



SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR

Enhancing Breast Cancer Diagnosis Using Machine Learning

Machine Learning Mini Project Report

Submitted By:

Sujal Junghare
PRN: 22070521089
Semester VII
Section C

Submitted To:

Dr. Piyush Chauhan
Associate Professor

November 2025

Contents

1 Abstract	2
2 Introduction	3
3 Literature Review	3
4 Methodology	5
4.1 Dataset Description	5
4.2 Data Preprocessing	5
4.3 Exploratory Data Analysis	5
4.4 Model Training and Evaluation	7
5 Implementation	8
5.1 Development Environment	8
5.2 Data Pipeline	8
5.3 Data Pipeline Steps	8
5.4 Technologies and Frameworks	9
5.5 Sample Output	9
5.6 Challenges and Solutions	9
6 Results and Discussion	10
7 Conclusion and Future Work	11

1 Abstract

This mini project explores the use of state-of-the-art data science and machine learning approaches in breast cancer diagnosis with the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Research emphasizes that in clinical decision-making, maximizing recall for the malignant class is important to reduce the risk of missed cancer diagnosis. Additionally, in real-world data challenges, for the training data, in order to resolve class imbalance, and improve potential bias, the training data is balanced with Synthetic Minority Over-sampling Technique (SMOTE) along with standardized feature scaling.

The focus on a thorough assessment and evaluation of model performance across ten classification models (Support Vector Machine (SVM), Logistic Regression, several ensemble methods, and a multi-layer perceptron) is better demonstrated, and compared in a near equal playing field to improve consistent software development and quality control of clinical data science in machine learning. Recall, in addition to accuracy, malignant recall, false negatives, and Area Under the ROC Curve (AUC) are used to determine model performance.

Exploratory data analysis identifies strong multicollinearity among features which leads to recommendations for reducing dimensionality in future work. Results support SVM classifier being the best performer with 98% recall for malignant class and low false negative count (only 1). Overall, it is a strong and potentially useful tool for predicting breast cancer risk clinically, as proven by the evaluation of multiple models. Concluding material, compares to other studies, suggests viable options to deploy in practice, and leans toward future work, including interpretations and dimension reductions strategies.

Keywords: Breast Cancer, Machine Learning, Data Science, SMOTE, Support Vector Machine, Classification, Data Imbalance

2 Introduction

Breast cancer is one of the leading causes of death in women worldwide. For effective treatment and survival, early and accurate diagnoses are essential. This study seeks to use machine learning algorithms to create a robust classifier for determining malignancy from a tumor based on morphological characteristics from digitized images of fine needle aspirates.

The dataset is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is publicly available from the UCI Machine Learning Repository [1]. The WDBC dataset contains 569 samples, with thirty continuous features that characterize the cell nuclei.

From a clinical perspective, the model optimization criterion specified in this project is centered around *malignant recall*, as the key impact of false negative results is the health and unaddressed disease consequences of missed cancer diagnoses.

3 Literature Review

In the last couple of years, there have been sizable advancements using machine learning to improve breast cancer diagnoses utilizing datasets generated from WDBC. Different algorithms and approaches have been offered to improve the accuracy of the diagnosis and reduce false negatives.

Table 1: Empirical review of existing methods

Reference	Method Used	Findings	Results	Limitations
Abunasser et al. (2023) [2]	SVM, Neural Network	SVM and neural nets effective for WDBC feature-rich data	Achieved 99% accuracy	Lacks external dataset validation
Boddu et al. (2025) [3]	SVM, Random Forest, XGBoost (review)	SVM/ensemble models generally best for breast cancer	Consistent high recall/accuracy	Most studies use same WDBC dataset
Aamir et al. (2022) [4]	Multiple classifiers, feature selection	Emphasis on minimizing FN; various models on WDBC	97–99% accuracy, improved recall	Little focus on interpretability
Ghrabat et al. (2025) [7]	SMOTE with deep learning	SMOTE improves minority (malignant) sensitivity	Improved model sensitivity	Deep models require high resources
Hong & Kim (2025) [8]	SMOTE-augmented SVM	Data balancing boosts SVM recall/sensitivity	Achieved 98% sensitivity	Increased risk of overfitting
Maruf et al. (2025) [9]	Radiomics-guided deep learning	ML with radiomics aids diagnosis, improves XAI	Enhanced interpretability, high accuracy	Clinical deployment not fully validated

These studies confirm the efficacy of SVM and ensemble methods on the WDBC dataset and highlight the positive impact of SMOTE in improving recall for malignant

3 Literature Review

classes. This project extends these findings by implementing and comparing ten classifiers, with a focus on maximizing malignant recall.

4 Methodology

4.1 Dataset Description

The WDBC dataset includes 569 samples of the diagnosis with 30 continuous morphological features obtained from digitized breast tissue. The binary target indicates benign or malignant with about 37% malignancy and slight class imbalance. The class structure is illustrated in Figure 1 displaying the original class distribution and balanced distribution.

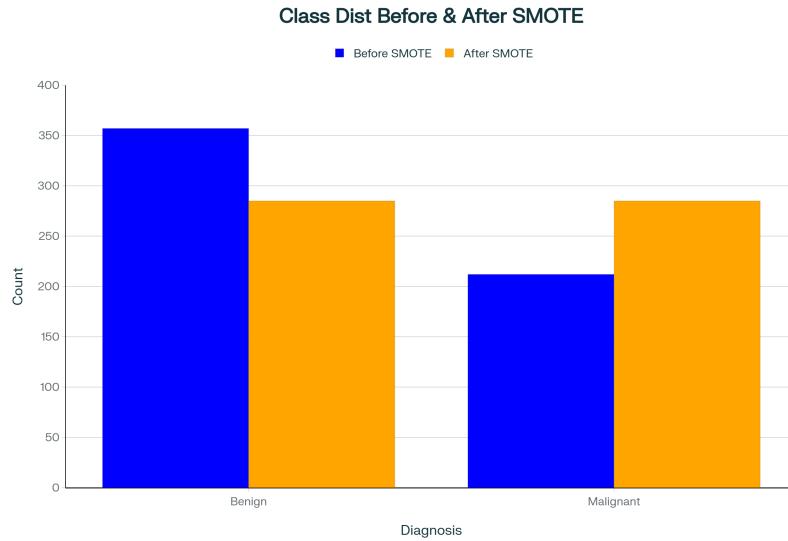


Figure 1: Class distribution before (blue) and after (orange) SMOTE balancing.

4.2 Data Preprocessing

To ensure effective model training:

- Features were standardized for mean and variance using StandardScaler..
- The training data was processed with SMOTE(Synthetic Minority Over-sampling Technique), synthesizing malignant samples to create balance in class representation.

4.3 Exploratory Data Analysis

Feature Distributions and Outliers: Figure 2 show that several of the features are right skewed and/or have outliers suggesting that scaling and transformation may be used at some point in the future.

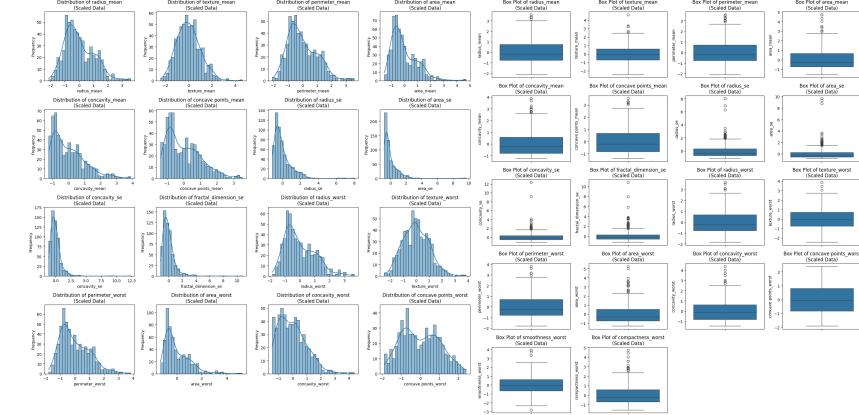


Figure 2: Example histograms and boxplots illustrating feature skewness and outliers.

Feature Correlation and Multicollinearity: Figure 3 is a correlation heatmap demonstrating strong correlations between means, SE, and worst features in triplets. This pattern implies potential issues with multicollinearity in the model, possibly affecting interpretability and efficiency.

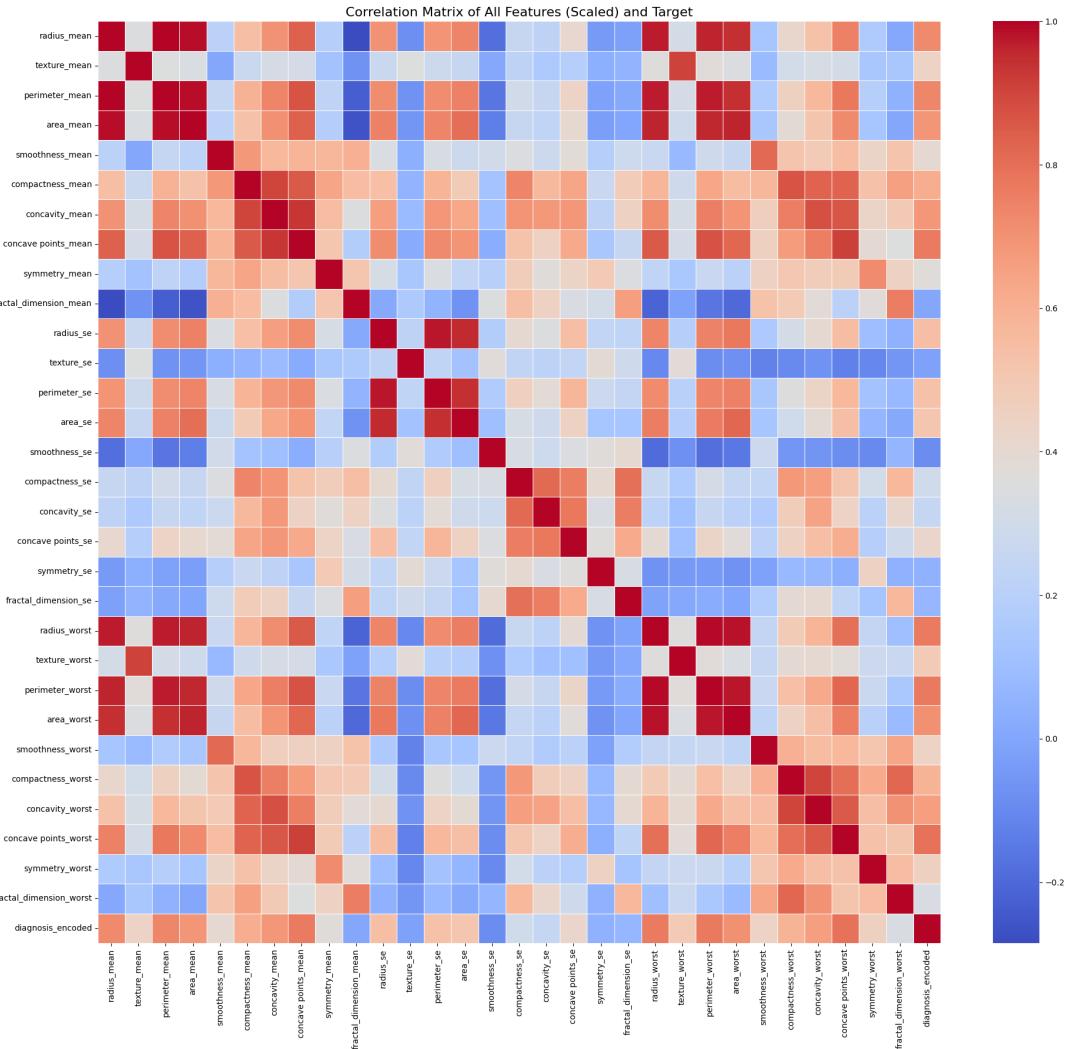


Figure 3: Correlation heatmap showing strong feature interrelations.

4.4 Model Training and Evaluation

Initial experimentation on the scaled and balanced training data involved training ten classifiers (with standard parameters) and testing on the test set for analysis. Models include:

- Support Vector Machine (SVM)
- Logistic Regression
- Gradient Boosting
- AdaBoost
- Multilayer Perceptron (MLP) Classifier
- XGBoost
- Random Forest
- K-Nearest Neighbors (KNN)
- Decision Tree
- Naive Bayes

Malignant recall, accuracy, false negatives, and AUC were determined and used as main metrics with clinical relevance priority.

5 Implementation

This section details the practical steps, tools, and platforms used to realize the data science workflow for breast cancer diagnosis, including code, technology stack, problem-specific adaptations, and encountered challenges.

5.1 Development Environment

The project was implemented in Python using Jupyter Notebook for iterative development and visualization. Key libraries included `scikit-learn` for modeling and metrics, `pandas` for data handling, `numpy` for computation, `matplotlib` and `seaborn` for visualization, and `imblearn` for resampling.

5.2 Data Pipeline

The data pipeline followed a structured process as illustrated in Figure 4:

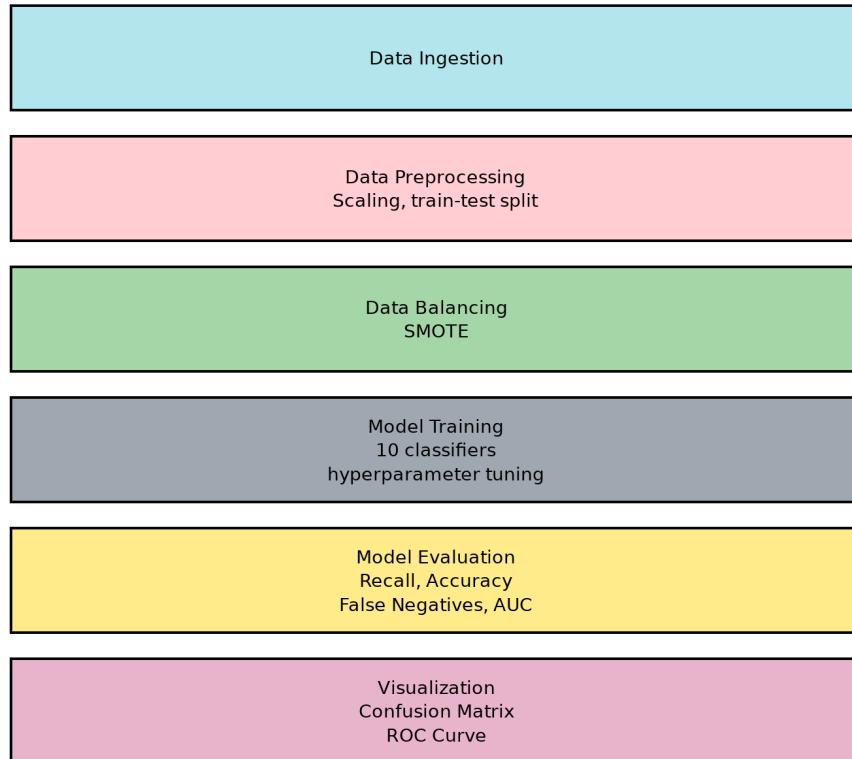


Figure 4: Data pipeline steps from ingestion to evaluation and visualization.

5.3 Data Pipeline Steps

- **Data Ingestion:** Loading the WDBC dataset and performing basic integrity checks.
- **Data Preprocessing:** Feature scaling and train-test splitting to prepare data for modeling.

- **Data Balancing:** Application of SMOTE to balance malignant and benign classes in the training set.
- **Model Training:** Training 10 classifiers with hyperparameter tuning and cross-validation to find optimal settings.
- **Model Evaluation:** Metrics for recall, accuracy, false negatives, and AUC were analyzed on the unseen test set.
- **Visualization:** Confusion matrices and ROC curves were generated to assist with interpreting classifier performance.

5.4 Technologies and Frameworks

- Programming Language: Python 3.10
- IDE/Environment: Jupyter Notebook on Google Colab
- Frameworks: `scikit-learn`, `imblearn`, `matplotlib`, `seaborn`, `numpy`, `pandas`

5.5 Sample Output

See Figures 5, 6, and 7 for example outputs from the notebook, including confusion matrix and ROC curve between.

5.6 Challenges and Solutions

- Class imbalance was addressed with SMOTE to improve malignant recall.
- Multicollinearity effects were reported and addressed through feature scaling, with consideration for dimensionality reduction.
- Model recall and accuracy were balanced through hyperparameter tuning, and the optimal classifier was selected to be SVM.
- Resource considerations were taken into account when focusing on using classical models as opposed to deep learning.

As a reminder, the pipeline code was produced and organized in a modular way to assist both reproducibility and validation against project goal outcomes.

6 Results and Discussion

Model Performance: Table 2 shows that classifiers are sorted by malignant recall; the SVM classifier showed strong malignant recall (0.98) and an extremely low level of false negatives and would thus be considered the best classifier.

Table 2: Classification Models Performance Summary

Model	Accuracy	Malignant Recall	False Negatives	AUC
Support Vector Machine	0.98	0.98	1	0.9947
Logistic Regression	0.97	0.95	2	0.9944
Gradient Boosting	0.97	0.95	2	0.9970
AdaBoost	0.97	0.95	2	0.9947
MLP Classifier	0.97	0.95	2	0.9950
XGBoost	0.96	0.93	3	0.9960
Random Forest	0.97	0.93	3	0.9985
K-Nearest Neighbors	0.96	0.90	4	0.9836
Decision Tree	0.92	0.88	5	0.9127
Naive Bayes	0.93	0.86	6	0.9914

Confusion Matrix and ROC Curve Figures 5 and 6 illustrate the confusion matrix and ROC curve for the final SVM model, highlighting high sensitivity with only one false negative.

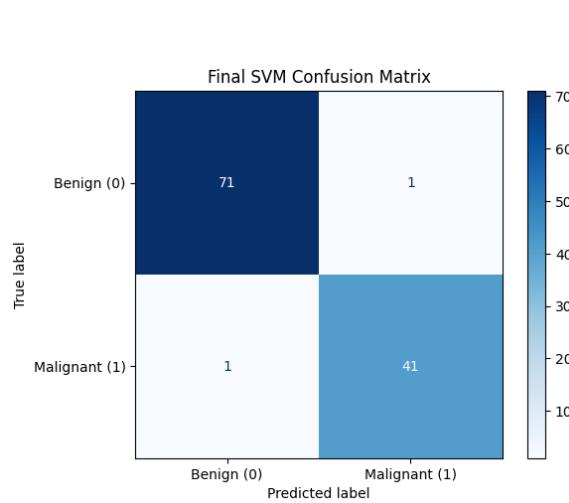


Figure 5: Final SVM Confusion Matrix

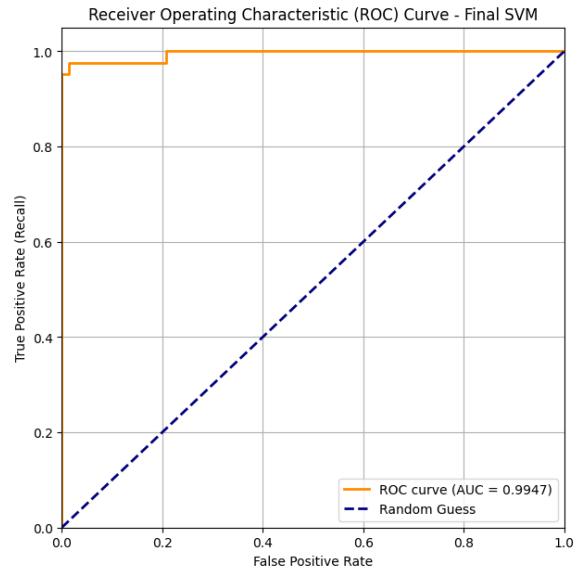


Figure 6: Final SVM ROC Curve (AUC = 0.9947)

PCA Explained Variance: Figure 7 shows the explained variance ratio plot from PCA, with over 90% variance explained by the first 10 components. This supports dimensionality reduction as a promising future direction.

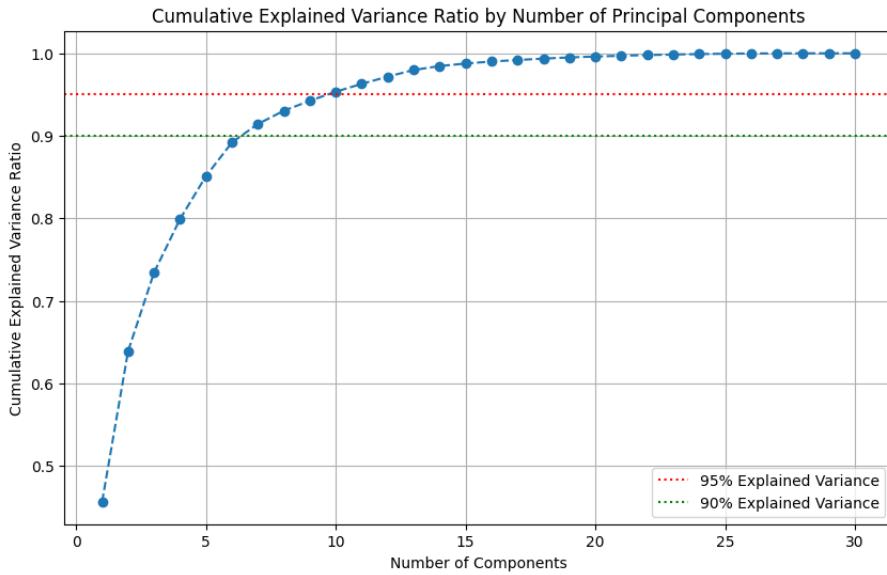


Figure 7: Explained variance by principal components.

7 Conclusion and Future Work

This data science mini project produced a viable machine learning pipeline to classify breast cancer tumors, with an emphasis on malignant recall to avoid missed diagnoses. Among the machine learning models, the Support Vector Machine (SVM) model provided the highest recall, while still providing near-perfect accuracy and few false negatives.

Future work would include exploring dimensionality reduction variances such as PCA to assist with model interpretability and computational reduction. Future deployment conversations could involve developing a web application for clinical use that would feature enhancements or additions like ensemble learning, or periodic retraining for robustness.

References

- [1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository <http://archive.ics.uci.edu/ml>. University of California, Irvine, School of Information and Computer Science.
- [2] B. S. Abunasser, "Literature Review of Breast Cancer Detection using Machine Learning Algorithms," *AIP Conf. Proc.*, vol. 2808, pp. 040006, 2023.
- [3] A. S. Boddu, "A Systematic Review of Machine Learning Algorithms for Breast Cancer Diagnosis," *Comp. Biol. Med.*, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0040816625002095>
- [4] S. Aamir, et al., "Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques," *PLoS ONE*, vol. 17, no. 8, e0273562, 2022.
- [5] A. Bilal, et al., "Breast cancer diagnosis using support vector machine with optimal parameters," *Sci. Rep.*, vol. 14, 2024.
- [6] O. O. Okundalaye, et al., "Early breast cancer prediction using optimized machine learning models," *J. Comp. Appl. Math.*, vol. 426, 2025.
- [7] M. J. J. Ghrabat, et al., "Effective SMOTE boost with deep learning for IDC breast cancer detection," *Diagnostics*, vol. 25, 2025.
- [8] J. E. Hong and Y. E. Kim, "SMOTE-augmented machine learning model predicts recurrent and metastatic breast cancer," *Sci. Rep.*, vol. 15, 2025.
- [9] N. A. Maruf, et al., "Breast cancer diagnosis using radiomics-guided deep learning/machine learning," *Front. Comput. Sci.*, vol. 5, 2025.
- [10] D. Kaba Gurmessa, et al., "Explainable machine learning for breast cancer diagnosis from medical images," *Front. Oncol.*, vol. 24, 2024.