DEBRE BERHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPATRMENT OF SOFTWARE ENGINEERING

Name: Eyerusalem Chernet

ID: 1401229

Course Code: SEng4091

Submitted to: Derbew F.

Submission Date: Feb 2, 2017

Breast Cancer Prediction and Analysis.

Table of Contents

Introduction

Breast cancer remains one of the most widespread and serious health challenges faced by women across the globe. It's a disease that touches countless lives, not only those diagnosed but also their families and communities. The good news is that advancements in medical science and technology have shown that early detection can significantly improve outcomes. When breast cancer is caught in its initial stages, treatment options are more effective, less invasive, and survival rates increase dramatically. This underscores the importance of developing tools that can help identify potential cases as early as possible.

In this project, we explore how modern technology, specifically machine learning, can play a pivotal role in the fight against breast cancer. By analyzing medical imaging data, we aim to create a predictive model capable of distinguishing between benign (non-cancerous) and malignant (cancerous) tumors. This isn't just about building a sophisticated algorithm—it's about creating a tool that can support healthcare professionals in making faster, more accurate diagnoses.

This approach involves using real-world medical data to train and refine a classification model. We'll experiment with various machine learning techniques to ensure the model is not only accurate but also interpretable, so doctors and clinicians can understand and trust its predictions. Ultimately, our goal is to contribute to the ongoing efforts to improve healthcare technology, making a meaningful difference in the lives of those affected by breast cancer. Through this work, we hope to take one step closer to a future where early detection and effective treatment are accessible to all.

Problem Definition

Breast cancer diagnosis is a critical process that currently depends on a combination of medical examinations, including biopsies and imaging techniques like mammograms. While these methods are widely used and have saved countless lives, they are not without their challenges. Manual diagnosis, for instance, can be influenced by subjectivity—different practitioners may interpret the same data in varying ways. Human error, whether due to fatigue, oversight, or other factors, can also creep into the diagnostic process.

Additionally, the variability in experience and expertise among healthcare professionals can lead to inconsistencies in diagnosis, potentially delaying treatment or causing unnecessary anxiety for patients.

This is where the power of machine learning comes into play. By developing a robust and reliable machine learning model, we can provide medical professionals with a valuable tool to complement their expertise. Think of it as a second opinion—one that is data-driven, consistent, and free from human bias. Such a model has the potential to reduce diagnostic errors, improve accuracy, and ultimately lead to better patient outcomes.

The goal is to create a system that doesn't replace doctors but empowers them. By leveraging the strengths of machine learning, we can help streamline the diagnostic process, giving healthcare providers more confidence in their decisions and ensuring that patients receive the timely and accurate care they deserve. This is the problem we're tackling: bridging the gap between human expertise and technological innovation to make breast cancer diagnosis more reliable and accessible for everyone.

Why Is This Important?

- **Early Detection Saves Lives**: Early identification of malignant tumors significantly improves patient outcomes.
- **Reducing Misdiagnosis**: Machine learning models can provide a consistent and objective analysis of medical data.
- **Scalability**: Automating tumor classification can help scale diagnostic services, especially in areas with a shortage of specialists.

Approach

The aim is to develop a supervised classification model capable of predicting whether a given tumor is benign (harmless) or malignant (cancerous) based on extracted medical imaging features. We will:

- Collect and preprocess the dataset
- Explore and visualize the data to understand key patterns
- Train a **Logistic Regression** classification model
- Evaluate the model's effectiveness using relevant performance metrics
- Deploy the model as an API for real-world usage

Data Acquisition

Dataset Source

The dataset used for this project was obtained from **Kaggle**, a popular platform for data science competitions and datasets.

License and Usage Terms

The dataset is publicly accessible and intended for educational and research purposes. No commercial usage restrictions apply.

Dataset Overview

- **Number of Instances**: 569 patients
- **Number of Features**: 32 (including ID and diagnosis label)
- **Target Variable**: Diagnosis (M = Malignant, B = Benign)
- **Feature Breakdown**:
    - 30 numerical features extracted from digitized medical images.
    - Features include radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension.

Data Understanding and Exploration

We start by loading the dataset and performing exploratory data analysis (EDA) to identify key patterns and relationships.

Data Loading

```
import pandas as pd

# Load the dataset

df = pd.read_csv('breast-cancer.csv')
```

Data Cleaning

```
#Drop the 'id' column as it is not useful for prediction

df.drop(columns=['id'], inplace=True)

#Convert diagnosis to binary values: M = 1, B = 0

df['diagnosis']= df['diagnosis'].map({'M': 1, 'B': 0})
```

Data Visualization

```
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

# Correlation heatmap
plt.figure(figsize=(12, 10))
sns.heatmap(df.corr(), cmap='coolwarm', linewidths=0.5)
plt.title('Feature Correlation Heatmap')
plt.show()
```

Data Preprocessing

Before training the model, we preprocess the dataset by scaling numerical features

```
from sklearn.preprocessing import StandardScaler

Separate features and target

X = df.drop(columns=['diagnosis']) y = df['diagnosis']

Standardize the features

scaler = StandardScaler()

 X_scaled = scaler.fit_transform(X)
```

Model Implementation and Training

We train a **Logistic Regression** model for binary classification.

from sklearn.model_selection import train_test_split from sklearn.linear_model import LogisticRegression

Split the data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

Train the Logistic Regression model

model = LogisticRegression() model.fit(X_train, y_train)

Model Evaluation and Analysis

We evaluate model performance using accuracy, confusion matrix, and ROC curve.

```
from sklearn.metrics import accuracy_score, confusion_matrix, roc_curve, auc

Predict on the test set

y_pred = model.predict(X_test)

Accuracy

accuracy = accuracy_score(y_test, y_pred) print(f'Accuracy: {accuracy:.2f}')

Confusion Matrix

conf_matrix = confusion_matrix(y_test, y_pred) sns.heatmap(conf_matrix, annot=True,
cmap='Blues') plt.title('Confusion Matrix') plt.xlabel('Predicted') plt.ylabel('Actual') plt.show()

ROC Curve

y_pred_prob = model.predict_proba(X_test)[:, 1] fpr, tpr, _ = roc_curve(y_test, y_pred_prob)
roc_auc = auc(fpr, tpr)

plt.plot(fpr, tpr, label=f'ROC curve (area = {roc_auc:.2f})') plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('False Positive Rate') plt.ylabel('True Positive Rate') plt.title('ROC Curve')
plt.legend(loc='lower right') plt.show()
```

Model Deployment

The trained model is deployed using **FastAPI** on **Render**.

Conclusion

This project demonstrates the effectiveness of machine learning in medical diagnosis. The Logistic Regression model achieves high accuracy and provides valuable insights into breast cancer prediction. Future improvements could include:

- Exploring deep learning approaches for enhanced accuracy.
- Integrating additional medical imaging data.
- Expanding the API for real-world integration.

References

- Kaggle: Breast Cancer Dataset
- scikit-learn documentation
- FastAPI documentation
- Render cloud deployment platform