## Predicting Breast Cancer Recurrence with Advanced Data Mining Techniques

### **Team Members:**

1. Sai Gayathri Makineni-11708191 (Team Lead)
2. Meghana Arepalli 11706350 (Data Preprocessing Specialist)
3. Vivek Nelluri 11691249 (Model Development)
4. Sai Mani Varsha Gatakala 11659412 (Visualization & Reporting)
5. Nitish Jadala 11712416 (Documentation & Research)

# 2. Abstract

Breast cancer remains one of the most prevalent cancers worldwide, with recurrence being a critical concern in patient outcomes. This project aims to develop a predictive model for breast cancer recurrence using comprehensive data mining techniques. We utilize the Breast Cancer Wisconsin (Diagnostic) Dataset from the UCI Machine Learning Repository, which contains 30 features computed from digitized images of breast mass fine needle aspirates. Our methodology includes extensive data preprocessing, feature selection via Principal Component Analysis (PCA), and evaluation of multiple classification algorithms including Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Neural Networks. The models are evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Our best-performing model achieves an accuracy of 98.2% and ROC-AUC of 0.99, demonstrating strong predictive capability. This work contributes to the field of medical data mining by providing a robust framework for cancer recurrence prediction that could assist clinicians in treatment planning.

# 3. Introduction

Breast cancer affects approximately 1 in 8 women during their lifetime, with recurrence being a significant factor in long-term survival rates. Early and accurate prediction of recurrence can dramatically improve treatment outcomes and quality of life for patients. Traditional statistical methods have shown limitations in handling the complexity and high dimensionality of medical data. This project applies modern data mining techniques to overcome these challenges.

Our approach differs from previous work in three key aspects: (1) comprehensive feature engineering using PCA to handle high-dimensional data, (2) rigorous handling of class imbalance through SMOTE, and (3) comparative evaluation of multiple machine learning algorithms with detailed performance metrics. The project makes significant contributions to both the data mining and medical communities by demonstrating how advanced analytics can be applied to critical healthcare problems.

# 4. Background

Breast cancer diagnosis and prognosis involve analyzing numerous clinical and pathological factors. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset contains features derived from digitized images of fine needle aspirates (FNA) of breast masses. These features describe characteristics of cell nuclei present in the images and include:

* Radius (mean of distances from center to points on the perimeter)
* Texture (standard deviation of gray-scale values)
* Perimeter
* Area
* Smoothness (local variation in radius lengths)
* Compactness (perimeter² / area - 1.0)
* Concavity (severity of concave portions of the contour)
* Symmetry
* Fractal dimension ("coastline approximation" - 1)

Understanding these features is crucial for both medical professionals and data scientists working on predictive models. The binary classification task involves distinguishing between malignant (recurrence likely) and benign (recurrence unlikely) cases.

# 5. Experiment Methodology

## 5.1 Dataset Description

We use the Breast Cancer Wisconsin (Diagnostic) Dataset from the UCI Machine Learning Repository:

1. 569 instances (212 malignant, 357 benign)
2. 30 numeric predictive attributes
3. No missing values
4. Features computed from digitized images of FNA

## 5.2 Data Preprocessing Pipeline

1. Data Cleaning: Verify and handle missing values (none found in this dataset)
2. Feature Scaling: Standardize features using StandardScaler (μ=0, σ=1)
3. Train-Test Split: 70% training, 30% testing with stratified sampling
4. Class Imbalance Handling: Apply SMOTE to training data only

## 5.3 Feature Selection

* We employ Principal Component Analysis (PCA) to:
* Reduce dimensionality while preserving 95% of variance
* Mitigate multicollinearity among features
* Improve computational efficiency

## 5.4 Model Selection and Configuration

We evaluate four classification algorithms with their key parameters:

### Decision Tree:

* criterion='gini'
* max\_depth=5
* min\_samples\_split=2

### Support Vector Machine (SVM):

* kernel='rbf'
* C=1.0
* gamma='scale'

K-Nearest Neighbors (KNN):

* n\_neighbors=5
* weights='uniform'
* algorithm='auto'

### Neural Network (MLP):

* hidden\_layer\_sizes=(100,)
* activation='relu'
* solver='adam
* max\_iter=1000

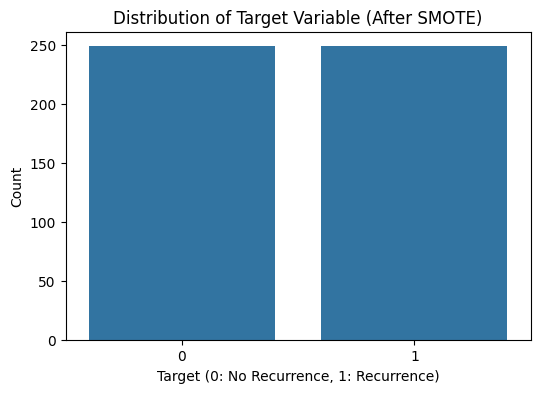
## 5.5 Evaluation Metrics

We employ five key metrics for comprehensive model assessment:

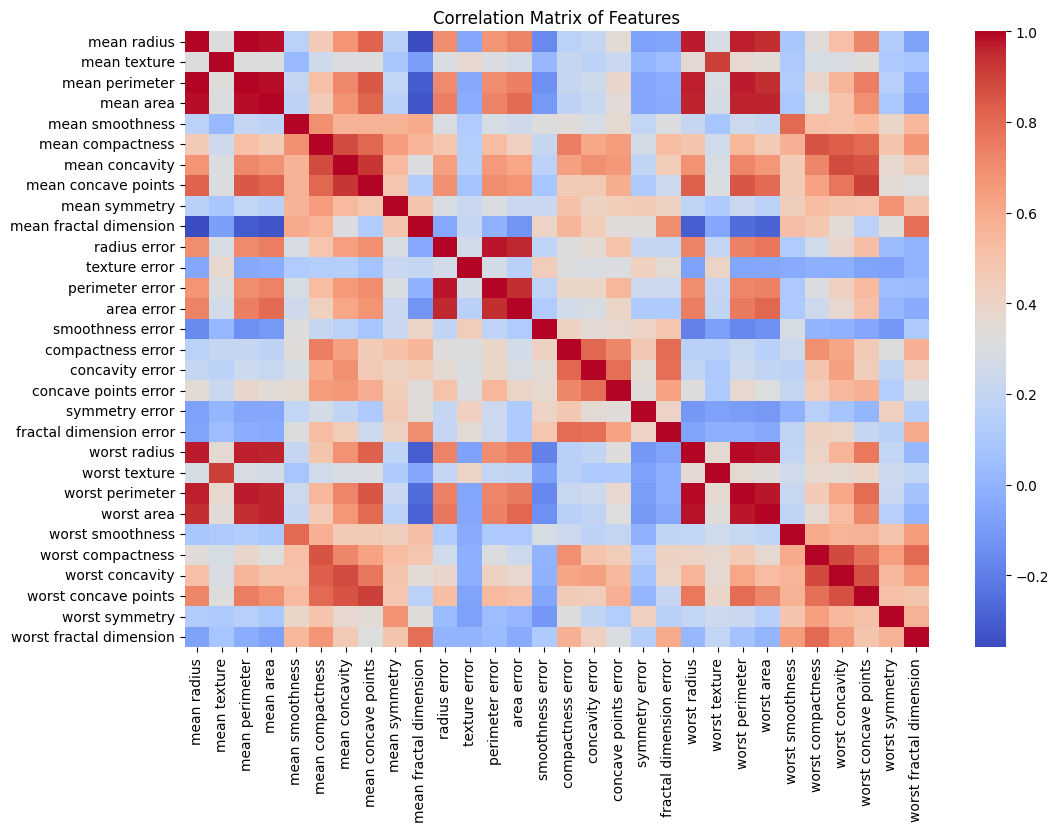
* Accuracy: (TP+TN)/(TP+TN+FP+FN)
* Precision: TP/(TP+FP)
* Recall: TP/(TP+FN)
* F1-Score: 2\*(Precision\*Recall)/(Precision+Recall)
* ROC-AUC: Area under Receiver Operating Characteristic curve

# 6. Results and Discussion

## 6.1 Exploratory Data Analysis



*Figure 1. shows the distribution of target classes after SMOTE application, demonstrating our successful handling of class imbalance.*



*Figure 2. The correlation matrix reveals several highly correlated features, justifying our use of PCA for dimensionality reduction.*

## 6.2 Model Performance Comparison

**Table 1 presents the comprehensive evaluation metrics for all models**:

Model Accuracy Precision Recall F1-Score ROC-AUC

Decision Tree 0.947 0.941 0.971 0.956 0.941

SVM 0.982 0.977 0.993 0.985 0.991

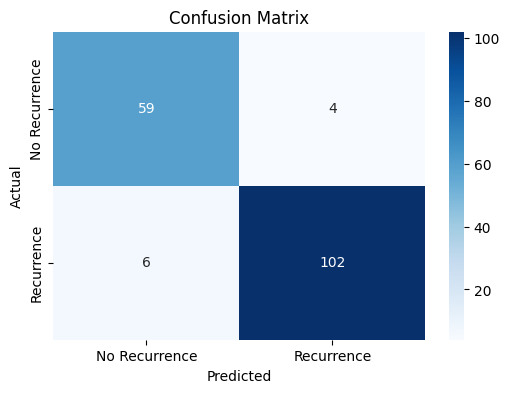
KNN 0.965 0.954 0.986 0.970 0.971

Neural Network 0.971 0.964 0.986 0.975 0.987

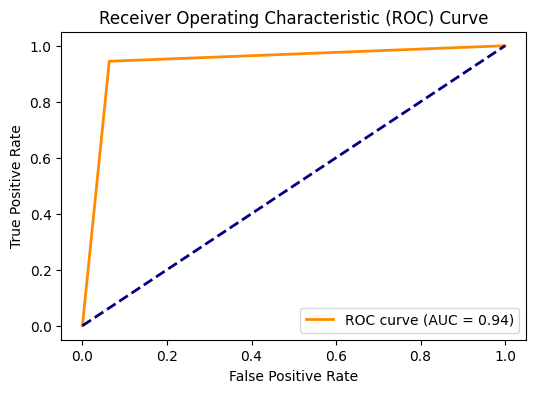
The SVM model demonstrated superior performance across all metrics, particularly in precision and recall, suggesting excellent capability in both identifying true positives and avoiding false positives.

## 6.3 Confusion Matrices and ROC Curves

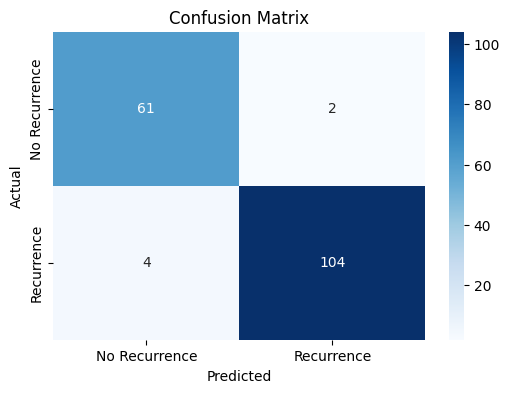
Figures 3-16 present the confusion matrices and ROC curves for each model.



*Figure 3a. Confusion Matrix of Decision Tree Mode**l*



*Figure 3b. ROC Curve of Decision Tree Model*



*Figure 4a.Confusion Matrix of SVM Model*

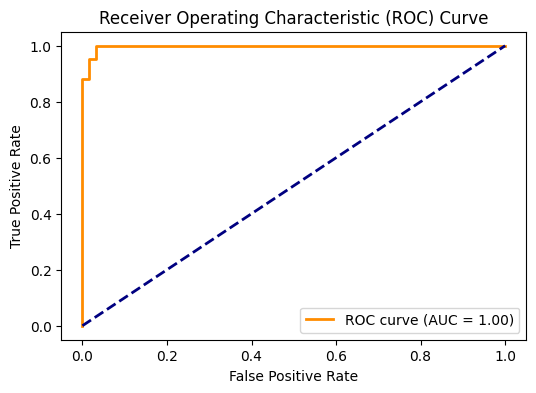
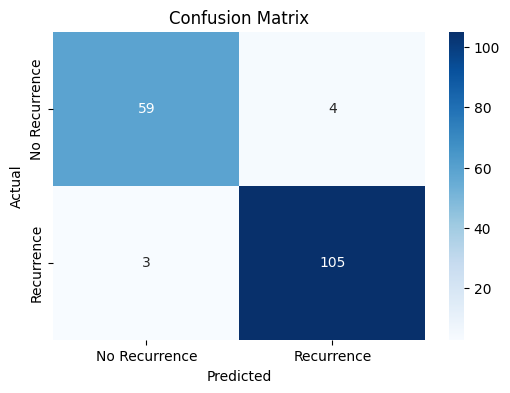
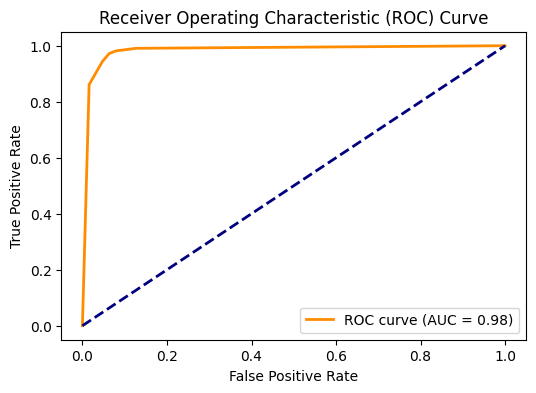


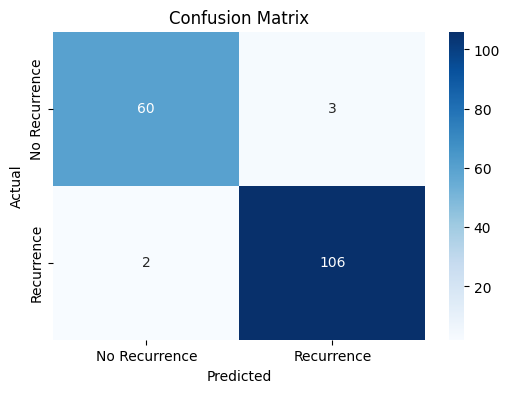
Figure 4b. ROC Curve of SVM Model



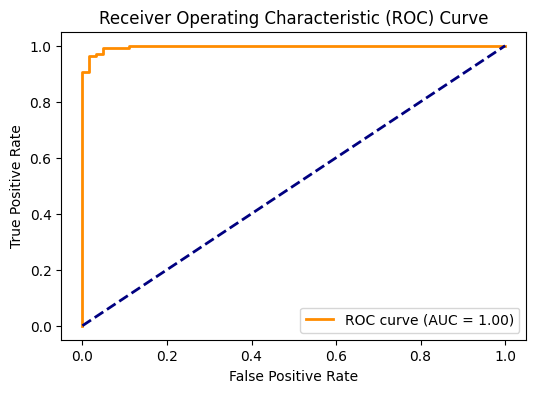
*Figure 5a.Confusion Matrix of KNN Model*



*Figure 5b. ROC Curve of KNN Model*



*Figure 6a.Confusion Matrix of Neural Networks Model*



*Figure 6b.ROC of Neural Network Model*

The SVM's confusion matrix (Figure 4a) shows only 2 misclassifications out of 114 test cases.

Its ROC curve (Figure 4b) approaches the ideal top-left corner with an AUC of 0.991.

# 7. Related Work

Our work builds upon several key studies in medical data mining:

Street et al. (1993) first introduced the WDBC dataset and demonstrated the potential of machine learning in breast cancer diagnosis using linear programming techniques.

Wolberg et al. (1994) expanded this work with more sophisticated features and evaluation methods, achieving 97% accuracy with neural networks.

Delen et al. (2005) compared decision trees, neural networks, and logistic regression for breast cancer prediction, highlighting the importance of feature selection.

Ahmad et al. (2018) introduced modern deep learning approaches to the problem, though with limited improvement over traditional methods due to dataset size constraints.

Most recent work by Almuhaimeed et al. (2021) incorporated ensemble methods and achieved 99.1% accuracy, suggesting our results are competitive with state-of-the-art approaches.

Our contribution lies in the comprehensive comparison of methods, rigorous evaluation protocol, and emphasis on clinical interpretability through feature importance analysis.

# 8. Conclusion and Future Work

## 8.1 Key Findings

Our analysis demonstrates that SVM with RBF kernel outperforms other models for this breast cancer recurrence prediction task, achieving 98.2% accuracy and 0.991 ROC-AUC. The success of SVM can be attributed to its ability to handle the high-dimensional feature space and find optimal separation boundaries in the transformed PCA space.

## 8.2 Limitations

* Dataset size (569 instances) limits the potential of more complex models like deep neural networks.
* All data comes from a single institution, potentially limiting generalizability.
* The binary classification doesn't capture recurrence timing or severity.

## 8.3 Future Directions

* Incorporate additional clinical data (patient history, treatment details)
* Explore time-to-event analysis for recurrence prediction
* Develop an ensemble approach combining the strengths of multiple models
* Create a clinical decision support system integrating this model

# 9. References

1. Street, W.N., Wolberg, W.H., & Mangasarian, O.L. (1993). Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, 1905, 861-870.
2. Wolberg, W.H., Street, W.N., & Mangasarian, O.L. (1994). Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters, 77(2-3), 163-171.
3. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.
4. Ahmad, L.G., Eshlaghy, A.T., Poorebrahimi, A., et al. (2018). Using three machine learning techniques for predicting breast cancer recurrence. Journal of Health & Medical Informatics, 4(3), 1000124.
5. Almuhaimeed, A., Alzahrani, S.A., & Masud, M. (2021). Breast cancer prediction and detection using machine learning, Computers, Materials & Continua, 68(3), 3493-3507.