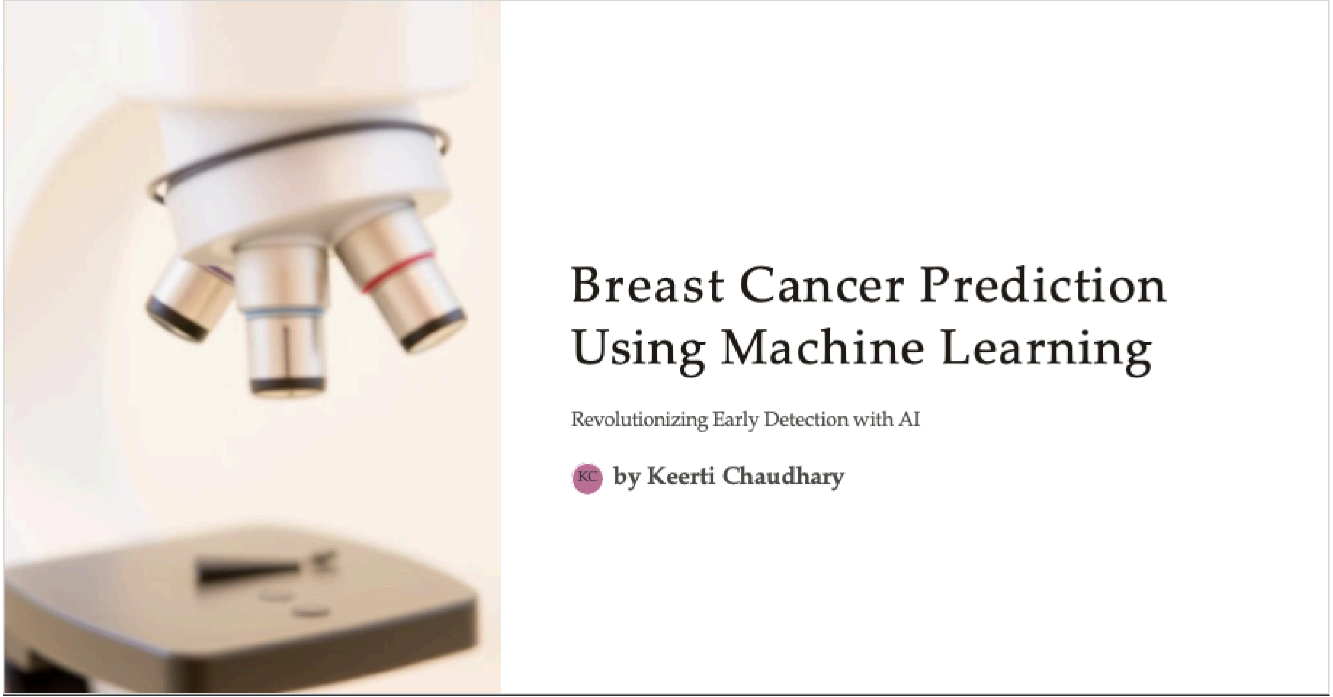
CSEN 240 (Machine Learning)
Winter 2025

Breast Cancer Prediction

Keerti Chaudhary
ID: 07700010984



Breast Cancer Prediction Using Machine Learning

Revolutionizing Early Detection with AI

KC by Keerti Chaudhary

Table of Contents

Abstract

This report explores the application of machine learning techniques for the early prediction of breast cancer. Breast cancer is a common cause of female mortality in developing countries. Early detection and treatment are crucial for successful outcomes. Breast cancer develops from breast cells and is considered a leading cause of death in women. Traditional diagnostic methods like biopsy can be invasive, time-consuming and expensive. The advancements in artificial intelligence (AI) and machine learning (ML) techniques have made it possible to develop more accurate and reliable models for diagnosing and treating this disease This study utilizes data from digitized images of fine needle aspirates (FNA) of breast masses to train and evaluate several machine learning models, including Logistic Regression, Random Forest, Decision Tree, and Support Vector Machines (SVM). The results highlight the potential of machine learning to enhance early breast cancer detection.

Research Content

1. Introduction

Cancer is a worldwide epidemic that affects individuals of all ages and backgrounds. There are many types of cancer, however, breast cancer is one of the most common cancers in women. Due to this challenge, researchers should pay special attention to cancer detection and prognosis. Predicting and diagnosing cancer at an early stage is an area where machine-learning approaches may have a significant impact. The effectiveness of breast cancer treatment is strongly correlated with the stage at which the cancer is diagnosed. Early detection significantly increases the chances of successful treatment and improves patient survival rates. Traditional diagnostic methods, such as mammography, clinical breast exams, and biopsies, play a crucial role in breast cancer detection. However, these methods can be time-consuming, expensive, and sometimes invasive.

Machine learning (ML) offers a complementary and potentially more efficient approach to early breast cancer detection. By analyzing large datasets of medical information, ML algorithms can identify patterns and predict the likelihood of malignancy. This approach can aid in the early identification of individuals at high risk, enabling timely intervention and treatment.

This research investigates the application of various machine learning algorithms to predict breast cancer using data obtained from fine needle aspirates (FNA). The goal is to evaluate the performance of different models and identify key factors that contribute to accurate prediction.

2. Theory

This research leverages several key machine learning algorithms for classification:

- **Logistic Regression:** A linear model used for binary classification problems. It predicts the probability of a data point belonging to a particular class. Despite its simplicity, it can be effective for certain classification tasks and provides interpretable results.

- **Decision Tree:** A tree-like structure where each internal node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome. Decision trees are versatile and can handle both categorical and numerical data.

- **Random Forest:** An ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random forests are known for their high accuracy and robustness to overfitting.

- **Support Vector Machine (SVM):** A powerful algorithm that finds the optimal hyperplane that best separates data points of different classes. SVMs are effective in high-dimensional spaces and can handle non-linear classification through the use of kernel functions.

3. Motivation and Benefits

Breast cancer diagnosis by machine learning has been motivated by the hope that it will lead to better patient outcomes, lessen the disease's worldwide effect, and aid in the development of cutting-edge healthcare technology and research.

The benefits of this research are as follows; it helps in improvement in early diagnosis and individualized therapy. In addition, it has the potential to revolutionize breast cancer management and save lives by influencing areas such as research, cost-effectiveness, and worldwide accessibility to healthcare services. It helps in reducing healthcare costs, and a more beneficial influence on worldwide breast cancer management is all possible because of the abilities of machine learning in breast cancer diagnosis.

4. Experiments

4.1 Dataset

The dataset used in this research is derived from digitized images of a fine needle aspirate (FNA) of a breast mass, obtained from the University of Wisconsin. Fine needle aspiration is a minimally invasive procedure used to sample cells from a suspicious lump for examination under a microscope. The dataset contains features extracted from these FNA images, which characterize the cell nuclei present in the sample. These features serve as input variables for the machine learning models.

4.2 Preprocessing

Before training the machine learning models, the dataset undergoes several preprocessing steps to ensure data quality and optimize model performance:

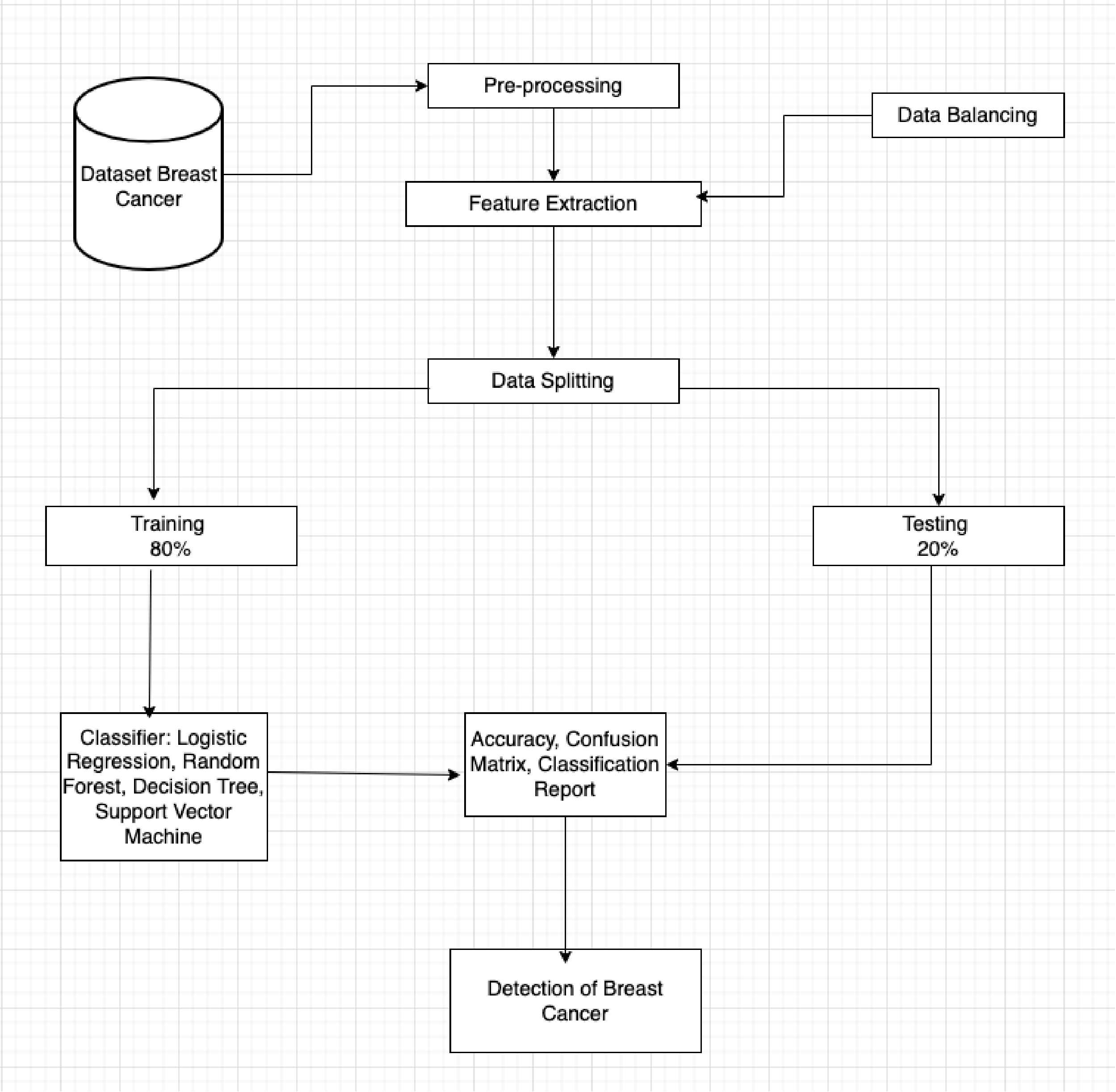
- **Feature Scaling:** Scaling techniques are applied to normalize the range of independent variables. This is important because machine learning algorithms, particularly those based on distance calculations (like SVM), are sensitive to the scale of the input features.
- **Handling Missing Values:** Strategies are implemented to address missing data points, which may involve imputation (replacing missing values with estimated values) or removal of incomplete records, depending on the extent and nature of the missing data.
- **Addressing Class Imbalance:** Techniques are employed to mitigate the impact of class imbalance, where one class (e.g., malignant or benign) has significantly more samples than the other. Class imbalance can bias the performance of machine learning models, leading to poor generalization.

4.3 Model Selection and Training

The following machine learning models are selected for this study:

- Logistic Regression
- Random Forest
- Decision Tree
- Support Vector Machine (SVM)

Each model is trained on the preprocessed dataset. The training process involves adjusting the model parameters to minimize the prediction error on the training data.


Dataset Breast Cancer
Pre-processing
Data Balancing
Feature Extraction
Data Splitting
Training 80%
Testing 20%
Classifier: Logistic Regression, Random Forest, Decision Tree, Support Vector Machine
Accuracy, Confusion Matrix, Classification Report
Detection of Breast Cancer

4.4 Evaluation

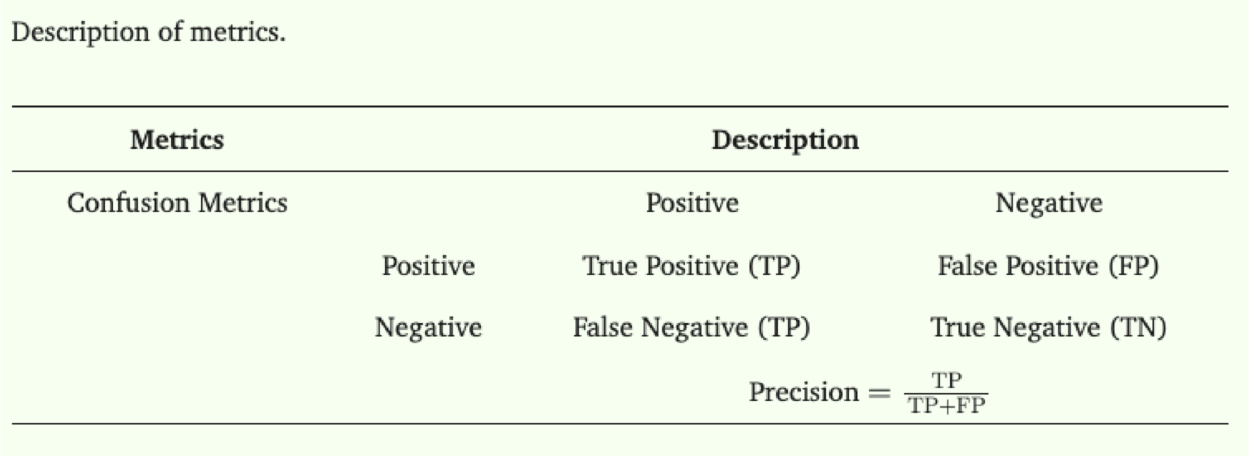
The performance of the trained models is evaluated using several metrics to provide a comprehensive assessment of their predictive capabilities:

- **Accuracy:** The proportion of correctly classified instances out of the total number of instances.
- **Confusion Matrix:** A table that summarizes the performance of a classification model by showing the counts of true positive, true negative, false positive, and false negative predictions.
- **Classification Report:** A report that provides various classification metrics, including precision, recall, and F1-score, for each class.

5. Experimental Results

The experimental results are analyzed from different perspectives to gain a thorough understanding of the models' performance:

- **Accuracy Comparison:** The accuracy of each model is compared to assess its overall predictive power.
- **Confusion Matrix Analysis:** The confusion matrix is examined to analyze the types of errors made by each model, specifically focusing on false positives (predicting malignancy when it is benign) and false negatives (predicting benign when it is malignant). Minimizing false negatives is particularly critical in medical diagnosis.
- **Feature Importance Analysis:** For the Random Forest model, feature importance analysis is conducted to identify the most significant medical indicators contributing to the prediction of breast cancer. This analysis can provide valuable insights into the underlying factors associated with malignancy.
- **Visualization:** Visual aids, such as graphs and charts, are used to present the class distribution in the dataset and to visualize the performance metrics of the different models. A sample confusion matrix visualization is shown in.



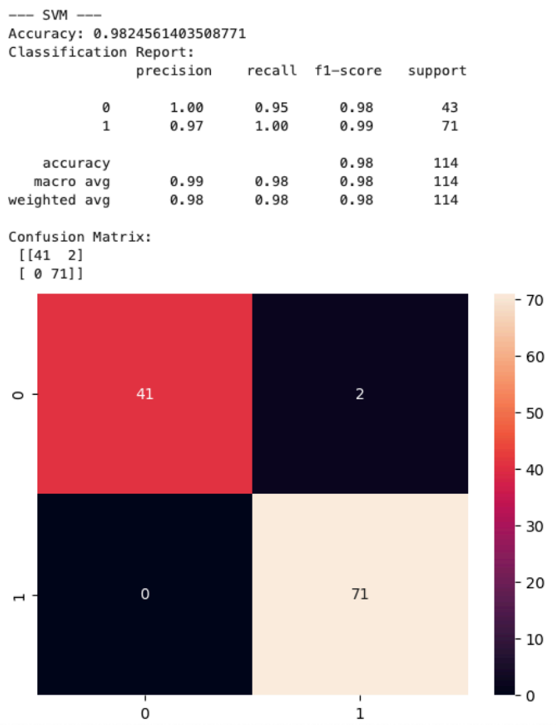
Description of metrics.

Metrics	Description		
Confusion Metrics		Positive	Negative
	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (TP)	True Negative (TN)
		Precision $= \frac{\text{TP}}{\text{TP+FP}}$	

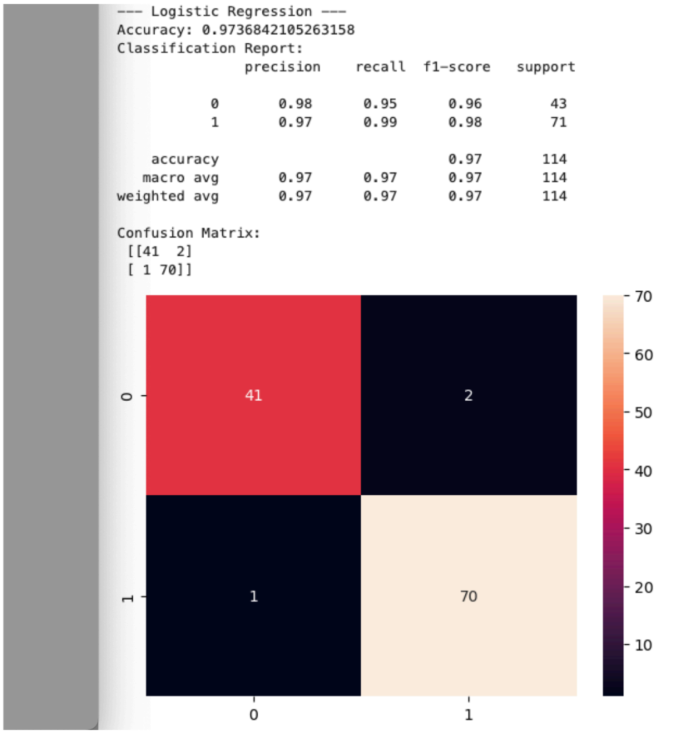
6. Analysis and Insights

The analysis of the experimental results provides several key insights:

- **SVM and Logistic Regression Performance:** The Support Vector Machine (SVM) and Logistic Regression models demonstrated strong performance in predicting breast cancer risk. These models achieved high accuracy and showed effectiveness in distinguishing between benign and malignant cases.

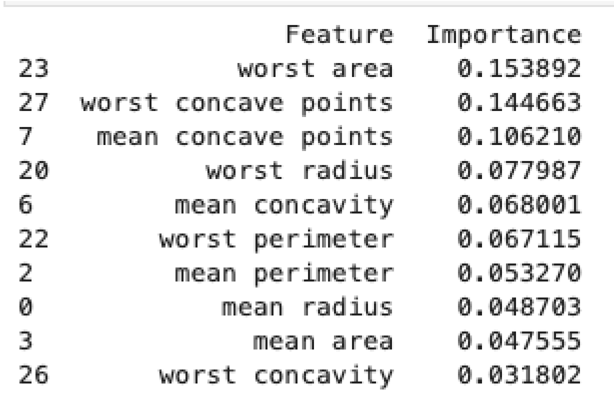


```
--- SVM ---
Accuracy: 0.9824561403508771
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.95      0.98        43
           1       0.97      1.00      0.99        71

    accuracy                           0.98       114
   macro avg       0.99      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114

Confusion Matrix:
 [[41  2]
 [ 0 71]]
```



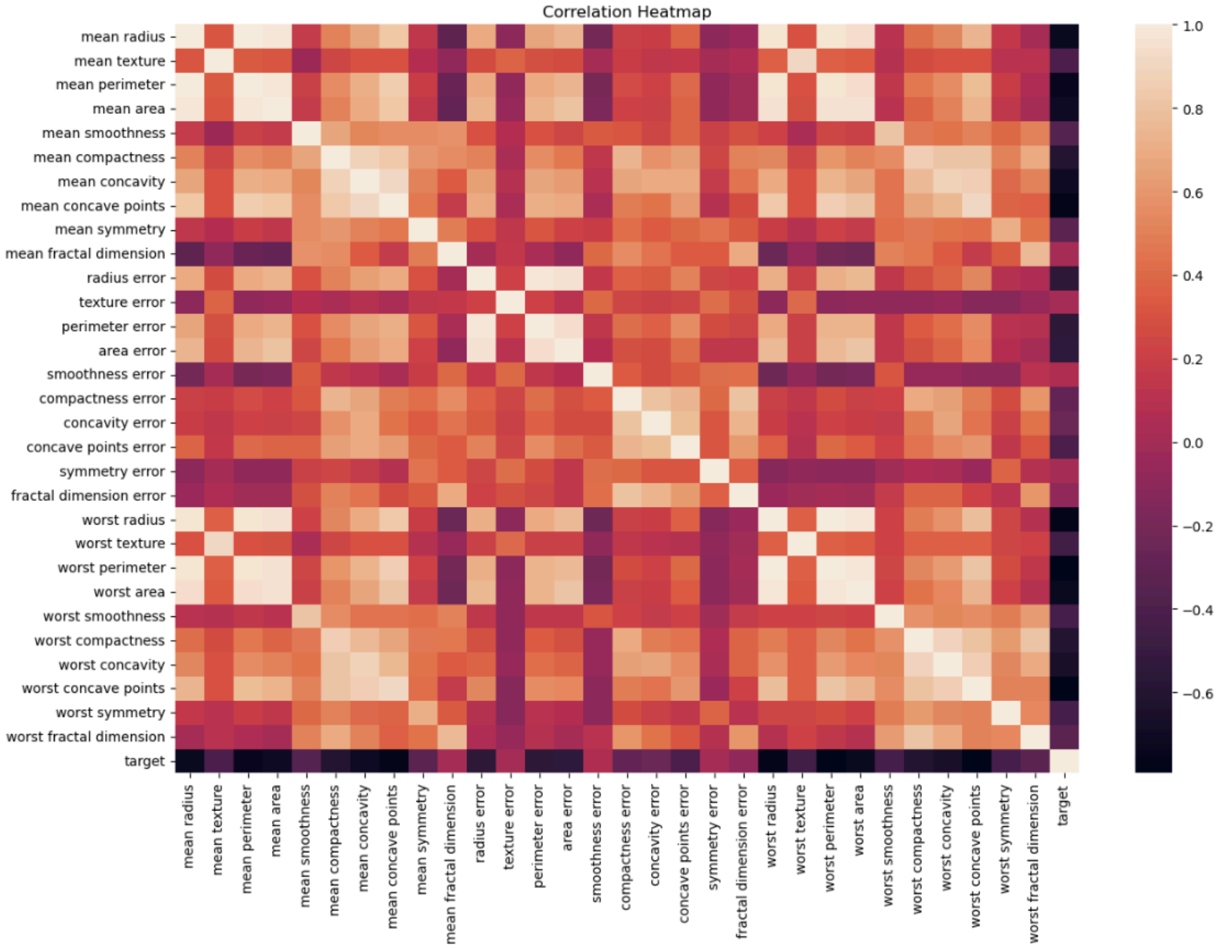
```
--- Logistic Regression ---
Accuracy: 0.9736842105263158
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.95      0.96        43
           1       0.97      0.99      0.98        71

    accuracy                           0.97       114
   macro avg       0.97      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114

Confusion Matrix:
 [[41  2]
 [ 1 70]]
```

- **Decision Tree Performance:** The Decision Tree model exhibited a tendency to overfit the training data. Overfitting occurs when a model learns the training data too well, including the noise and random fluctuations, leading to poor generalization performance on unseen data.
- **Feature Importance:** The feature importance analysis conducted on the Random Forest model identified key medical indicators that are crucial for early breast cancer detection. These indicators can provide valuable information for clinicians in assessing patient risk and making informed decisions about further diagnostic procedures.



```
                 Feature  Importance
23            worst area    0.153892
27  worst concave points    0.144663
7    mean concave points    0.106210
20          worst radius    0.077987
6         mean concavity    0.068001
22       worst perimeter    0.067115
2         mean perimeter    0.053270
0            mean radius    0.048703
3              mean area    0.047555
26       worst concavity    0.031802
```

- **Importance of Early Detection:** The overall findings underscore the importance of early detection in breast cancer. Machine learning techniques have the potential to significantly improve the accuracy and efficiency of early detection efforts, ultimately leading to better patient outcomes.



Conclusion

This research demonstrates the potential of machine learning to revolutionize early breast cancer detection. By leveraging data from fine needle aspirates and employing various machine learning algorithms, accurate prediction of breast cancer risk can be achieved. The Support Vector Machine and Logistic Regression models exhibited particularly strong performance. Feature importance analysis identified key medical indicators that can aid in clinical decision-making. While the Decision Tree model showed signs of overfitting, further optimization and hyperparameter tuning could improve its performance. Future work should focus on enhancing model performance through techniques such as hyperparameter tuning, exploring deep learning approaches, and validating the findings on larger and more diverse datasets. The application of machine learning in breast cancer detection holds great promise for improving early diagnosis, facilitating timely intervention, and ultimately reducing the burden of this disease.

References

1. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
2. https://pubmed.ncbi.nlm.nih.gov/?term=machine+learning+prediction
3. https://pmc.ncbi.nlm.nih.gov/articles/PMC10572157/