TEXAS TECH UNIVERSITY SYSTEM™

Breast Cancer Prediction System Utilizing Machine Learning Algorithms

Presented By:

Shreyas Prabhakar	R11894057
Muhammad Talha Jabbar	R11914715
Teja Sri Dharma Reddy Vanukuri	R11974522
Rahul Chowdary Maddineni	R11976773
Sathwik Tatiparthi	R11976772

Under the guidance of Prof. Dr. Jingjing Yao

November 25, 2024

TABLE OF CONTENTS

Introduction

Leading cause of cancer-related deaths among women worldwide.

Early Detection Importance: Reduces mortality rates and improves treatment success.

Prediction System Goal: Develop an enhanced system using machine learning for accurate and reliable breast cancer prediction.

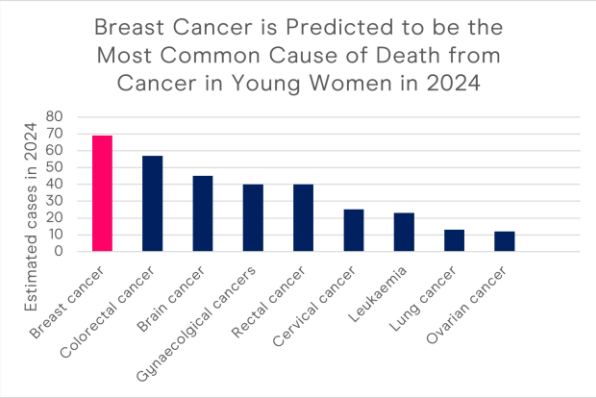


Figure 1: Source: Australian Institute of Health and Welfare (AIHW) 2024 Cancer Data in Australia.

Problem Statement

Challenges in Current Diagnosis Methods:

Traditional diagnostic techniques have limitations. Biopsies are time-consuming and stressful for patients, and imaging techniques may not detect tumors early. Existing prediction models are inefficient, with moderate accuracy and high false-positive and false-negative rates.

To improve prediction accuracy, we need to enhance the model and incorporate advanced algorithms to capture complex data patterns. This will reduce the risk of misdiagnosis, enable early intervention, and optimize treatment planning.

Figure 1: Image of Medical Professionals Performing a Biopsy

Previous work

1. Previous Research Goal:
   - Enhance early detection and prognosis of breast cancer using machine learning
2. Methods Used:
   - Algorithms like Random Forest, SVM, and Gradient Boosting, along with data preprocessing, feature selection, cross-validation, and ensemble approaches.
3. Evaluation Metrics:
   - Precision, recall, F1-score, and AUC-ROC.

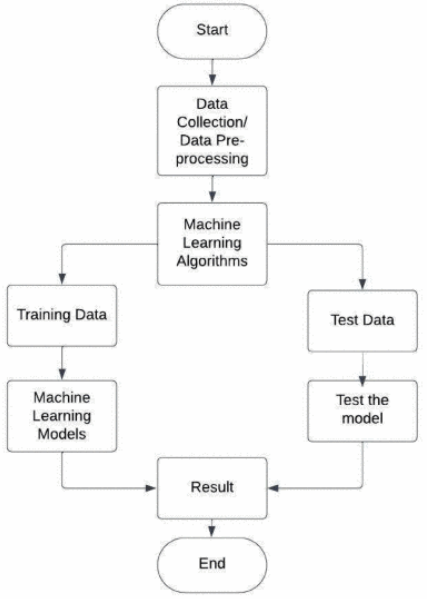


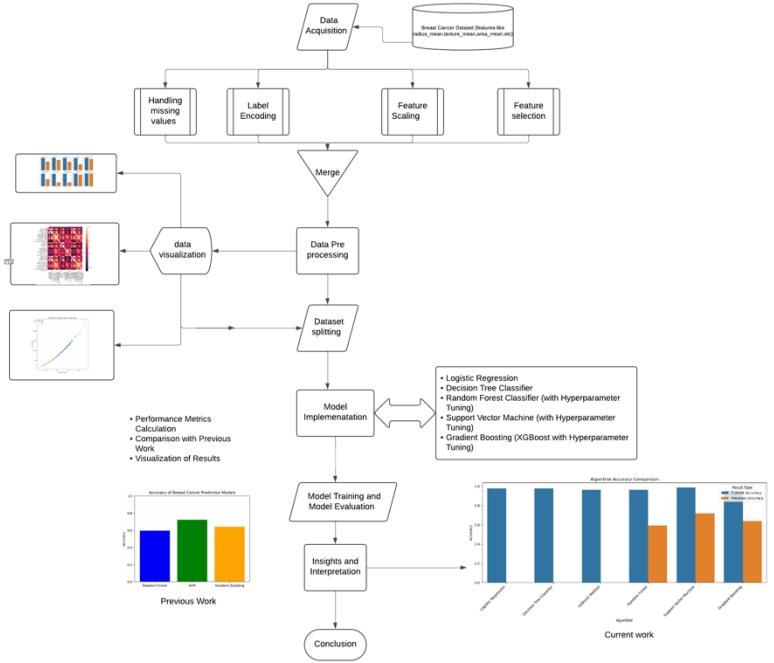
Fig3: working mechanism

Proposed Methodology

Aspect	Description
Improvement over Previous Studies	Incorporating more detailed features to enhance tumor classification.
Data Preprocessing	Normalization, label encoding, and handling missing values.
Feature Selection and Analysis	Utilizing scatter plot matrix, bar plot analysis, and correlation heatmap to identify key features and their correlation.
Algorithm Comparison	Compare Logistic Regression, Decision Tree, SVM, Random Forest, Gradient Boosting, and XGBoost, highlighting accuracy improvements and reasons for them.
Visualization of Results	Use bar plots, line plots, and scatter plots to visualize accuracy improvements across algorithms and discuss their significance.
Model Robustness and Efficiency	Discuss the robustness and efficiency of the proposed models, summarizing improvements in accuracy and reliability.

Improved Algorithms

- ✓ **Data Preprocessing Techniques:** Handling missing values, scaling and normalization, label encoding, addressing missing values and outliers, and data standardization.
- ✓ **Model Enhancement Techniques:** Implementing a Random Forest ensemble with optimized hyperparameters, applying various kernels for SVM, tuning hyperparameters, and enhancing decision boundaries in SVM.
- ✓ **Feature Engineering:** Employing visualization techniques to identify key features.



Experimental Results

Sl.No.	Algorithm	Current Accuracy	Previous Accuracy	Difference
1	Logistic Regression	0.976608	NaN	NaN
2	Decision Tree Classifier	0.976608	NaN	NaN
3	XGBoost Method	0.964912	NaN	NaN
4	Random Forest	**0.964912**	0.593	0.371912
5	Support Vector Machine	**0.988304**	0.720	0.268304
6	Gradient Boosting	**0.953216**	0.640	0.313216

Conclusion

- Model Performance: Optimized Support Vector Machine achieved 98.83% accuracy in breast cancer diagnosis prediction.
- Methodology: Enhanced accuracy through advanced machine learning, data preprocessing, hyperparameter tuning, and feature selection.
- Clinical Impact: Reliable and accurate predictive systems aid clinicians in informed decision-making and improve patient outcomes.

References and Acknowledgements

- We would like to express our sincere gratitude to the Doctors and National Breast Cancer Foundation for sharing invaluable insights into breast cancer globally. Their contributions have been invaluable.
- Additionally, we would like to acknowledge the UC Irvine Machine Learning repository for providing the Wisconsin Breast Cancer Dataset (WBCD). This dataset proved to be a crucial resource for our project.
- Texas Tech University for Providing access to HPCC

Q&A Slides

TEXAS TECH UNIVERSITY SYSTEM™