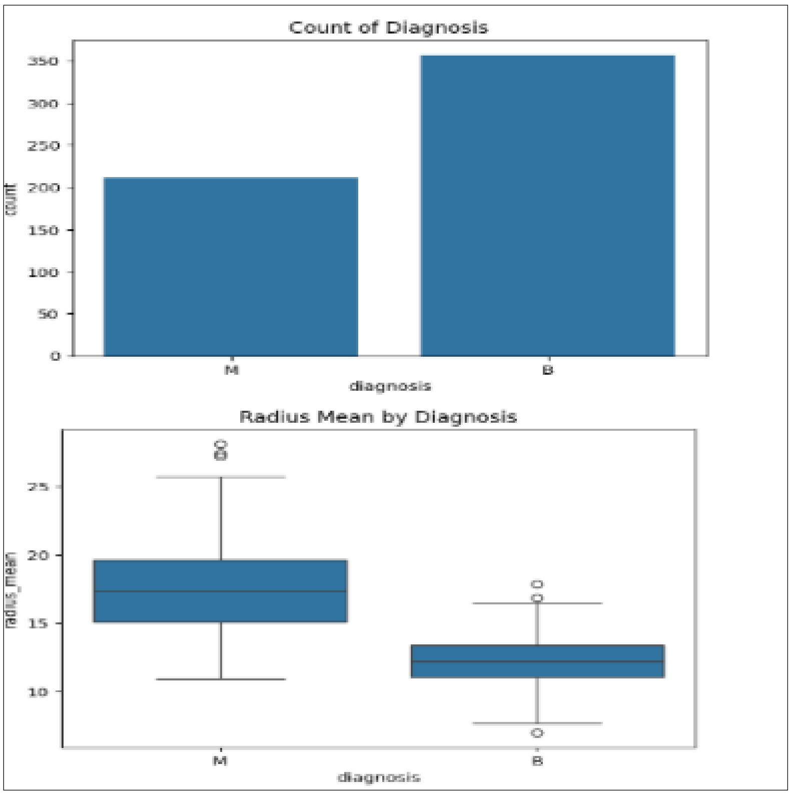


Figure 5 Count of diagnosis and radius mean by diagnosis

3.3. Confusion Matrix

Table 1 This chart shows what the model got right and what it got wrong

	Predicted Benign (0)	Predicted Malignant (1)
Actual Benign (0)	70 (correct)	1 (false positive)
Actual Malignant (1)	2 (false negative)	41 (correct)

- True positive (41): These represent the malignant tumors that are correctly detected.
- True negatives (70): These represent the benign tumors that are correctly detected.
- False positive (1): Benign tumor that are incorrectly marked as malignant
- False negative (2): Malignant tumor that are incorrectly marked as benign. (very dangerous)

$$\boxed{\text{Accuracy} = \frac{TP + TN}{Total} = \frac{70 + 41}{114} \approx 97.4\%}$$

This is a very high accuracy.

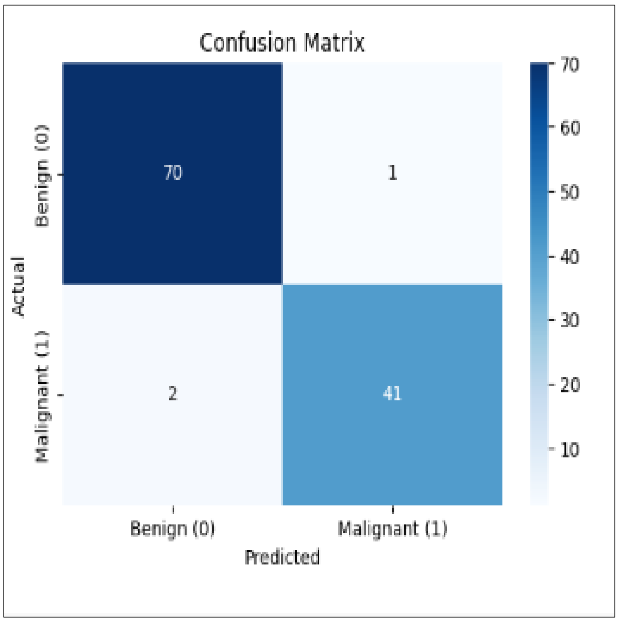


Figure 6 Confusion matrix

3.4. ROC Curve (Logistic Regression)

This graph shows how well the model separates benign and malignant tumors.

3.4.1. How to read it

- Y-axis = True Positive Rate (Sensitivity)
- X-axis = False Positive Rate
- A curve closer to the top-left corner means better performance.

3.4.2. Result

- The ROC curve is very close to the top left.
- AUC (Area Under Curve) = 1.00 → This is perfect performance.
- The model is excellent at telling the two classes apart.

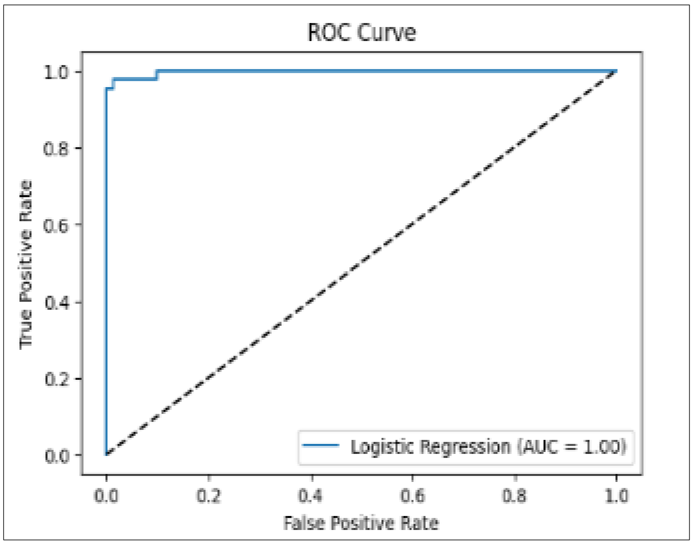


Figure 7 ROC curve

3.5. ROC Curve Comparison: Logistic, SVM, Random Forest

3.5.1. This chart compares the ROC performance of three models

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest

3.5.2. All models perform equally well

- All have AUC = 1.00.
- All ROC curves nearly overlap.
- This means each model is highly capable of distinguishing between tumor types.

3.5.3. You might now choose based on

- Speed (Logistic is fast),
- Accuracy under noise (Random Forest handles it well),
- Interpretability (Logistic is easy to explain).

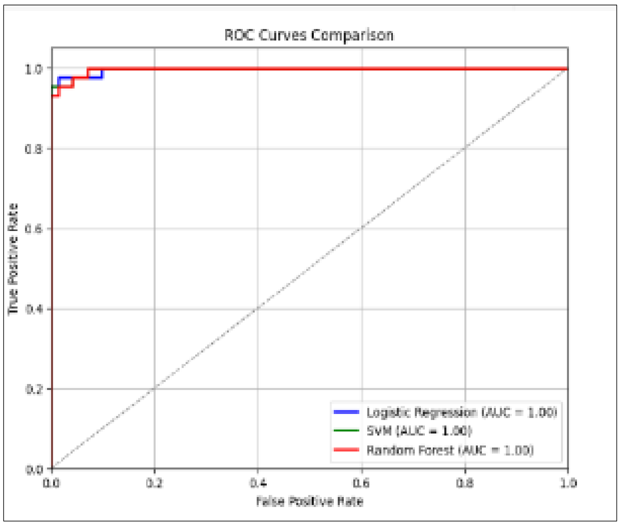


Figure 8 ROC curve comparision

3.6. Feature Importance (from Random Forest)

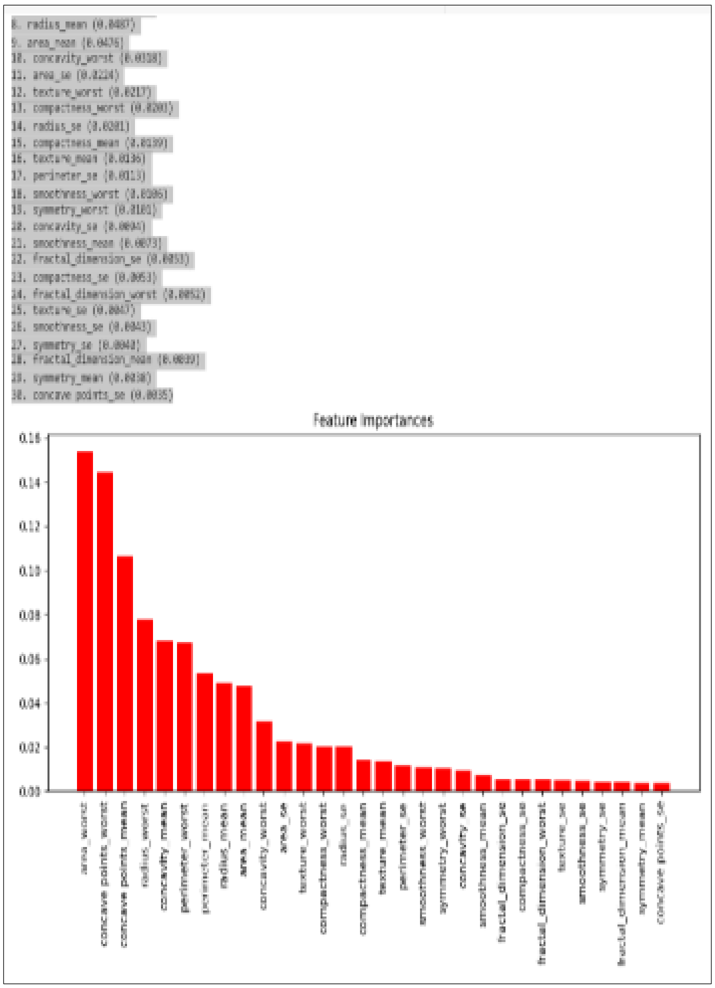
This chart shows which tumor features helped the model most when making predictions.

3.6.1. Top Features

- area_worst
- concave points_worst
- concave points_mean
- radius_worst
- perimeter_worst

3.6.2. Meaning

- Features with _worst show tumor characteristics in their most extreme (worst) form — these are more useful in spotting malignancy.
- Larger and more irregular tumors (more concave points, bigger radius/area) are likely to be malignant.
- The model focuses on tumor size and irregular shape to predict cancer.



```
8. radius_mean (0.0487)
9. area_mean (0.0476)
10. concavity_worst (0.0318)
11. area_se (0.0224)
12. texture_worst (0.0217)
13. compactness_worst (0.0201)
14. radius_se (0.0201)
15. compactness_mean (0.0139)
16. texture_mean (0.0136)
17. perimeter_se (0.0113)
18. smoothness_worst (0.0106)
19. symmetry_worst (0.0101)
20. concavity_se (0.0094)
21. smoothness_mean (0.0073)
22. fractal_dimension_se (0.0053)
23. compactness_se (0.0053)
24. fractal_dimension_worst (0.0052)
25. texture_se (0.0047)
26. smoothness_se (0.0043)
27. symmetry_se (0.0040)
28. fractal_dimension_mean (0.0039)
29. symmetry_mean (0.0038)
30. concave points_se (0.0035)
```

Figure 9 Feature importances

3.7. area_worst – Tumor Area (Worst Case)

3.7.1. Histogram

- Benign tumors: Mostly have area_worst around 400–600.
- Malignant tumors: Often have area_worst values above 1000.

3.7.2. Boxplot

- Malignant tumors show a higher median, more variation, and some extreme values.
- Benign tumors are tightly packed at lower values.

3.7.3. Conclusion: Larger tumor area is strongly linked to malignancy.

area_worst is a powerful indicator of cancer.

3.8. concave points_worst – Most Irregular Points on Tumor Edge

3.8.1. Histogram

- Malignant tumors: Concentrated between 0.15–0.25.
- Benign tumors: Mostly below 0.1.

3.8.2. Boxplot

- Clear separation in medians.

- Malignant tumors have a wider range.
- Conclusion: Tumors with more concave points (more irregular shapes) are often malignant.

3.9. concave points_mean – Average Irregularity

3.9.1. Histogram

- Malignant: Values around 0.10–0.15.
- Benign: Mostly under 0.05.

3.9.2. Boxplot

- Clear visual difference in medians.
- Benign tumors show some outliers, but are mostly grouped low.
- Conclusion: This feature is also a strong predictor of malignancy.

3.10. radius_worst – Maximum Tumor Radius

3.10.1. Histogram

- Benign: Mostly below 17.
- Malignant: Spread across 17 and above.

3.10.2. Boxplot

- Malignant tumors have higher radius values, suggesting they tend to be larger.
- Conclusion: Tumor size (measured by radius) helps differentiate benign from malignant.

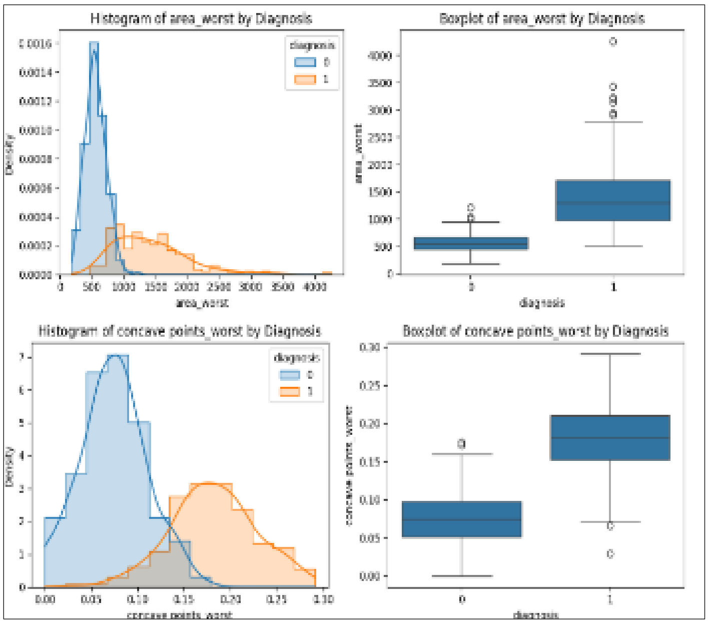
3.11. concavity_mean – Average Tumor Indentation Depth

3.11.1. Histogram

- Benign: Mostly between 0 and 0.05.
- Malignant: Peaks around 0.15.

3.11.2. Boxplot

- Malignant tumors: Higher medians, greater spread.
- Benign tumors: Low values and less variation.
- Conclusion: More deep indentations (concavity) usually mean a tumor is malignant.



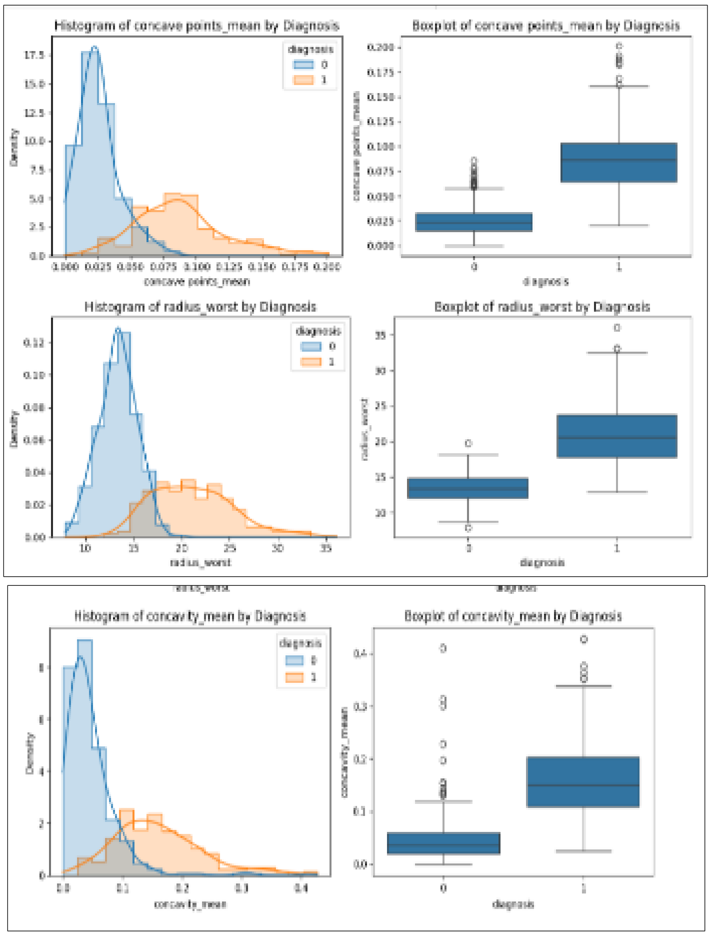
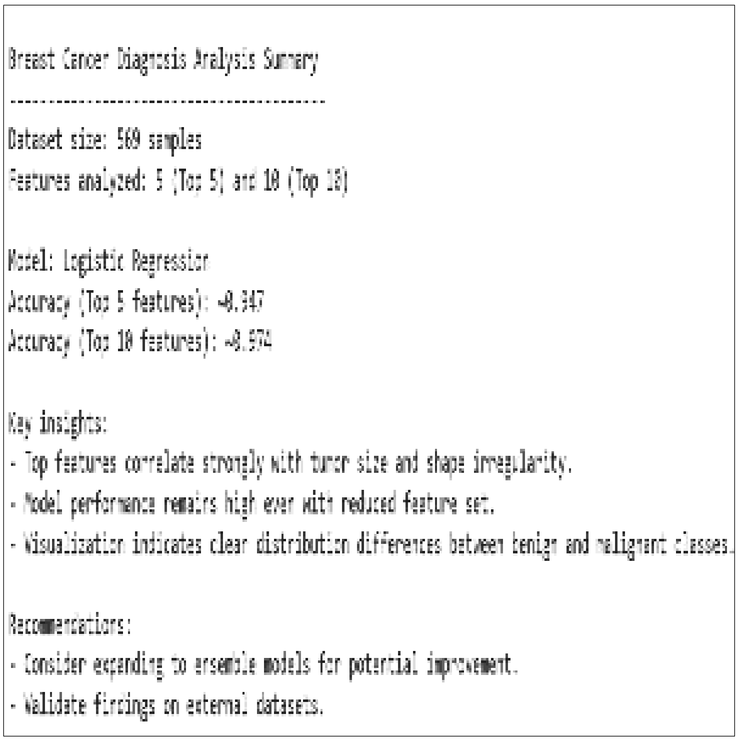


Figure 10 Histogram and box plot of features

4. Discussion

We have established that machine learning, specifically logistic regression, is effective when using breast cancer data based on tumor characteristics. We analysed a classic UCI Machine Learning Repository dataset. Even a reduced set of purposely selected features deliver very accurate results. Our models trained on the top 5 and top 10 most relevant features achieved good results. The logistic regression model using the top 10 features achieved an accuracy of ~97.4% compared to the ~94.7% accuracy using the top 5 features. This indicates the utility of providing models with as much useful features as possible without introducing too much noise into the model. The confusion matrix and classification report are reliable methods of providing evidence of the precision and recall of malignant and benign classes, respectively. The visualization tools, count plots, boxplots, and correlation heatmaps were useful in achieving intuitively easier understandings of which features were most important and how the features diagnosed malignant and benign classes.



```
Breast Cancer Diagnosis Analysis Summary
----------------------------------------
Dataset size: 569 samples
Features analyzed: 5 (Top 5) and 10 (Top 10)

Model: Logistic Regression
Accuracy (Top 5 features): ~0.947
Accuracy (Top 10 features): ~0.974

Key insights:
- Top features correlate strongly with tumor size and shape irregularity.
- Model performance remains high even with reduced feature set.
- Visualization indicates clear distribution differences between benign and malignant classes.

Recommendations:
- Consider expanding to ensemble models for potential improvement.
- Validate findings on external datasets.
```

5. Conclusion

Using logistic regression becomes a trustworthy and interpretable model for early detection of breast cancer when combined with appropriate preprocessing and feature selection. Future work could assess alternative methods through ensemble methods (e.g., Random Forest, Gradient Boosting) and perform external validation to confirm generalizability of the model. With additional development, these types of models could help clinicians become faster and more accurate in making diagnostic decisions.

Compliance with ethical standards

Acknowledgments

I would like to express my sincere gratitude to RV College of Engineering, Bengaluru, for providing the resources and support necessary for conducting this research.

I extend my heartfelt thanks to my faculty guide and mentors for their valuable guidance, encouragement, and constructive feedback throughout the project. Their insights greatly enhanced the quality of this work.

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] W. H. Wolberg, O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, 1990.

[2] S. Dua, U. R. Acharya, "Machine Learning in Healthcare Informatics", Springer, 2014.

[3] Elter, R. Schulz-Wendtland, T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process", Medical Physics, 2007.

[4] N. M. Nayak, "Diagnosis of breast cancer using logistic regression and SVM classifier", International Journal of Engineering and Technology, 2019.

[5] D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, 2005.

[6] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection", Journal of Machine Learning Research, 2003.

[7] L. Breiman, "Random forests", Machine Learning Journal, 2001.

[8] P. Cortez, A. Cerdeira, F. Almeida, "Modeling wine preferences by data mining from physicochemical properties", Decision Support Systems, 2009.

[9] T. Fawcett, "An introduction to ROC analysis", Pattern Recognition Letters, 2006.

[10] M. Kaur, S. Gosain, "Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise", International Journal of Computer Applications, 2013.

[11] S. Chaurasia, S. Pal, "A novel approach for breast cancer detection using data mining techniques", International Journal of Innovative Research in Computer and Communication Engineering, 2014.

[12] K. Polat, S. Güneş, "Breast cancer diagnosis using least square support vector machine", Digital Signal Processing, 2007.

[13] U. Akhil, T. D. Sudhakar, "Breast cancer detection using logistic regression", International Journal of Engineering Research and Applications, 2020.

[14] M. Sokolova, G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing & Management, 2009.

[15] P. S. Hiremath, B. S. Kolekar, "Feature extraction and classification of mammographic images using wavelet transform and SVM", International Journal of Computer Applications, 2010.

[16] B. Zadrozny, C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates", Proceedings of the Eighth ACM SIGKDD, 2002.

[17] L. Jiang, D. Wang, "An improved KNN text classification algorithm based on clustering", Journal of Computers, 2015.

[18] Y. Bengio, "Learning deep architectures for AI", Foundations and Trends in Machine Learning, 2009.

[19] M. Mohammadi, A. Al-Fuqaha, "Deep learning for IoT big data and streaming analytics: A survey", IEEE Communications Surveys & Tutorials, 2018.

[20] J. M. Gómez, A. M. Gómez, "Breast cancer diagnosis based on logistic regression", International Journal of Computer Applications, 2016.