Breast Cancer Malignancy Prediction Using Machine Learning

Group C Project for UCSC
Training course on Data Science
with Python

Abstract—

Early detection of breast cancer significantly improves patient outcomes. This study explores machine learning approaches to classify breast tumors as benign or malignant using the Wisconsin Breast Cancer Dataset (WBCD). We evaluate Logistic Regression, Decision Tree, Random Forest, Neural Network, and Support Vector Machine models after standard preprocessing. The Support Vector Machine (SVM) achieved the best performance with an accuracy of 97.3% and an ROC AUC of 0.9841. Feature analysis identified radius_mean, concavity_mean, and area_mean as top predictors. The results indicate that ML models can provide accurate, non-invasive decision support for early breast cancer detection.

Index Terms— Breast cancer, machine learning, SVM, Random Forest, Wisconsin Breast Cancer Dataset, classification

I. INTRODUCTION

Breast cancer is a leading cause of cancer-related mortality among women worldwide. Early and accurate detection increases the chances of successful treatment. Traditional diagnostic techniques such as mammography and biopsies can be invasive, costly, or require specialist interpretation. Machine learning (ML) offers a complementary, data-driven approach to support clinicians by providing fast and consistent risk stratification.

II. PROBLEM DEFINITION

This work focuses on the binary classification task: predicting whether a tumor is benign or malignant. The aim is to develop and compare multiple ML models that can assist in early diagnosis and reduce false negatives.

III. DATASET USED

We used the Wisconsin Breast Cancer Dataset (WBCD) from the UCI Machine Learning Repository, comprised of 569 patient records with 30 real-valued features computed from digitized images of fine needle aspirate (FNA) of breast masses. Common features include radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, and concavity_mean. The target variable encodes diagnosis as M (malignant) or B (benign). Data preprocessing included cleaning, scaling (standardization), and label encoding prior to model training.

IV. METHODS

Five supervised learning models were implemented and evaluated: Logistic Regression, Decision Tree, Random Forest, a feedforward Neural Network, and Support Vector Machine (SVM). Hyperparameters for each model were tuned using cross-validation. Evaluation metrics used were accuracy, precision, recall, F1-score, and ROC AUC. Feature importance was examined for tree-based models, and coefficients/weights were inspected for linear models to identify relevant predictors.

V. RESULTS AND DISCUSSION

Among the models tested, the SVM achieved the highest accuracy at 97.3% and an ROC AUC of 0.9841. Precision, recall, and F1-score were approximately balanced around 97% for the top-performing model. Tree-based models (Random Forest and Decision Tree) showed strong performance and enabled feature importance analysis, which highlighted radius_mean, concavity_mean, and area_mean as the most predictive features.

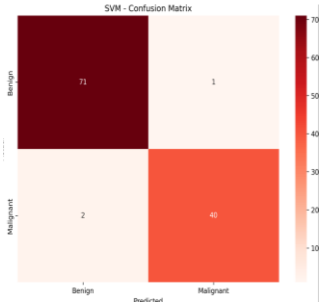

Fig 1 : Confusion Matrix for SVM

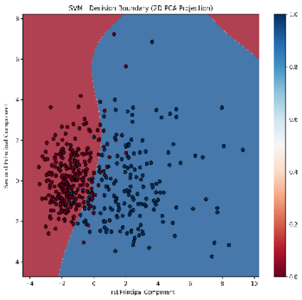

Fig 2 : Decision Boundary Curve for SVM

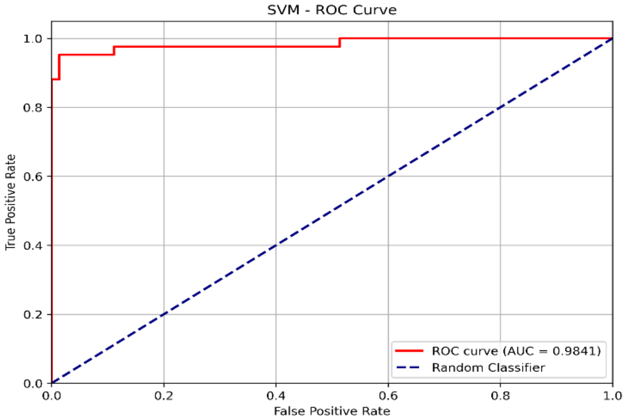

Fig 3 : Receiver operating characteristic (ROC) Curve for SVM

The high performance of SVM indicates its suitability for margin-based separation in this dataset; however, model selection should consider interpretability and deployment constraints in clinical settings. Tree-based models offer easier interpretability via feature importance, while neural networks and ensemble methods may offer gains on larger, more diverse image datasets. Model calibration and external validation on diverse populations are necessary before clinical deployment.

VI. CONCLUSION

This study demonstrates that machine learning models, particularly SVM, can accurately classify breast tumors using the WBCD features. Such models have potential as non-invasive decision-support tools to help prioritize patients for further testing. Future work will focus on image-based deep learning, explainable AI techniques (e.g., SHAP, LIME), expanding training data to improve generalization, and integrating models into clinical workflows for prospective testing.

VII. FUTURE ENHANCEMENTS AND RESEARCH DIRECTIONS

Future Enhancements:

- Employ convolutional neural networks (CNNs) for image-based analysis.
- Apply explainable AI methods such as SHAP and LIME to improve interpretability.
- Develop hybrid SVM–Neural Network architectures and ensembled models.
- Train on larger, more diverse datasets for robust generalization.
- Build real-time diagnostic tools with continuous learning and model monitoring.

ACKNOWLEDGMENT

Acknowledgements — The authors would like to thank peers and mentors who provided feedback during model development. (Group C)

REFERENCES

[1] D. Dua and C. Graff, UCI Machine Learning Repository, 'Wisconsin Diagnostic Breast Cancer (WDBC) Data Set'.

[2] F. Pedregosa et al., 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, 2011.

[3] G. James et al., 'An Introduction to Statistical Learning', Springer, 2013.