A Project Report

on

BREAST CANCER DETECTION

by

<TVISHA SHETTY>

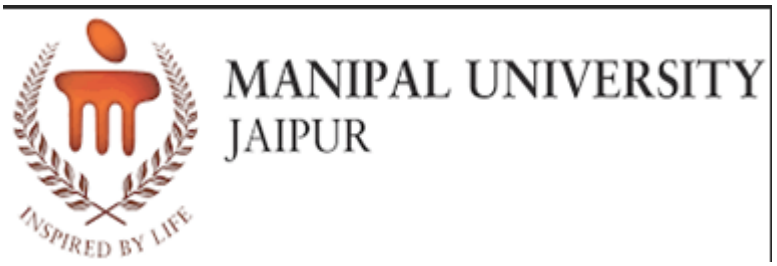
<23FE10CDS00151>

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering (Data Science)



School of Information, Security and Data Science

Department of Data Science and Engineering

MANIPAL UNIVERSITY JAIPUR, JAIPUR

RAJASTHAN, INDIA

Nov 2024

CERTIFICATE

Date:20 NOVEMBER 2024

This is to certify that the project based learning, project titled BREAST CANCER DETECTION is a record of the bonafide work done by TVISHA SHETTY (23FE10CDS00151) submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering (Data Science) of Manipal University Jaipur, Jaipur during the academic year 2024-25.

<GAURAV KUMAWAT >

<MENTOR >

Project Guide, Department of Data Science and Engineering

Manipal University Jaipur, Jaipur

Abstract

Breast cancer continues to be a major health concern worldwide, making early detection crucial for improving survival rates. Current diagnostic tools, while effective, can lead to misdiagnoses due to limitations such as false positives and false negatives. Machine learning (ML) has shown great potential in assisting breast cancer detection by analyzing complex patterns in medical data. However, existing ML approaches often struggle with challenges like selecting the most relevant features and preventing overfitting, which can limit their accuracy and reliability.

This project aims to develop an efficient breast cancer detection system using machine learning models such as Logistic Regression, Random Forest, and XGBoost. The system leverages publicly available datasets to classify malignant and benign tumors based on various diagnostic features. By comparing the performance of these models in terms of accuracy, precision, recall, and F1 score, we identify the most suitable model for reliable breast cancer prediction. This study provides a practical solution for aiding early diagnosis, contributing to better treatment outcome

LIST OF TABLES

LIST OF FIGURES

Table of Contents

Introduction

- Breast cancer is one of the most prevalent and life-threatening diseases among women worldwide, accounting for a significant number of cancer-related deaths annually. Early detection plays a crucial role in improving patient outcomes, as it enables timely intervention and treatment. While traditional diagnostic methods such as mammography and biopsy are widely used, they can sometimes produce inaccurate results, leading to false positives or false negatives. This not only impacts the patient's treatment plan but also adds to the emotional and financial burden.

- Machine learning (ML) has emerged as a powerful tool in healthcare, offering the ability to analyse vast amounts of data and identify patterns that may be missed by conventional diagnostic techniques.

- Recent advancements help develop an efficient breast cancer detection system using machine learning models such as Logistic Regression, Random Forest, and XGBoost.

Problem Statement

- Despite significant advancements in breast cancer detection, existing diagnostic techniques face ongoing challenges in delivering consistently accurate results.

- Addressing this gap, the project aims to create a breast cancer detection system that uses machine learning to provide a more accurate, reliable, and practical solution for early detection. This project will focus on making these improvements feasible within the scope of student implementation while offering better diagnostic outcomes.

Figure 1. The growing trend for the applications of ML in the medical field.

Objective

- The objective of this project is to develop an advanced breast cancer detection system that significantly improves diagnostic accuracy by integrating machine learning (ML) .

- The system will focus on optimizing feature selection, which is crucial for improving the accuracy of the diagnostic model. By selecting the most relevant features from medical datasets, the system will be better equipped to differentiate between cancerous and non-cancerous cases, reducing the chances of false positives and false negatives.

- Ultimately, the goal of this project is to contribute to ongoing research efforts in breast cancer detection by delivering a system that enhances both accuracy and efficiency.

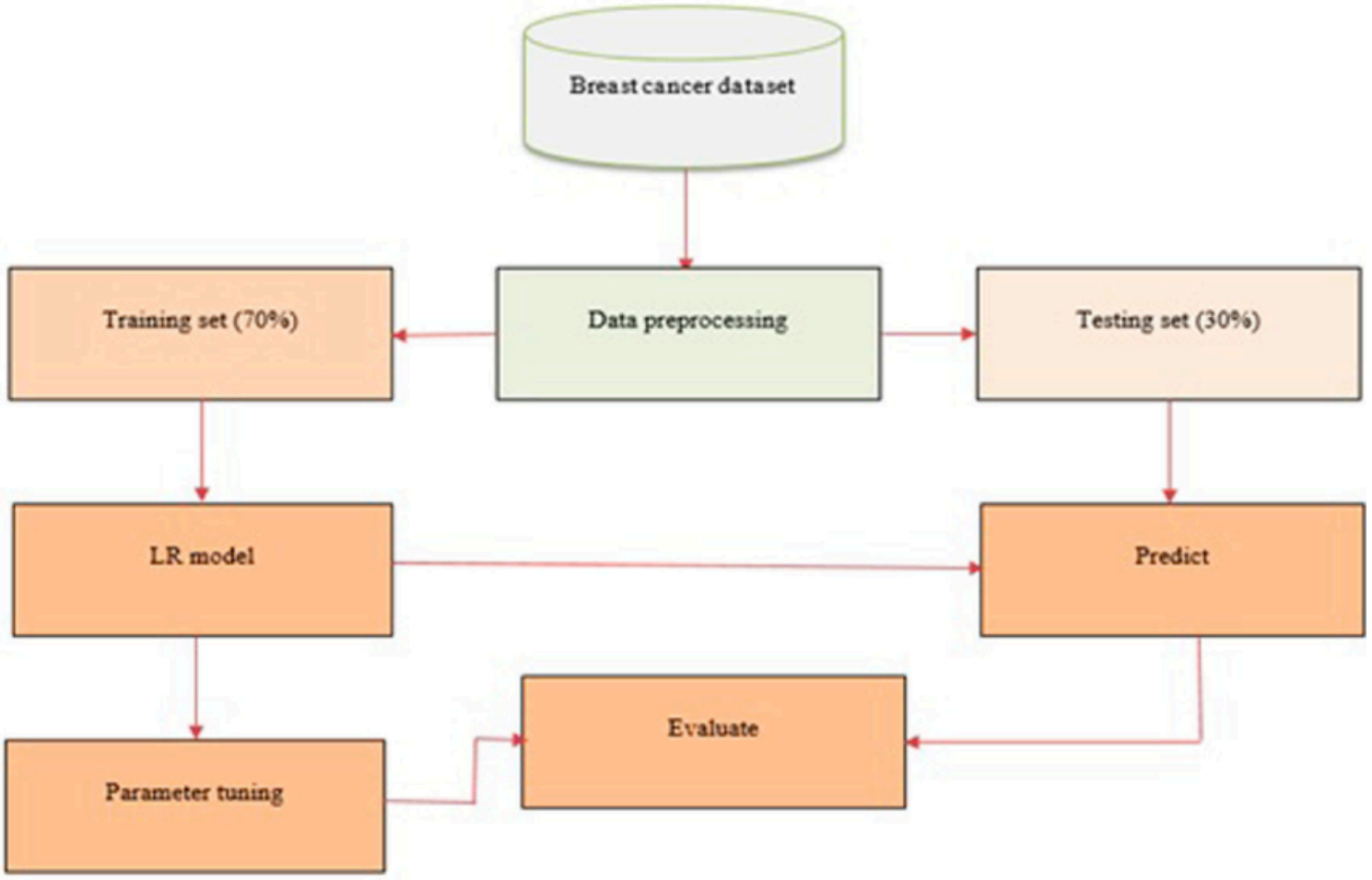
SCOPE OF THE PROJECT

- The scope includes the development of a project that aims to use machine learning models like Logistic Regression, Random Forest, and XGBoost to classify breast tumors as benign or malignant. It involves training the models on a breast cancer dataset, comparing their performance, and identifying the most accurate model for early detection. The results will help in developing a practical tool to assist in breast cancer diagnosis.

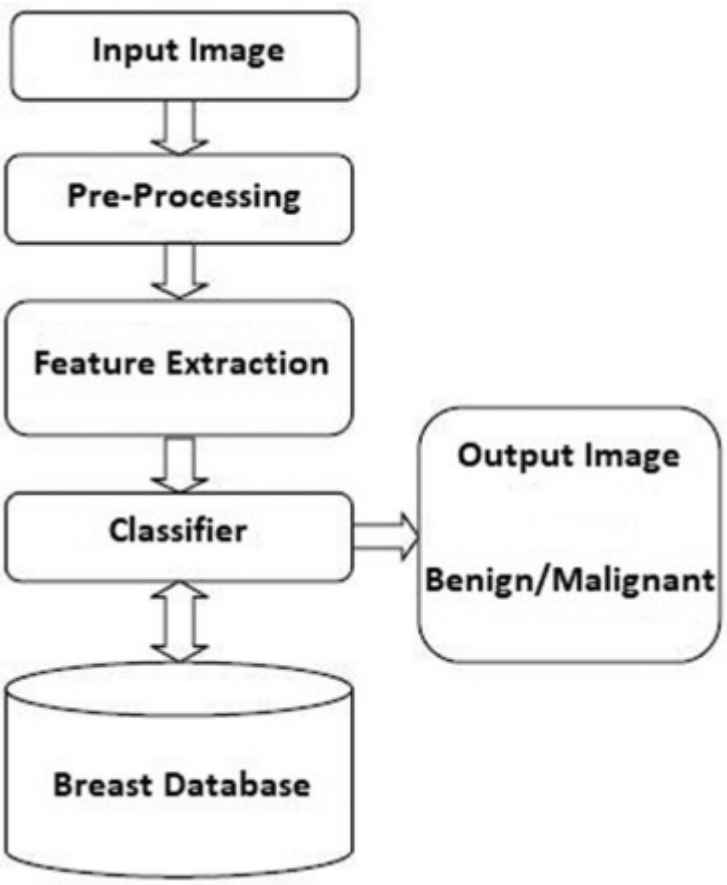
Literature Review and Gap Analysis

Deep learning models from Google Health have demonstrated that the system may perform badly on women from different backgrounds when training data is primarily composed of photographs from one population, such as white women. For under-represented populations, this bias may result in incorrect diagnoses. This problem is made worse by subtle differences in breast tissue density and composition between ethnic groups, which affect the precision of detection. In the meanwhile, integrating iCAD's SecondLook technology into current electronic health record (EHR) systems could involve extra stages, which frequently calls for major modifications to IT infrastructure and workflow in radiology departments. Another degree of complication is created by the absence of standardisation in the data formats and interfaces between CAD systems and EHRs, which makes interoperability difficult to achieve. Similar issues with storage within a hospital's IT infrastructure arise from the enormous data files generated by Hologic's 3D Mammography (Tomosynthesis). Data management and archiving expenses rise as a result of these increased storage requirements. These issues can be lessened by putting effective data compression strategies into practice and using cloud-based storage choices; however, when using cloud-based alternatives for sensitive medical photos, data security and privacy issues need to be carefully considered.

block diagram for proposed system



Design Flow Chart for the proposed problem



Conclusion and Future Plan

This project aims to develop the solution of our gap analysis to reduce diagnostic errors in breast cancer detection, contributing to better early detection and ultimately improving patient outcomes.

The project is designed to be technically feasible while still offering valuable insights into the application of advanced algorithms in medical diagnostics. The goal is to classify whether a tumor is benign or malignant based on medical data such as tumor size, texture, and other diagnostic features. By training these models on labeled data, we aim to help doctors and medical professionals in diagnosing breast cancer early, improving the chances of successful treatment. We will compare the performance of the models to find the one that works best for accurate and reliable prediction.

Future Plan

The following steps are planned for future development

1. **Data Collection and Preprocessing:** Additional datasets will be explored to enhance the model's robustness.

2. **Algorithm Refinement:** The nature-inspired optimization algorithms will be further fine-tuned to improve performance. Parameters will be adjusted to ensure optimal feature selection, and alternative algorithms may be tested for comparison.

3. **Model Evaluation:** The model will be evaluated on multiple performance metrics such as accuracy, precision, recall, and F1-score, ensuring that it performs well across a variety of breast cancer cases.

4. **Testing and Validation:** Rigorous testing will be conducted on the final model to assess its generalisation ability on unseen data.

5. **Report Finalization and Documentation:** The final report will document the project’s progress, methodology, results, and potential for future improvements.

References

Abou Tabl A, Alkhateeb A, ElMaraghy W, and Ngom A. 2017. Machine learning model for identifying gene biomarkers for breast cancer treatment 805 survival. Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; p. 607–11.

A Kamboj, P Tanay, A Sinha, et al.
Breast cancer detection using supervised machine learning: a comparative analysis
Springer, Singapore (2021), pp. 263-269

Alam, M. S. et al. Statistics and network-based approaches to identify molecular mechanisms that drive the progression of breast cancer. Comput. Biol. Med. 145, 105508 (2022).

Mihaylov I, Nisheva M, Vassilev D. Machine learning techniques for survival time prediction in breast cancer. Lecture Notes in Computer Science Springer, Cham. 2018: 186–94, doi:10.1007/978-3-319-99344-7_17.

World Health Organization. Breast Cancer (World Health Organization, 2021)