INDIRA COLLEGE OF SCIENCE AND COMMERCE

Breast Cancer Detection

S.Y M.Sc.(Comp.Sci) – Sem III

Team members:

1) SIDDHI GAVHANE (71)
2) SHIVANEE HASE (69)

Table Of Contents

1. Introduction

a) Background and Context

Breast cancer is one of the leading causes of death among women globally. Early detection is crucial for improving survival rates. Traditional diagnostic methods such as biopsies and mammograms, while effective, can be time-consuming, costly, and prone to human error. With the advancement of technology, machine learning (ML) offers a promising approach for automated, accurate, and early detection of breast cancer. ML algorithms can analyze large datasets to identify patterns and correlations that may be invisible to the human eye.

b) Purpose and Objectives

The main purpose of this study is to build an intelligent system capable of accurately detecting breast cancer using machine learning models.

Objectives:

- To understand the features that differentiate malignant from benign tumors.
- To explore and apply various machine learning algorithms on breast cancer datasets.
- To evaluate and compare model performance using metrics like accuracy, precision, recall, and F1-score.
- To develop a predictive tool for healthcare professionals to assist in early diagnosis.

c) Research Questions

- What features in medical data are most predictive of breast cancer?
- Which ML algorithms are most effective for classification of breast cancer?
- How can ML help reduce false positives/negatives in breast cancer diagnosis?
- Can a generalized model be developed that is both accurate and efficient?

d) Significance

The project has the potential to significantly aid in the early detection of breast cancer, improving treatment outcomes and survival rates. Automated systems reduce the burden on radiologists and clinicians, ensuring quicker, data-driven decision-making.

2. Literature Review

a) Overview of Previous and Current Research

The use of Machine Learning (ML) for early breast cancer detection has expanded rapidly in recent years. Notable research trends include:

1.Traditional ML Approaches:

- Early studies used classical ML models like Logistic Regression, SVM, Decision Trees, and Random Forests on feature-based datasets like WDBC (Wolberg et al., 1992).
- Dheeba et al. (2014) explored SVMs with image-based features and showed promising results in mammogram-based detection.

2. Deep Learning and CNN-based Methods (2020–2025)

- Recent studies have used Convolutional Neural Networks (CNNs) to analyze mammograms and histopathology images.
- A 2023 study by Zhou et al. used ResNet-50 combined with attention mechanisms, achieving ~97% accuracy in multi-class classification tasks.

3. Attention-Based Transformers (2024–2025)

- In 2024, Vision Transformers (ViTs) began gaining popularity in medical imaging tasks due to their ability to capture long-range dependencies.
- A 2025 paper by Kumar et al., published in *Nature Machine Intelligence*, used ViT-HealthNet for breast cancer subtype classification and achieved state-of-the-art AUCs (0.98+).

4. Explainable AI (XAI)

- Due to ethical and medical accountability, recent research focuses on explainability using methods like LIME and SHAP to visualize tumor-related features.

b) Latest Datasets (Women/Girls above Age 28)

The following datasets are more updated and age-relevant than the classic WDBC dataset used in your paper:

Dataset Name	Year	Age Focus	Description
TCGA-BRCA	Ongoing	28–80+	From The Cancer Genome Atlas. Includes genomic + clinical data for women only with detailed age metadata.
CBIS-DDSM	Updated 2022	25–90	Digitized mammograms; includes BI-RADS category, age, and pathology. Includes women over 28.
BreakHis	2023	25–75	Microscopic biopsy images. Labeled as benign or malignant. Images categorized by age and subtype.
VinDr-Mammo	2021–2024	30+	Vietnamese dataset with 5,000+ full-field digital mammograms and verified labels.

Most of these datasets focus only on women and contain explicit age metadata, often segmented in the 28–80+ age range.

c) Theoretical Framework – 2025 Perspective

Current breast cancer detection models are rooted in Supervised Learning with an increasing trend toward Self-Supervised Learning (SSL) and Transfer Learning. Modern models utilize:

- Pretrained Models (e.g., ResNet, EfficientNet, ViT)
- Ensemble Techniques for performance boosting
- XAI for ethical medical interpretation

3. Methodology

a) Research Design

The research follows a data-driven experimental design using quantitative methods. The process includes:

1. Data acquisition
2. Data preprocessing
3. Model training
4. Evaluation and validation

b) Data Analysis

- **Dataset:** Wisconsin Diagnostic Breast Cancer (WDBC) dataset
- **Preprocessing:** Handling missing data, normalization, train-test split
- **Feature Selection:** Correlation matrix and recursive feature elimination

Algorithms Used:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- K-Nearest Neighbor (KNN)
- Naive Bayes
- Neural Networks (MLP)

Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1 Score

- ROC-AUC Curve

c) Ethical Considerations

- Only publicly available datasets used
- Data anonymization ensured
- Compliance with ethical research practices
- Transparency in reporting results and limitations

4. Expected Results

a) Hypotheses

- ML models will achieve prediction accuracy between 90% to 99%.
- Ensemble methods and deep learning models will outperform traditional models.

b) Predicted Outcomes

- Identification of key predictive features such as radius mean, texture mean, concavity, etc.
- Random Forest and Neural Network models are expected to provide highest accuracy.
- Development of a lightweight, deployable tool for hospital use.
- Reduced false positives and negatives through improved model calibration.

5. Conclusion

a) Summary

This project uses machine learning for breast cancer detection. It utilizes clinical datasets to build and test various ML models to determine which algorithms provide the most reliable diagnosis support.

b) Implications

Accurate prediction tools can greatly improve patient outcomes by enabling timely treatment. This work also contributes to the growing field of medical AI.

c) Conclusion

The study demonstrates the feasibility and value of machine learning in the medical domain. It highlights the importance of data quality, feature engineering, and algorithm selection in achieving high performance in predictive healthcare models.

6. References

1. • **Chen & Martel (2025)**: *Enhancing breast cancer detection on screening mammogram using self-supervised learning and a hybrid deep model of Swin Transformer and Convolutional Neural Network*.
   Introduces **HybMNet**, a hybrid SSL + Swin-CNN model achieving AUCs of ~0.864 (CMMD dataset) and 0.889 (INbreast), demonstrating improved detection performance using limited labeled data Wikipedia+15arXiv+15Nature+15.
2. • **Park et al. (2025)**: *A Multi-Modal AI System for Screening Mammography: Integrating 2D and 3D Imaging to Improve Breast Cancer Detection*.
   Trained on ≈500k exams (later expanded), this system integrates FFDM and DBT, achieving AUC ≈ 0.945, reducing recalls by ~32%, and lowering radiologist workload ~44% while maintaining 100% sensitivity arXiv.
3. • **Patel et al. (2025)**: *GAIN-BRCA: a graph-based AI-net framework for breast cancer subtype classification*.
   Uses graph learning on multi-omic (mRNA, miRNA, methylation) data from TCGA, improving subtype prediction in precision oncology contexts Oxford Academic.
4. • **Su et al. (2025)**: *Computational Pathology for Accurate Prediction of Breast Cancer Subtypes Using TCGA-BRCA*.
   Explores high-dimensional genomic and histopathology integration, covering 1,006 patients including HR+/HER2− cases ETASR+10ScienceDirect+10portal.gdc.cancer.gov+10.
5. • **Kallah-Dagadu et al. (2025)**: *Breast cancer prediction based on gene expression data (1,208 samples, 3,602 genes)*.
   Combines KNN, RF, SVM with feature selection and explainable ML (SHAP, ALE) to identify and interpret predictive genomic markers Nature+2Nature+2Nature+2.
6. • **Chowdhury et al. (2025)**: *An efficient and interpretable ML model for classifying BRCA subtypes using RNA-seq data*.
   Proposes an interpretable framework using high-dimensional transcriptomics for subtype classification with explainability features ETASR+1Nature+1.
7. • **Ghasemi et al. (2024/2025)**: *Explainable Artificial Intelligence in Breast Cancer Detection and Risk Prediction: A Systematic Scoping Review*.
   Synthesizes XAI techniques across breast cancer ML research (covering 2017–2023), highlighting that **SHAP** is the most commonly used model-agnostic technique arXiv+1Nature+1.
8. • **Khan et al. (2025)**: *A comprehensive review of machine learning and deep learning techniques in breast cancer detection and diagnosis*.
   Delivers a broad meta-analysis addressing intra-class variance in various imaging and ML pipelines ScienceDirect.

9. • **Das et al. (2025)**: *Comprehensive bioinformatics and machine learning utilizing TCGA gene expression data for breast cancer detection and categorization*.
   Uses large-scale transcriptomics to build predictive models linking genomics with clinical diagnosis Oxford Academic.
10. • **Nasir et al. (2025)**: *Breast Cancer Detection Using Convolutional Neural Networks: Recent Advances*.
   Reviews CNN architectures such as ResNet, EfficientNet, demonstrating their high accuracy and deployment possibilities in clinical workflows