IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Early Detection Of Breast Cancer Using Different Techniques

[1]Dr. Sachinkumar, [2]Ms.Radhika Mithari, [3]Mr.Omkar Budruk, [4]Mr. Saket Raj Singh, [5]Ms. Saakshi Shetty
[1]Associate Professor, [2]UG Student, [3]UG Student,[4]UG Student,[5]UG Student
[1]Department of Computer Science & Engineering,
[1]Jain College of Engineering, Belagavi, Karnataka, India

Abstract: Breast cancer is one of the most prevalent and life-threatening diseases among women, necessitating early detection for better treatment outcomes. With advancements in machine learning (ML) and deep learning (DL), automated systems for breast cancer prediction have gained significant attention. This literature review explores various methodologies employed for breast cancer detection, including both traditional ML algorithms and cutting-edge deep learning techniques. A variety of approaches, such as tumor segmentation using Conditional Generative Adversarial Networks (cGANs), feature extraction with Convolutional Neural Networks (CNNs), and hybrid models like CNN-SVM, have been extensively studied. Additionally, techniques such as transfer learning, whale optimization algorithms, and autoencoders have been used to improve model accuracy and efficiency. While these methods demonstrate high potential, challenges related to dataset quality, interpretability, computational cost, and generalization remain. This review highlights the strengths and limitations of these approaches, providing a comprehensive understanding of the current landscape in breast cancer prediction. The findings suggest that hybrid models, optimization techniques, and deep learning-based feature extraction offer promising results, but addressing computational and data-related challenges is crucial for real-world implementation.

CHAPTER 1 INTRODUCTION

Cancer is a global health issue caused by abnormal and uncontrolled cell growth that can spread through the body. It contributes to millions of deaths annually, with 990 new cases diagnosed daily. Over 100 types of cancer exist, classified by origin and cell type:

- Carcinomas: Most common; affect skin and internal organs (e.g., breast, lung).
- Sarcomas: Arise in connective tissues (e.g., bones, muscles).
- Leukemia: Blood cancer disrupting normal blood cell production.
- Lymphomas/Myelomas: Affect immune system cells.
- CNS Cancers: Originate in the brain/spinal cord.

Causes include genetic mutations, carcinogen exposure (e.g., tobacco, UV), unhealthy lifestyles, and infections like HPV and Hepatitis B/C.

Breast Cancer Overview

Breast cancer, affecting ducts or lobules, accounts for 15% of global cancer cases and is a leading cause of cancer-related deaths in women. Risk factors include age, genetic mutations (BRCA1/2), hormonal therapy, obesity, and family history. Symptoms include breast lumps, changes in size/shape, skin dimpling, and nipple discharge.

Emerging Treatments

Advances in surgery, chemotherapy, and radiation improve survival but often have side effects. Cannabinoids (e.g., CBD oil) are being studied for their anti-inflammatory, anti-tumor, and pain-relief properties, showing promise in inhibiting cancer growth and improving treatment outcomes. Holistic approaches like CBD may offer additional relief for breast cancer patients.

CHAPTER 2 LITERATURE SURVEY

- Tumor Segmentation and Classification Using Conditional GANs and CNNs by VK Singh et al. (2020): Utilizes cGANs for segmentation and CNNs for tumor shape classification but requires large labeled datasets.
- SVM with Median Filtering and Wavelet Transform by A.M. Ibraheem et al. (2021): Employs SVM with feature extraction techniques but struggles with noisy datasets.
- Transfer Learning with Pre-Trained Models by S. Khan et al. (2020): Leverages CNNs like ResNet and VGGNet for feature extraction but is computationally expensive.
- Deep Learning Integrated with SVM by D.A. Ragab et al. (2019): Combines CNN-based feature extraction with SVM classification but is limited to binary classification.
- CNN for Histopathology Image Analysis by S.A. Alanazi et al. (2022): Achieves high accuracy in detecting malignant regions but requires substantial computational resources.
- Whale Optimization Algorithm with MLP by H. Fang et al. (2018): Optimizes MLP for mammogram analysis but is sensitive to hyperparameter tuning.
- Autoencoders for Breast Lesion Detection by M. G120 et al. (2021): Utilizes autoencoders for feature learning but requires extensive data for training.
- Hybrid CNN-SVM Model by N. Chouhan et al. (2019): Combines CNN for feature extraction with SVM for classification, achieving higher accuracy.
- Breast Cancer Classification Using XGBoost by Liu et al. (2019): Demonstrates robust predictions but requires careful hyperparameter tuning.
- Ensemble Techniques for Breast Cancer Prediction by Various Authors (2020): Highlights Random Forests and XGBoost for accuracy but notes computational intensity.

CHAPTER 3 PROBLEM IDENTIFICATION

Gap Identification

Breast cancer remains a leading cause of mortality among women. Key gaps in current diagnostics include:

1. **Limited Accessibility**: High costs and limited availability of advanced diagnostic tools.
2. **Data Complexity**: Difficulty in manually analyzing large, complex datasets.
3. **Accuracy Issues**: High rates of false positives and negatives in existing methods.
4. **Underutilized ML Potential**: Lack of scalable and accurate machine learning models for prediction. This project aims to address these gaps with an accurate, accessible, and interpretable ML-based prediction system.

Problem Statement

Current breast cancer diagnostics are costly, invasive, and inefficient for handling high-dimensional data, limiting large-scale use in resource-constrained settings. This project proposes a robust ML system to classify malignant and benign tumors using feature selection, preprocessing, and algorithm optimization for practical applications.

Scope

1. **Data**: Use public datasets like the UCI Breast Cancer Wisconsin dataset.
2. **Algorithms**: Implement and evaluate ML models (e.g., logistic regression, SVM, random forests).
3. **Feature Selection**: Focus on relevant features for better accuracy and reduced complexity.
4. **Optimization**: Apply hyperparameter tuning and cross-validation.
5. **Usability**: Build a scalable system for healthcare integration.

CHAPTER 4 OBJECTIVES

1. **Data Understanding and Preprocessing:**
   - Load, clean, and preprocess the breast cancer dataset to ensure it is suitable for model training and evaluation.
2. **Exploratory Data Analysis (EDA):**
   - Use statistical and visual methods to explore relationships in the data, such as correlations between features and the target variable.
3. **Feature Engineering:**
   - Select or derive the most relevant features for building the prediction model.
   - Normalize or scale data to improve model performance.
4. **Model Building:**
   - Train various machine learning models (e.g., Random Forest, Logistic Regression, etc.) to classify breast cancer cases.
5. **Model Evaluation:**
   - Use metrics like accuracy, precision, recall, F1-score, and ROC-AUC to evaluate the models' performance.
6. **Deployment and Prediction:**
   - Save the best-performing model for real-world deployment.
   - Enable predictions on new, unseen data.
7. **Improve Decision Support:**
   - Assist clinicians by reducing diagnostic errors and improving the reliability of result.

CHAPTER 5 SYSTEM REQUIREMENTS

- **SOFTWARE REQUIREMENTS**
  - Operating System: Windows 10+, Linux Ubuntu 20.04+, macOS Catalina+.
  - Programming: Python, HTML/CSS.
  - Frameworks: Flask/Django, scikit-learn, XGBoost, pandas, matplotlib.
  - Database: SQLite.

- **HARDWARE REQUIREMENTS**
  - Minimum: Dual-core 2.5 GHz processor, 8GB RAM, 256GB SSD.
  - Recommended: Quad-core 3.0 GHz processor, 16GB RAM, CUDA-enabled GPU, 512GB SSD.

- **FUNCTIONAL REQUIREMENTS**
  - Input CSV datasets for preprocessing and model training.
  - Train and compare models (Logistic Regression, SVM, etc.).
  - Predict outcomes ("Benign/Malignant") with confidence scores.
  - Visualize results (confusion matrices, ROC curves, feature importance).
  - Provide real-time predictions through a web app/API.

- **NON-FUNCTIONAL REQUIREMENTS**
  - Performance: Predictions in ≤2 seconds; training ≤10 minutes for ≤10,000 rows.
  - Scalability: Support datasets up to 100,000 rows.
  - Security: End-to-end encryption, GDPR/HIPAA compliance.
  - Usability: Intuitive UI, minimal input requirements.
  - Reliability: 99% uptime.
  - Portability: Deployable locally or on cloud environments.

CHAPTER 6 METHODOLOGY & IMPLEMENTATION

- **DATA COLLECTION AND PREPROCESSING**
  - Dataset Acquisition: Obtain breast cancer datasets with tumor characteristics from credible sources like UCI Machine Learning Repository or Kaggle.
  - Data Cleaning: Handle missing values using imputation techniques (mean, median) or remove records with excessive missing data. Remove duplicates to ensure data integrity.
  - Encoding Categorical Variables: Convert categorical data like "diagnosis" into numerical form using label encoding or one-hot encoding.
  - Outlier Detection: Use statistical methods like Z-score or IQR to identify and remove data anomalies.

- **EXPLORATORY DATA ANALYSIS (EDA)**
  - Visualization: Use histograms, box plots, and density plots to examine feature distributions and class balance.
  - Correlation Analysis: Generate a heatmap to identify relationships among features, flagging multicollinearity.
  - Class Imbalance: Address imbalances using techniques like SMOTE if the data is skewed between benign and malignant cases.

- **FEATURE SELECTION AND SCALING**
  - Feature Importance: Use algorithms like Random Forest to identify key features influencing predictions.
  - Dimensionality Reduction: Apply PCA to reduce redundancy while retaining essential variance.

- Scaling: Standardize features using StandardScaler for models like SVM or normalize using MinMaxScaler for neural networks.

- **MACHINE LEARNING MODEL DEVELOPMENT**
  - Algorithm Selection:
  - Logistic Regression as a baseline.
  - Advanced models like SVM, Random Forest, and Gradient Boosting algorithms (e.g., XGBoost, LightGBM).
  - Neural Networks for non-linear relationships.
  - Training: Train models on 80% of the dataset and validate using cross-validation.
  - Hyperparameter Tuning: Optimize model parameters using Grid Search or Random Search.

- **MODEL EVALUATION AND COMPARISON**
  - Evaluation Metrics: Use accuracy, precision, recall, F1-score, and ROC-AUC for model assessment.
  - Result Visualization: Plot confusion matrices, ROC curves, and precision-recall curves for detailed analysis.
  - Model Selection: Choose the model with the highest performance and lowest computational cost.

- **WEB APPLICATION DEVELOPMENT**
  - Frontend Design: Develop an intuitive, responsive interface using Bootstrap.
  - Backend Integration: Build RESTful APIs with Django for handling model predictions and user inputs.
  - Database Setup: Use SQLite to store user interactions and model output for easy retrieval.

- **SECURITY MEASURES**
  - Data Protection: Implement end-to-end encryption for sensitive patient data.
  - Role-Based Access: Restrict data access based on user roles (e.g., admin, doctor).
  - Audits: Conduct regular system audits to ensure compliance with data security standards.

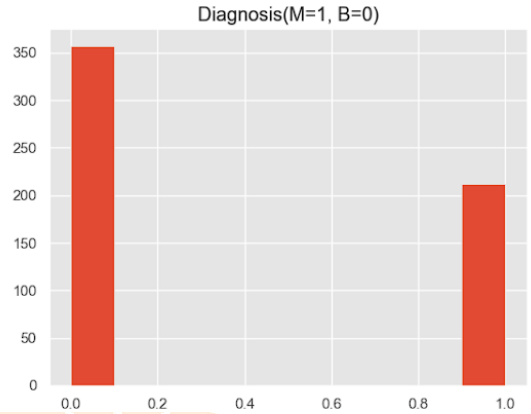
TESTING AND VALIDATION

- Unit Testing: Validate individual components like preprocessing modules, APIs, and model predictions.
- Integration Testing: Ensure seamless interaction between frontend, backend, and database.
- System Testing: Assess the complete workflow for performance and reliability.

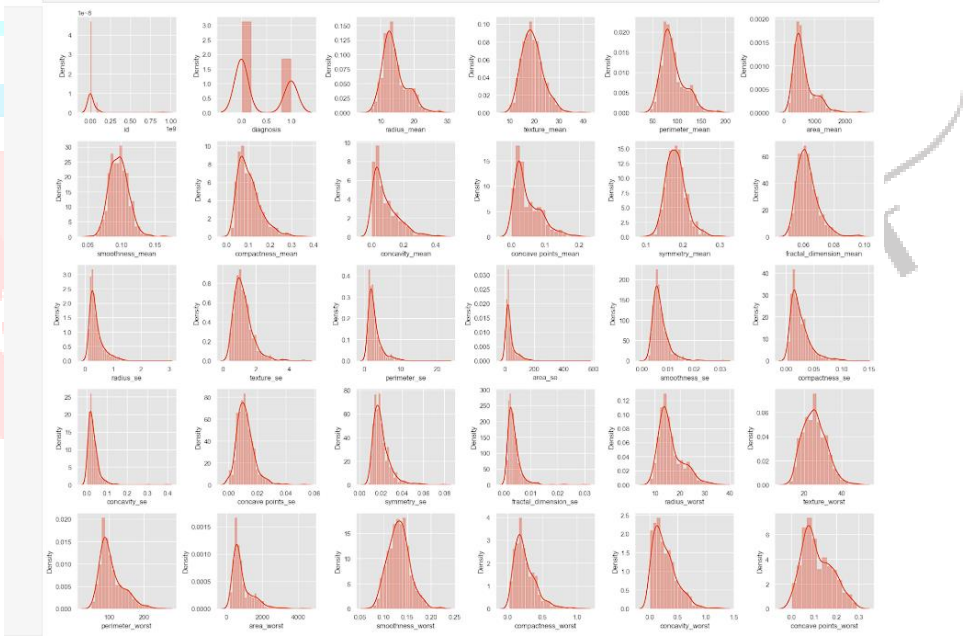
- **DEPLOYMENT**
  - Cloud Deployment: Deploy the application on cloud platforms like AWS or Azure for scalability and availability.
  - Containerization: Use Docker to create isolated, consistent deployment environments.
  - Monitoring: Implement logging and monitoring tools to ensure the system operates smoothly post-deployment.

CHAPTER 7 RESULTS

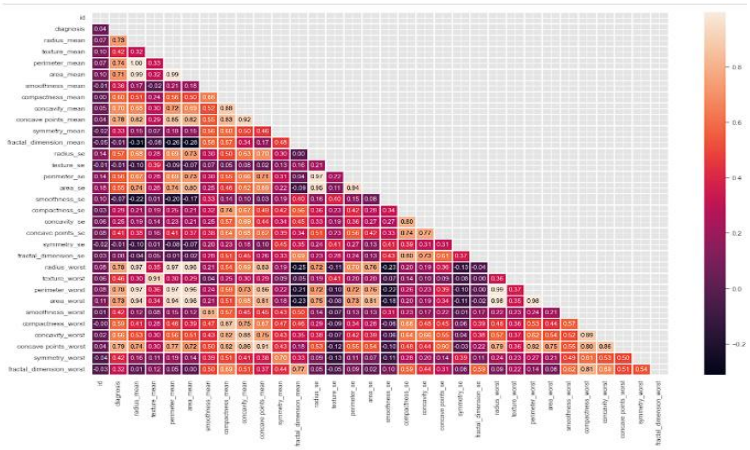
➢ **Diagnosis**



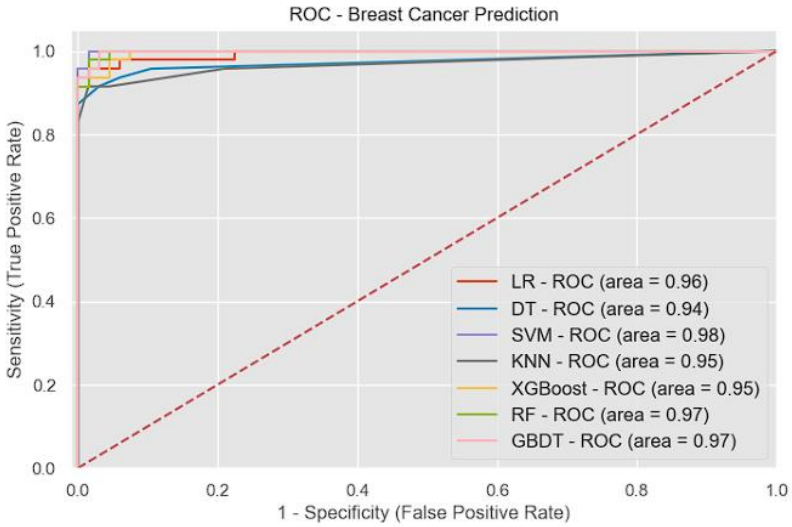
➢ **Density graph**



- **Heatmap**

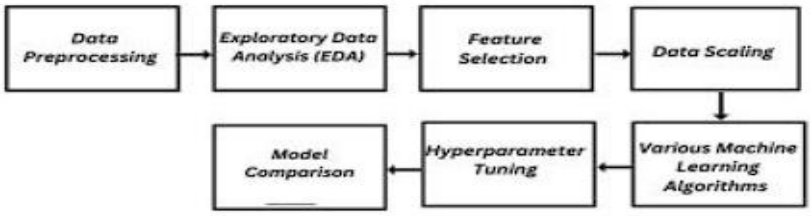


- **ROC**

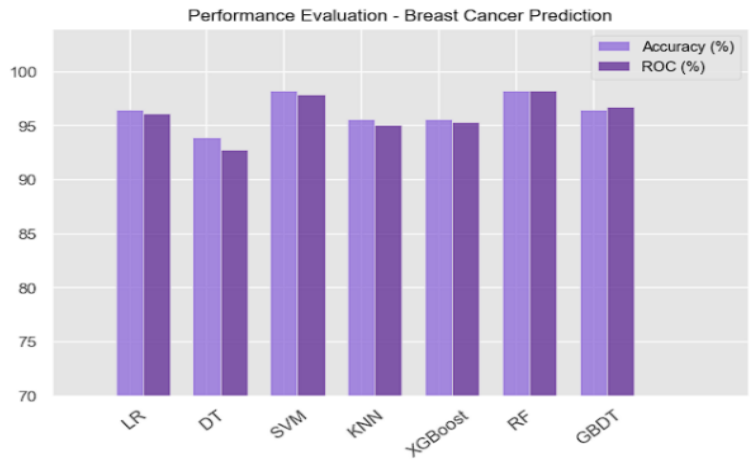

ROC - Breast Cancer Prediction
Sensitivity (True Positive Rate)
1 - Specificity (False Positive Rate)
0.0
0.2
0.4
0.6
0.8
1.0
LR - ROC (area = 0.96)
DT - ROC (area = 0.94)
SVM - ROC (area = 0.98)
KNN - ROC (area = 0.95)
XGBoost - ROC (area = 0.95)
RF - ROC (area = 0.97)
GBDT - ROC (area = 0.97)

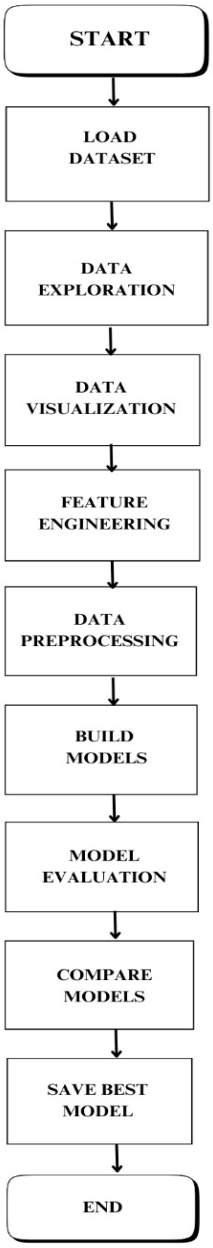
- **Performance Evaluation**

CHAPTER 8 DIAGRAMS



- **Flowchart:**





CHAPTER 9 CONCLUSION

- Accurate Diagnosis: The implementation of machine learning models has demonstrated significant potential in predicting breast cancer, offering a reliable distinction between malignant and benign tumors.

- Feature Importance: The analysis highlights key features, such as tumor size, texture, and smoothness, that contribute most to the predictive performance.

- Performance Metrics: Metrics like accuracy, precision, recall, and F1-score confirm the effectiveness of the chosen model(s). This validates the applicability of AI in assisting medical professionals with early and accurate diagnosis.

- Impact on Healthcare: The model's predictions can augment clinical workflows, reducing diagnostic errors and facilitating quicker decision-making, ultimately leading to better patient outcomes.

CHAPTER 10 APPLICATION & FUTURE SCOPE

➢ APPLICATIONS

1. Diagnosis Assistance: Enhances early breast cancer detection and accuracy.
2. Data Analysis: Identifies key features and trends in medical data.
3. Predictive Modelling: Assesses risk and compares cancer prediction models.
4. Education: Demonstrates machine learning applications in healthcare.
5. Healthcare Optimization: Automates tasks and prioritizes urgent cases.
6. Personalized Medicine: Tailors treatment and disease management.
7. Clinical Trials: Improves patient selection and predicts treatment outcomes.
8. Healthcare Strategy: Supports cancer screening programs and public health planning.
9. Digital Integration: Integrates AI in EHRs and telemedicine for real-time insights.
10. Cost Efficiency: Reduces unnecessary procedures and operational costs.
11. R&D: Identifies new biomarkers and refines ML algorithms.
12. Community Impact: Promotes awareness and improves access in underserved areas.

➢ FUTURE SCOPE

1. Enhance accuracy with advanced methods like deep learning.
2. Integrate genetic, lifestyle, and environmental data for better predictions.
3. Develop explainable AI models to increase trust.
4. Integrate with EHRs for real-time diagnostics in clinical settings.
5. Deploy in low-resource areas using cloud/mobile solutions.

CHAPTER 11 REFERENCES

- **Dataset Sources**
  1. UCI Machine Learning Repository. “Breast Cancer Wisconsin (Diagnostic) Dataset.” [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)]
- **Machine Learning Frameworks**
  1. Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research 12 (2011): 2825-2830. [https://scikit-learn.org/]
  2. TensorFlow Team. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." [https://www.tensorflow.org/]
- **Visualization and Data Analysis Tools**
  1. Hunter, J. D. "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering 9.3 (2007): 90-95.[https://matplotlib.org/]Waskom, M., et al. "Seaborn: Statistical Data Visualization."
  2. [https://seaborn.pydata.org/]

- **Textbooks and Articles**
  1. Hastie, T., Tibshirani, R., and Friedman, J. "The Elements of Statistical Learning." Springer Series in Statistics, 2009. [https://web.stanford.edu/~hastie/ElemStatLearn/]
- **Other Online Resources**
  1. Kaggle. "Breast Cancer Detection with Machine Learning." [https://www.kaggle.com/]