# Predictive Analytics for Breast Cancer Survival and Metastasis Prediction Using Machine Learning Models

A Comparative Analysis of Predictive Models


**Submitted by:-**

Shruti Raj-22051892

Vikas Priyadarshi-22053651

# Introduction:-

- Breast cancer is the most common cancer among women globally, accounting for a significant number of cancer-related deaths.
- The progression of breast cancer can vary significantly from patient to patient, making early detection and accurate prognosis essential.

## Importance of Early Detection and Prediction:

- Early detection of breast cancer dramatically improves the chances of successful treatment and increases survival rates.
- Accurate prediction of survival and metastasis can help in tailoring personalized treatment plans, ensuring better patient outcomes.

# Objective of the Study:-

## Primary Objective:

To apply predictive analytics and machine learning models to predict breast cancer survival and metastasis.

### Specific Goals:

Compare the performance of different machine learning algorithms in predicting survival and metastasis.

Improve the accuracy of predictions through hyperparameter tuning.

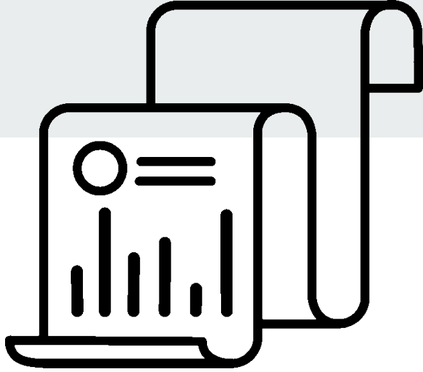
### Models Evaluated:

- Support Vector Machine (SVM)
- Gradient Boosting
- Random Forest
- K-Nearest Neighbors (KNN)
- Logistic Regression

### Dataset Overview:

The study was conducted on a dataset comprising **286** instances, divided into two classes: survival (201 cases) and metastasis (85 cases).

# Data Preprocessing:-



**1. Handling Missing Values:**

Missing data was imputed using the median for continuous variables and the mode for categorical variables.

**2. Normalization:**

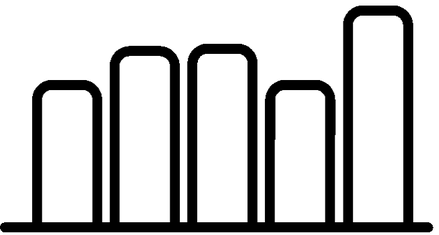
Continuous features (e.g., age, tumor size) were normalized to a uniform scale to ensure fair comparison across models.

**3. Encoding Categorical Variables:**

Binary categorical variables (e.g., hormone receptor status) were converted into numerical format using label encoding.

**4. Handling Class Imbalance:**

The Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the representation of the two classes (survival and metastasis) in the training data.



# Machine Learning Models Implemented:-

## Support Vector Machine (SVM):

Effective for high-dimensional data.
RBF kernel was used to capture non-linear relationships.

## Gradient Boosting:

An ensemble method that builds sequential models to minimize error.
Strong at handling imbalanced datasets and complex patterns.

- Each model was fine-tuned using GridSearchCV to optimize performance.

## Logistic Regression:

A statistical model used for binary classification.
Maps feature combinations to probabilities using the sigmoid function.

## Random Forest:

An ensemble technique that reduces overfitting by averaging predictions from multiple decision trees.
Key hyperparameters (number of trees, tree depth) were optimized.

## K-Nearest Neighbors (KNN):

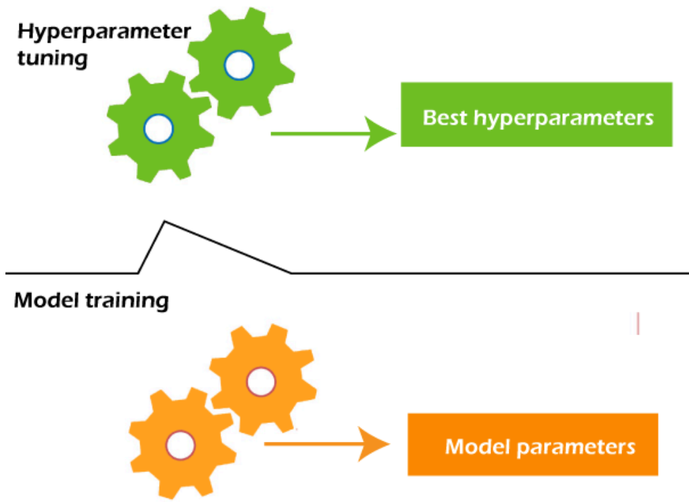
A simple algorithm that classifies data points based on the majority vote of their nearest neighbors.

# Hyperparameter Tuning:-

- Hyperparameter tuning optimizes model performance by selecting the best settings.
- GridSearchCV was used to test multiple hyperparameter combinations via cross-validation.

## Key Hyperparameters Tuned:

✅ SVM: C (regularization), gamma (kernel coefficient).
✅ Gradient Boosting: Learning rate, number of estimators, max depth.
✅ Random Forest: Number of trees, max depth, min samples per split.
✅ KNN: Number of neighbors (K), distance metric.
✅ Logistic Regression: Regularization strength (C), solver.



# Model Evaluation Metrics:-

**To assess model performance, the following metrics were used:**

✅ **Accuracy** – Measures overall correctness of predictions.

- 🔹 **Formula:** TP+TN/TP+TN+FP+FN

✅ **Precision** – Measures how many predicted positive cases were actually positive.

- 🔹 **Formula:** TP/FP+TP

✅ **Recall (Sensitivity)** – Measures how well the model identifies actual positives.

- 🔹 **Formula:** TP/TP+FN

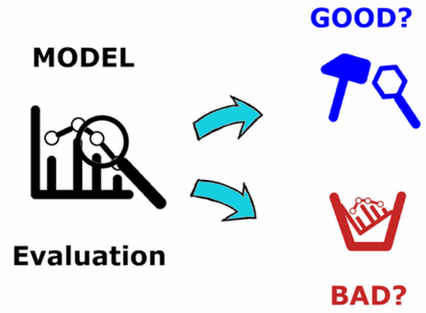
✅ **F1-Score** – Harmonic mean of precision and recall for balanced evaluation.

- 🔹 **Formula:** 2×( Precision*Recall/Precision+Recall )

✅ **AUC-ROC (Area Under Curve - Receiver Operating Characteristic)**

- 🔹 Measures model's ability to distinguish between classes, higher is better.

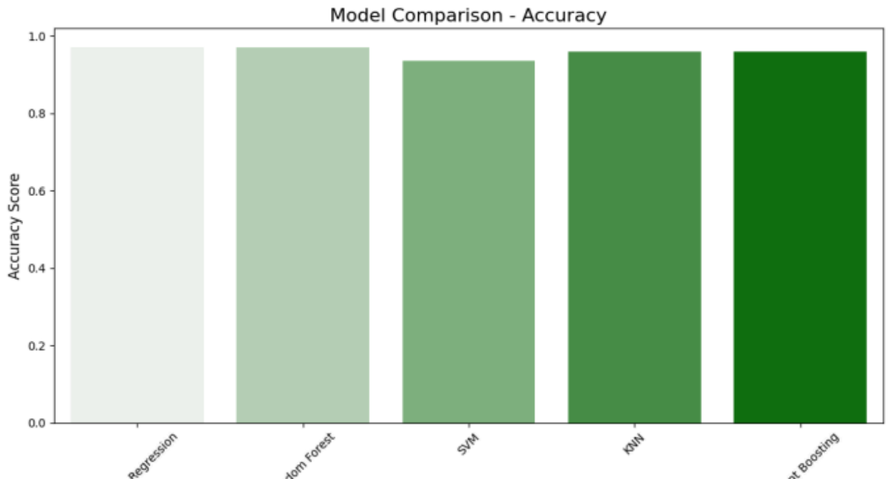
**Key Insight:** Random Forest achieved the highest accuracy and AUC, making it the best model for this dataset.

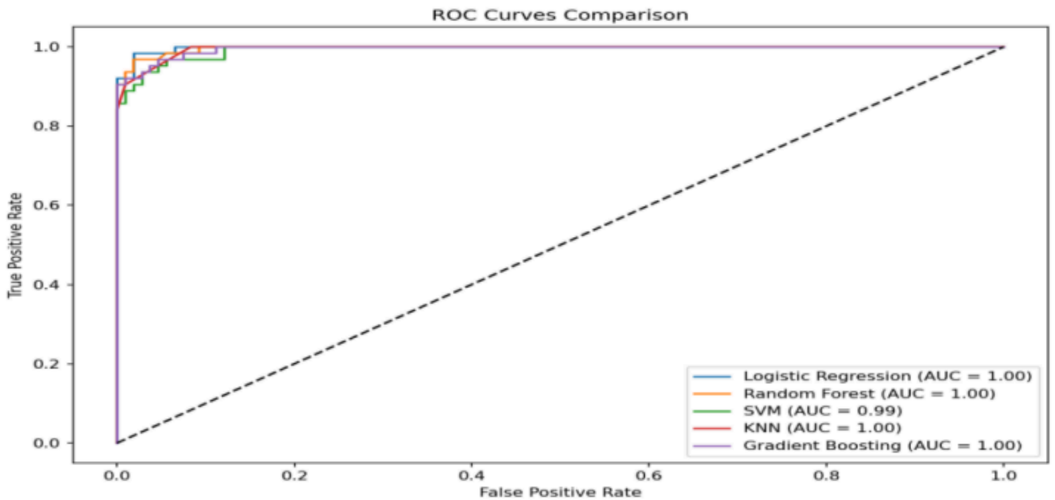


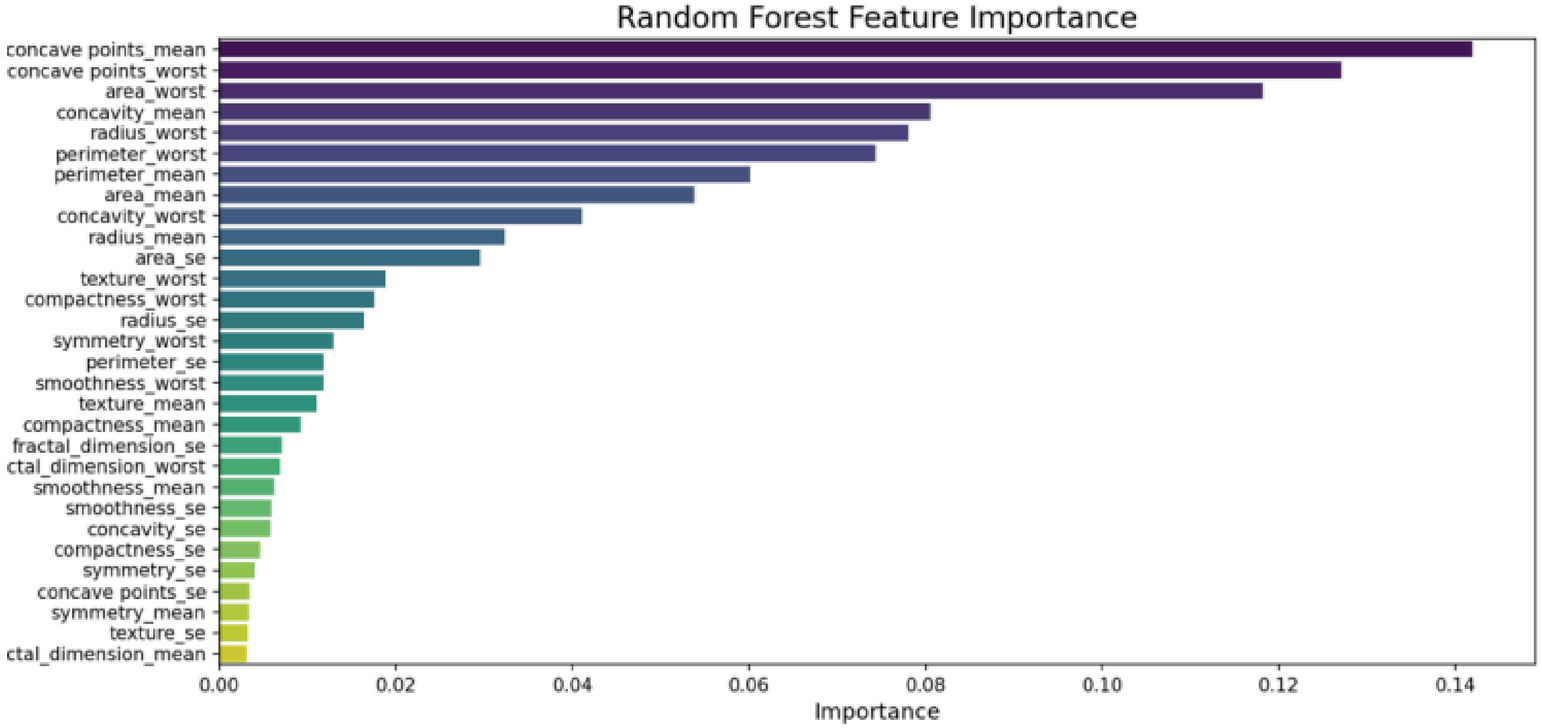
# Outputs:-

Evaluation Metrics for Different Classifiers:

| | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.98125 | 0.948454 | 1.000000 | 0.973545 | 0.992989 |
| Random Forest | 0.98625 | 0.981818 | 0.978261 | 0.980036 | 0.997801 |
| SVM | 0.98125 | 0.948454 | 1.000000 | 0.973545 | 0.990002 |
| Gradient Boosting | 0.98625 | 0.968198 | 0.992754 | 0.980322 | 0.997265 |
| K-Nearest Neighbors | 0.98375 | 0.974729 | 0.978261 | 0.976492 | 0.997438 |
| Naive Bayes | 0.95500 | 0.897351 | 0.981884 | 0.937716 | 0.987274 |






The feature importance chart, shows the ranked significance of various features used by the Random Forest classifier. The top features, such as "concave points_mean" and "concavity_mean," have the highest influence on the model's predictions. This helps identify the key factors contributing to the classification, which can be particularly useful in medical applications, where understanding feature relevance can guide clinical decision-making.

# Results & Discussion:-

- **Model Performance:**
  - Random Forest achieved the highest accuracy (98%) and AUC score (0.997), making it the most effective model.
  - Gradient Boosting performed well, particularly in recall, which is crucial for identifying metastasis risk.
  - SVM had lower accuracy, likely due to the complex, non-linear relationships in the dataset.

- **Key Takeaways:**
  - Ensemble models (Random Forest & Gradient Boosting) outperformed others due to their ability to capture intricate patterns.
  - Hyperparameter tuning played a crucial role in optimizing model performance.
  - The results indicate machine learning's potential in aiding clinicians with breast cancer prognosis.

# Future Work:-

**Enhancing Model Performance:**

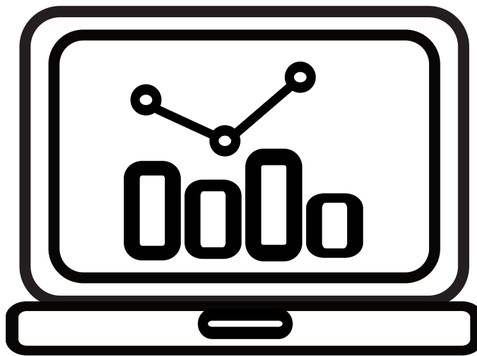
Incorporate additional clinical data, such as genomic markers, to improve prediction accuracy.
Address class imbalance with more diverse and larger datasets for better generalization.

**Exploring Advanced Techniques:**

Implement deep learning models (e.g., CNNs, LSTMs) for more precise predictions.
Investigate feature selection techniques to identify the most critical predictors.

**Clinical Integration:**

Develop a user-friendly decision support system for real-world clinical applications.
Validate models with real-world patient data to ensure robustness and reliability.



# Thank You