SYNOPSIS

Report on

BREAST CANCER DETECTION

by

Ekata Gupta (2300290140062)
Drishty (2300290140061)

Session:2024-2025(IV Semester)

Under the supervision of

Mr. R.N.Panda
Associate Professor

KIET Group of Institutions, Delhi-NCR, Ghaziabad

DEPARTMENT OF COMPUTER APPLICATIONS
KIET GROUP OF INSTITUTIONS, DELHI-NCR, GHAZIABAD-201206
(2023-2025)

ABSTRACT

Breast cancer is one of the most prevalent and life-threatening diseases affecting women worldwide. Early detection significantly increases survival rates and improves treatment outcomes. This project presents a machine learning-based approach to breast cancer detection using the Breast Cancer Wisconsin (Diagnostic) Dataset. The dataset consists of key features extracted from digitized breast mass images, allowing classification into benign or malignant categories.

The study explores multiple classification models, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and K-Nearest Neighbors (KNN). The methodology involves data preprocessing, feature selection, model training, and hyperparameter tuning to achieve optimal accuracy. Model performance is evaluated using key metrics such as accuracy, precision, recall, and F1-score to determine the most effective classification technique.

The objective of this project is to develop a reliable and automated system that assists healthcare professionals in detecting breast cancer at an early stage. By leveraging machine learning techniques, the proposed solution enhances diagnostic efficiency, reduces human error, and facilitates timely medical intervention. The results of this study demonstrate that machine learning can be a valuable tool in medical diagnostics, potentially improving patient outcomes and advancing healthcare technology.

TABLE OF CONTENTS

Introduction

Breast cancer is one of the leading causes of cancer-related deaths in women globally. It occurs due to the uncontrolled growth of abnormal cells in the breast tissue, which can be categorized as benign (non-cancerous) or malignant (cancerous). Early and accurate detection of breast cancer is crucial for effective treatment and increased survival rates.

Traditional diagnostic methods, such as mammography and biopsy, although effective, often require expert evaluation and can be time-consuming. With advancements in technology, machine learning has emerged as a powerful tool in medical diagnostics, offering a faster and more reliable approach to detecting breast cancer. Machine learning models can analyze large datasets, identify patterns, and provide high-accuracy predictions, aiding healthcare professionals in making informed decisions.

This project utilizes the **Breast Cancer Wisconsin (Diagnostic) Dataset** to train and evaluate multiple machine learning algorithms, such as **Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, and K-Nearest Neighbors (KNN)**. By implementing advanced data preprocessing techniques and model optimization strategies, the objective is to achieve a robust and efficient breast cancer detection system that enhances early diagnosis and supports clinical decision-making.

Literature Review

Several studies have highlighted the importance of machine learning in medical diagnostics, particularly in breast cancer detection. **Support Vector Machines (SVM), Decision Trees, and Random Forest** have been widely used in previous research due to their high classification accuracy. Studies have demonstrated that combining multiple algorithms can enhance the overall predictive performance.

Feature selection plays a crucial role in improving model accuracy and reducing computational complexity. Research has shown that selecting relevant features from datasets like the **Breast Cancer Wisconsin (Diagnostic) Dataset** significantly improves classification results. Additionally, **data balancing techniques** such as SMOTE (Synthetic Minority Over-sampling Technique) have been employed to handle class imbalance, ensuring models do not favor majority class predictions.

Furthermore, studies comparing deep learning and traditional machine learning methods suggest that while deep learning models like CNNs can provide high accuracy, they require extensive computational power and large datasets. In contrast, traditional models such as **Random Forest and SVM** offer a balance between accuracy and computational efficiency, making them suitable for medical applications.

This project builds upon these findings by implementing multiple classification algorithms, applying feature selection techniques, and optimizing hyperparameters to develop an accurate and efficient breast cancer detection model.

Project Objective

The primary objective of this project is to develop a machine learning-based breast cancer detection system that enhances early diagnosis and supports medical professionals in decision-making. The key objectives are:

- Data Acquisition and Preprocessing: Collect and preprocess the Breast Cancer Wisconsin (Diagnostic) Dataset, ensuring data quality and feature selection for optimal performance.
- Model Development: Implement various machine learning algorithms, including Logistic Regression, SVM, Decision Tree, Random Forest, and KNN, and evaluate their effectiveness in classifying tumors.
- Feature Engineering and Optimization: Utilize feature selection techniques and hyperparameter tuning to enhance model accuracy and reduce computational complexity.
- Performance Evaluation: Compare model accuracy, precision, recall, and F1-score to determine the most effective classification method.
- Deployment and Integration: Develop a user-friendly interface that allows healthcare professionals to input patient data and receive real-time diagnostic results.
- Medical Application and Impact: Ensure the developed model contributes to early detection, reducing human error and improving the efficiency of breast cancer diagnosis in clinical settings.

By fulfilling these objectives, this project aims to leverage machine learning to provide an efficient, accurate, and scalable solution for breast cancer detection, ultimately aiding in timely medical intervention and improved patient outcomes.

Project Flow

The project follows a systematic flow, starting with data collection from the publicly available Breast Cancer Wisconsin (Diagnostic) Dataset. The collected data undergoes preprocessing, including handling missing values, feature normalization, and selection to improve model efficiency.

Next, multiple machine learning models such as Logistic Regression, SVM, Decision Tree, Random Forest, and KNN are trained and evaluated. The models' performance is compared using metrics like accuracy, precision, recall, and F1-score. Hyperparameter tuning is performed to optimize the best-performing model.

Once an optimal model is identified, a user-friendly interface is developed to allow healthcare professionals to input patient data and receive real-time diagnostic results. Finally, thorough testing and validation ensure the model's reliability before deployment.

1. Data Collection: Using the publicly available Breast Cancer Wisconsin (Diagnostic) Dataset.
2. Data Preprocessing: Handling missing values, normalizing features, and feature selection.
3. Model Implementation: Training multiple classification algorithms (Logistic Regression, SVM, Decision Tree, Random Forest, KNN).
4. Model Evaluation: Comparing accuracy, precision, recall, and F1-score.
5. Optimization: Hyperparameter tuning for performance enhancement.
6. Deployment: Developing a user-friendly interface for diagnosis predictions.

Project Outcome

The outcome of this project is an efficient and automated breast cancer detection system that aids in early diagnosis. The machine learning model successfully classifies tumors as benign or malignant, providing high accuracy and reliability.

A comparative analysis of different machine learning algorithms determines the most effective classification approach. The integration of feature selection and hyperparameter tuning enhances the model's precision and reduces misclassification rates.

Additionally, a user-friendly interface is developed to assist healthcare professionals in making informed decisions. This project contributes to medical diagnostics by reducing human error, improving early detection, and enabling timely treatment. The final system serves as a cost-effective, scalable, and efficient tool for breast cancer detection, with the potential for future enhancements using larger datasets and deep learning techniques.

Proposed Time Duration

The project will be completed in approximately 6 weeks. The first 2 weeks will be dedicated to data collection and preprocessing, where the dataset will be cleaned, normalized, and relevant features selected.

During the next 2 weeks, multiple machine learning models such as Logistic Regression, SVM, Decision Tree, Random Forest, and KNN will be trained and evaluated. Their performance will be assessed based on accuracy, precision, recall, and F1-score.

Following this, 1 week will be spent on hyperparameter tuning and optimization to enhance model accuracy and efficiency. The optimal model will then undergo final testing and integration in the last week, ensuring its reliability and usability.

The project will conclude with documentation and report writing, summarizing the findings and preparing the system for potential enhancements. This structured timeline ensures an organized and efficient execution of the project.

REFERENCES/ Bibliography

[1] Shannon Doherty, Breast cancer analysis using lazy 2011learners https://www.webmd.com/breast-cancer/features/ shannen-doherty-breast-cancer

[2] M Navya Sri, ANIT, Analaysis of NNC and SVM for Machine Learning 2020

[3] N Gupta, Google Scholar, Prediction of Areolar cancer

[4] Jiaxin Li, Jilin University, 5year survival forpersonhaving-breast-cancer(2020).

[5] Mohammad Milan Islam, University of Waterloo, Prediction of residual diseases and breast cancer.2020 https:// link.springer.com/article/10.1007/s42979-020-00305-