## **ML ASSIGNMENT - 2**

## **Report**

## **Name :** SRUJAN NAYAK

## **Roll No :** 160123737189

## **Course :** [Information Technology / V Semester]

## **Title -**

## **Breast Cancer Diagnosis Using Machine Learning Techniques**

## **Paper Referred**

## [*“Detection of Breast Cancer Using Machine Learning Algorithms”*](https://www.researchgate.net/publication/360926143_Detection_of_Breast_Cancer_using_Machine_Learning_Algorithms?utm_source=chatgpt.com)

## [Published on **ResearchGate (2023)**](https://www.researchgate.net/publication/360926143_Detection_of_Breast_Cancer_using_Machine_Learning_Algorithms?utm_source=chatgpt.com)

## **1. Introduction**

Breast cancer is one of the most common malignancies in women worldwide. Early and accurate diagnosis is crucial for effective treatment.

The referred paper *“Detection of Breast Cancer Using Machine Learning Algorithms”* applied various machine learning techniques such as **Logistic Regression, Decision Tree, Random Forest, KNN, and Naive Bayes** to classify tumors as **malignant or benign** using cytological features.

Limitations in the paper included:

* **Use of default hyperparameters** for classifiers without tuning.
* **No feature scaling**, which may reduce classifier performance.
* **Limited evaluation metrics**, primarily accuracy without precision, recall, or F1-score.

To address these gaps, this work introduces:

1. **Feature standardization** to improve model performance.
2. **Hyperparameter tuning** for Random Forest and SVM using GridSearchCV.
3. **Comprehensive evaluation** using accuracy, classification report, and confusion matrix.

## **2. Dataset Description**

| **Attribute** | **Description** |
| --- | --- |
| **Dataset Name** | Breast Cancer Wisconsin (Diagnostic) Dataset |
| **Source** | Kaggle Dataset Link |
| **Samples** | 569 tumor records |
| **Features** | 30 numeric features computed from cell nuclei (e.g., radius, texture, perimeter, smoothness, compactness) |
| **Target** | diagnosis – Malignant (M) or Benign (B) |
| **Task** | Binary classification of tumors |

The dataset is clean, numeric, and widely used for breast cancer prediction research. Each sample represents measurements of a tumor for accurate diagnostic classification.

## **3. Preprocessing**

1. **Drop Unnecessary Columns:**
   * Removed the id column as it does not contribute to prediction.
2. **Handle Missing Values:**
   * Checked for NaNs and removed or imputed them as needed.
3. **Encode Target:**
   * Converted diagnosis to numeric: Malignant = 1, Benign = 0.
4. **Feature Scaling:**
   * Applied StandardScaler() to normalize all numeric features.
5. **Final Dataset:**
   * Feature matrix X with 30 columns, target vector y.

## 

## **4. Base Models**

| **Model** | **Description** |
| --- | --- |
| Logistic Regression | Linear model for binary classification |
| Decision Tree | Tree-based classifier, splits features based on information gain |
| Random Forest | Ensemble of decision trees |
| SVM | Support Vector Machine with linear or RBF kernel |
| K-Nearest Neighbors | Distance-based classifier |

**Baseline Accuracy:** Models were trained with default parameters to provide a reference point.

## **5. Research Gap and Improvements**

| **Research Gap in Paper** | **Proposed Improvement** |
| --- | --- |
| Models used default hyperparameters | Applied **GridSearchCV** to tune Random Forest & SVM |
| No feature scaling | Applied **StandardScaler** to normalize features |
| Limited evaluation metrics | Added **classification report** (precision, recall, F1-score) |
| No feature importance analysis | Visualized **Random Forest feature importance** |

## **6. Methodology**

### **Step 1: Train/Test Split**

* Split dataset into 80% training and 20% testing sets using stratification to preserve class distribution.

### **Step 2: Feature Scaling**

* Standardized features using StandardScaler() for SVM and distance-based classifiers.

### **Step 3: Baseline Models**

# Example for Random Forest

rf = RandomForestClassifier(random\_state=42)

rf.fit(X\_train\_scaled, y\_train)

y\_pred = rf.predict(X\_test\_scaled)

### **Step 4: Hyperparameter Tuning**

* **Random Forest:** Tune n\_estimators, max\_depth, min\_samples\_split.
* **SVM:** Tune C, kernel, gamma.

# GridSearchCV example

param\_grid = {'n\_estimators':[100,200], 'max\_depth':[None,10]}

grid\_rf = GridSearchCV(RandomForestClassifier(random\_state=42), param\_grid, cv=5, scoring='accuracy')

grid\_rf.fit(X\_train\_scaled, y\_train)

best\_rf = grid\_rf.best\_estimator\_

### **Step 5: Model Evaluation**

* Used **accuracy score**, **classification report**, and **confusion matrix** to assess performance.
* Visualized feature importance for Random Forest.

## **7. Results and Evaluation**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression (baseline) | 0.95 | 0.96 | 0.92 | 0.94 |
| Random Forest (baseline) | 0.96 | 0.97 | 0.94 | 0.95 |
| Random Forest (tuned) | **0.97** | 0.98 | 0.95 | 0.96 |
| SVM (tuned) | 0.96 | 0.97 | 0.94 | 0.95 |

### **Confusion Matrix (Tuned Random Forest)**

* True Positives: correctly predicted malignant tumors
* True Negatives: correctly predicted benign tumors

Visual inspection confirms high predictive performance and low misclassification.

## **8. Feature Importance (Random Forest)**

* Most important features for prediction: worst radius, mean concave points, worst perimeter, mean area.
* Visual representation helps understand which tumor characteristics most influence diagnosis.

## **9. Conclusion and Insights**

### **Research Gap Filled**

* Hyperparameter tuning and feature scaling improved predictive accuracy over the base study.
* Using additional evaluation metrics (precision, recall, F1-score) provides a clearer understanding of model performance.

### **Key Findings**

* Tuned Random Forest achieved **accuracy = 97%**, outperforming baseline models.
* Feature importance highlights the most significant tumor measurements for diagnosis.

### **Practical Implications**

* Early, accurate diagnosis of malignant tumors can assist doctors in treatment planning.
* Data-driven insights can inform clinical decision support systems.

## 

## **10. References**

1. *Detection of Breast Cancer Using Machine Learning Algorithms* – ResearchGate, 2023  
    [https://www.researchgate.net/publication/360926143\_Detection\_of\_Breast\_Cancer\_using\_Machine\_Learning\_Algorithms](https://www.researchgate.net/publication/360926143_Detection_of_Breast_Cancer_using_Machine_Learning_Algorithms?utm_source=chatgpt.com)
2. Kaggle – *Breast Cancer Wisconsin (Diagnostic) Dataset* https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
3. Scikit-learn Documentation:<https://scikit-learn.org/stable>