Title: Cancer Cell Classification

By: Ayush Nautiyal

Table of Contents

1. Abstract

Breast cancer is one of the most common and life-threatening diseases, making early detection crucial for effective treatment. This project focuses on the classification of breast cancer cells as either malignant or benign using machine learning techniques. Various machine learning models were trained and evaluated, with Support Vector Classifier (SVC) using a linear kernel achieving the best performance. The SVC model attained an accuracy of **98.61%** and a recall score of **97.87%**, indicating its high reliability in identifying cancerous cells. These results demonstrate the potential of machine learning in assisting medical professionals with accurate and efficient cancer diagnosis.

2. Introduction

Breast cancer is one of the most prevalent cancers worldwide and a leading cause of mortality among women. Early detection plays a vital role in improving survival rates, as timely diagnosis allows for effective treatment and management. Traditional diagnostic methods, such as biopsy and histopathological analysis, are time-consuming and subject to human interpretation, which can lead to inconsistencies.

Machine learning (ML) offers a promising approach to automating and enhancing the accuracy of breast cancer detection. By analyzing cellular features, ML models can

efficiently classify breast cancer cells as malignant or benign. This project explores various ML algorithms to determine the most effective model for breast cancer cell classification. After rigorous evaluation, the **Support Vector Classifier (SVC) with a linear kernel** was found to be the most accurate, achieving an **accuracy score of 98.61%** and a **recall score of 97.87%**. These results highlight the potential of ML-based methods in supporting medical professionals with reliable and efficient diagnostic tools.

---

3. Literature Review

Breast cancer classification has been an extensively researched topic in medical and computational fields, with machine learning (ML) playing a significant role in improving diagnostic accuracy. Traditional diagnostic methods, such as fine needle aspiration cytology (FNAC) and histopathological examination, require manual analysis by pathologists, which can be subjective and time-consuming. To address these limitations, researchers have explored various ML techniques to automate the classification process and enhance detection accuracy.

Several studies have demonstrated the effectiveness of ML algorithms in breast cancer detection. **Wolberg et al. (1994)** introduced the **Wisconsin Breast Cancer Dataset (WBCD)**, which has since been widely used in breast cancer classification research. Studies utilizing this dataset have shown that ML models can achieve high accuracy in distinguishing malignant from benign cells.

Support Vector Machines (SVM) have been widely applied in cancer classification due to their ability to handle high-dimensional data efficiently. **Cortes and Vapnik (1995)** introduced SVM, which has since been used extensively in medical diagnosis applications. Research has demonstrated that SVM, particularly with a **linear kernel**, performs well in classifying breast cancer cells due to its ability to find an optimal decision boundary.

Other ML models, including **Random Forest, Decision Trees, k-Nearest Neighbors (k-NN), and Neural Networks**, have also been studied for breast cancer classification. **Patel and Goyal (2019)** compared multiple ML models and found that **SVM outperformed other classifiers in terms of accuracy and generalization ability**. Deep learning approaches, such as Convolutional Neural Networks (CNNs), have also been explored, particularly for histopathological image classification. However, these models often require large datasets and significant computational resources.

Recent advancements in ML have also emphasized **feature selection and optimization techniques** to improve model performance. Studies have shown that **principal component analysis (PCA), recursive feature elimination (RFE), and genetic algorithms** can enhance classification accuracy by selecting the most relevant features.

Based on previous research, this project evaluates multiple ML models and identifies **Support Vector Classifier (SVC) with a linear kernel** as the most effective method for breast cancer classification, achieving **98.61% accuracy** and a **recall score of 97.87%**. These findings align with existing literature and demonstrate the continued effectiveness of SVM in medical diagnosis applications.

---

4. Methodology

4.1 Data Collection

- Dataset is imported by scikit-learn datasets library.
- Contains independent features: mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, radius error, texture error, perimeter error, area error, smoothness error, compactness error, concavity error, concave points error, symmetry error, fractal dimension error, worst fractal dimension.
- Target feature: Target.

4.2 Data Preprocessing

- No duplicates are found.
- Check for missing values and none are found.
- Plot the distribution plot and found that the data is not distributed evenly so used log_transform to give them a normal distribution shape.
- Outliers are detected by barplot so IQR(inter-quantile range) method was used.
- The data is highly multicollinear detected by high correlations between the data features.
- Split in train, valid, and test data.
- While using linear models use l1 regularization for multicollinearity.

4.3 Machine Learning Models Used

- Logistic Regression with l1 regularization
- SGD Classifier
- Linear SVC
- SVC (kernel= Linear)
- Bernoulli Naïve Bayes
- Gaussian Naïve Bayes
- K Neighbors Classifer

- Decision Tree Classifier
- Bagging Classifier (Estimator= SVC)
- Extra Tree Classifier
- Gradient Boosting
- Hist Gradient Boosting
- Random Forest Regressor

4.4 Model Evaluation Metrics

- Accuracy Score
- Confusion Matrix
- Classification Report
- Learning Curve
- Recall Score

---

5. Results and Analysis

Various models were tried for the analysis and the best result was the SVC model. In the analysis it was found that the linear model like logistic regression is overfitting the data and the high performance/ strong models like ensemble techniques are underfitting the data and were unable to catch the pattern due to the small dataset.

98.61% accuracy and a **recall score of 97.87%** was achieved by SVC.

Discussion

These high scores indicate that the model is highly reliable in distinguishing malignant cells from benign ones, making it a strong candidate for aiding in medical diagnosis.

One of the key advantages of the **SVC (linear kernel)** is its ability to find an optimal decision boundary in high-dimensional space. Given that breast cancer classification involves multiple cellular features, this property of SVC contributes to its superior performance. Additionally, the high recall score of **97.87%** suggests that the model effectively minimizes false negatives, which is critical in medical applications, where missing a cancer diagnosis can have severe consequences.

---

6. Conclusion and Future Work

In conclusion, this study reaffirms that **SVM with a linear kernel is a powerful and reliable tool for breast cancer classification**. By improving data quality,

incorporating advanced feature extraction methods, and exploring deep learning techniques, machine learning can significantly contribute to early detection and better patient outcomes in breast cancer diagnosis.
