**Decoding: Breast Cancer**

**Submitted for**

**Statistical Machine Learning CSET211**

Submitted by:

**(E23CSEU0383) Lakshay Gupta**

**(E23CSEU0372) Yash Verma**

**(E23CSEU0373) Anshuman Mathur**

Submitted to

**DR. NITIN ARVIND SHELKE**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

**INDEX**

|  |  |  |
| --- | --- | --- |
| Sr.No | Content | Page No |
| 1 | Abstract |  |
| 2 | Introduction |  |
| 3 | Methadology |  |
| 4 | Hardware/Software Required |  |
| 5 | Experimental Results |  |
| 6 | Conclusions |  |
| 7 | Future Scope |  |
| 8 | Github Link |  |

**Abstract:**

Breast cancer remains one of the leading causes of cancer-related deaths worldwide, making early detection crucial for improving survival rates. Traditional diagnostic methods, such as biopsies and mammograms, are time-consuming and often require expert interpretation, which can delay diagnosis. This project aims to develop a machine learning (ML)-based model to predict the likelihood of breast cancer in patients using medical data. The model utilizes a dataset containing key features such as tumor size, shape, texture, and other relevant factors to classify tumors as either benign or malignant. Various machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), and Random Forest, are employed to train and test the model, with performance evaluated using metrics such as accuracy, precision, recall, and F1-score. The goal is to achieve a high level of prediction accuracy to assist healthcare professionals in early and accurate breast cancer detection. By leveraging the power of machine learning, this project aims to enhance diagnostic efficiency, reduce the burden on medical practitioners, and contribute to improved patient outcomes through timely treatment.

**Introduction:**

Breast cancer is one of the most prevalent and deadly forms of cancer worldwide, affecting millions of individuals each year. Early detection plays a crucial role in improving survival rates, as it allows for timely intervention and more effective treatment options. However, traditional diagnostic methods, such as mammography and biopsies, can be invasive, time-consuming, and dependent on the skill and experience of healthcare professionals. As a result, there is a growing need for more efficient, automated, and accurate methods of detecting breast cancer.

Machine learning (ML) has emerged as a powerful tool in healthcare, offering the potential to revolutionize diagnostic procedures. By leveraging large datasets and advanced algorithms, ML models can recognize patterns and make predictions that may not be immediately obvious to human experts. This project aims to develop a machine learning-based model to predict the likelihood of breast cancer based on a variety of tumor characteristics and clinical factors. Using a well-established dataset, the model will be trained to classify tumors as either benign or malignant, providing a faster and potentially more accurate alternative to traditional diagnostic methods.

The primary objective of this study is to build a robust machine learning model capable of assisting healthcare professionals in the early detection of breast cancer. By reducing diagnostic time and providing more accurate predictions, the model could significantly improve patient outcomes and contribute to the ongoing effort to combat breast cancer. This research explores the use of different ML algorithms, evaluates their performance, and aims to identify the most effective approach for breast cancer prediction.

**Methodology:**

1. **Data Collection**
   1. **Dataset: This dataset consists of features derived from Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses.**
   2. **Features: The dataset includes 9 numerical features that describe tumor characteristics**
2. **Data Preprocessing**
   1. **Data Cleaning: Inspect the dataset for missing, inconsistent, or erroneous data. Any missing values will be imputed using appropriate techniques, such as mean or median imputation, or rows with missing data may be removed.**
   2. **Feature Selection: Select the most relevant features based on domain knowledge, correlation analysis, or feature importance techniques.**
3. **Model Selection**
   1. **Algorithm Selection: Several machine learning algorithms will be considered for the classification task:**
      1. **Logistic Regression: A simple but effective linear classifier for binary classification tasks**
      2. **K-Nearest Neighbors (KNN): A distance-based classifier that assigns a class based on the majority class among the nearest neighbors.**
4. **Model Training and Evaluation**
   1. **Training: Split the data into a training set (e.g., 80% of the data) and a testing set (e.g., 20% of the data). T**
5. **Evaluation Metrics**
   1. **Accuracy: The proportion of correctly classified instances out of the total instances**
   2. **F1-Score: The harmonic mean of precision and recall, providing a balance between the two metrics.**
   3. **Confusion Matrix: A matrix showing true positives, true negatives, false positives, and false negatives, which helps to assess the model’s performance in detail.**
6. **Model Deployment**

**Software Requirement:**

**Python 3.x, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn**

**Experimental Results:**

1. **Data Split:**
   1. **Training Split: 80 percent of the dataset**
   2. **Test Set: 20 percent of the dataset**
2. **Model Evaluation Metrics:**
   1. **Accuracy (KNN) = 97.81**
   2. **F1 Score (KNN) = 0.98**
   3. **Accuracy (LR) = 98.54**
   4. **F1 Score (LR) = 0.99**

**Summary:**

**This project developed a machine learning model for breast cancer detection, utilizing various algorithms such as Logistic Regression and KNN. The Logistic Regression model achieved 98.54% accuracy, and 99% F1-score, demonstrating excellent performance in distinguishing between breast cancer and not. While the model's performance is promising, it should complement, not replace, clinical expertise. Overall, the project highlights the potential of machine learning in improving early breast cancer detection and patient outcomes.**

**Future Scope:**

**The project can be scaled on to detecting the other various types of cancers, rather than just one.**

**It can also be scaled to be able to detect cancer through the x-ray images shown to it.**

**GitHub:**

**https://github.com/Dabby7/Decoding-Breast-Cancer**