Performance Analysis of ML Techniques for Identifying the Breast Cancer

Presented by:

Rekapalli Bhagya Srilakshmi

Department of Electronics and Communication Engineering
Koneru Lakshmaiah Education Foundation, Hyderabad, India

Co-authors -Bittu Kumar, Kanikireddy Harshitha, Pulikollu Suma Harshitha, Neelakantam Aashritha

2025 International Conference on Electrical and Electronics Engineering (ICE3-2025)

Contents

- Introduction
- Related work
- Dataset Description
- Proposed Methodology
- Evaluation Parameters
- Results and discussion
- Conclusion
- References

Introduction

- **Breast cancer** is the most common cancer among women worldwide.
- Early detection is important for improving **prognosis** and reducing mortality.
- **Risk factors**:
  - Non-modifiable – age, **BRCA1/BRCA2 mutations**, family history
  - Modifiable – obesity, alcohol consumption, hormonal therapy
- **Histological subtypes**:
  - Invasive Ductal Carcinoma **(IDC)** – 80% cases
  - Invasive Lobular Carcinoma **(ILC)** – 10–15% cases
  - Ductal Carcinoma in Situ **(DCIS)** – early, non-invasive
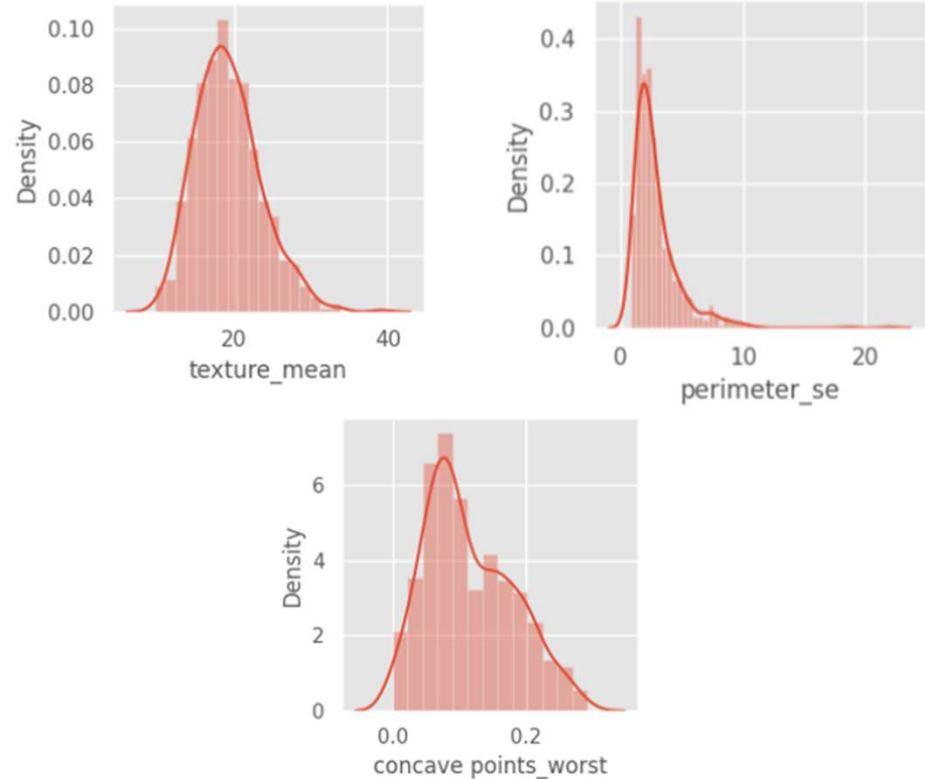- **Objective:** To compare **Machine Learning algorithms** for breast cancer classification.
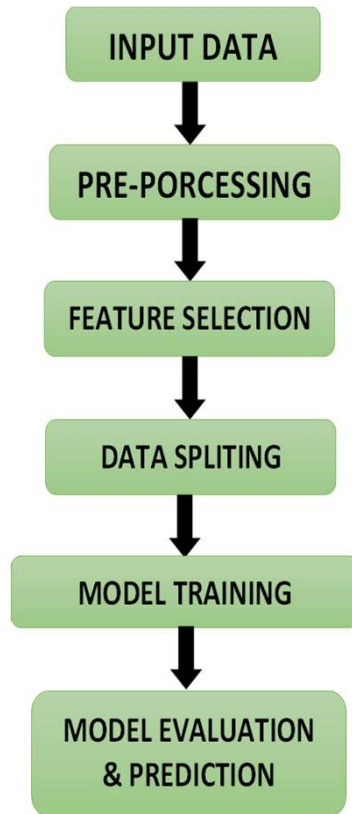
Related Work

- **Support Vector Machine (SVM)** reported as highly effective in breast cancer classification.
- **Convolutional Neural Networks (CNNs)** useful for analyzing medical imaging (e.g., mammography).
- Hybrid models combining feature selection and classifiers improve accuracy.
- **Ensemble methods** (e.g., **Random Forest, AdaBoost**) achieved accuracy up to **97.5%**.
- Research highlights importance of **model optimization** and **ensemble learning** for reliable results.

Dataset Description

•**Source:** Wisconsin Breast Cancer Dataset

•**Cases:** 570 → 210 malignant (M), 360 benign (B)

•**Format:** Excel file with 33 columns
  • 31 tumor-related attributes
  • 1 target column (diagnosis: M / B)
  • 1 ID column (not used)

•**Features grouped into:**
  • **Mean values** (radius mean, texture mean, etc.)
  • **Standard Error (SE)** values
  • **Worst-case values** (extreme tumor measurements)

Proposed Methodology

- **Input Data:** Tumor features (benign / malignant)
- **Preprocessing:** Z-score normalization, median imputation, label encoding
- **Feature Selection:** Correlation filter + mutual information
- **Data Split:** 80% train, 20% test (stratified)
- **Models:** LR, SVM, Random Forest, XGBoost, k-NN, Decision Tree
- **Metrics:** Accuracy, Precision, Recall, F1-score, ROC-AUC
- **Validation:** McNemar's test

Evaluation Parameters

- **Accuracy** – Overall correctness

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** – Reliability of positive (malignant) predictions

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity)** – Ability to detect true malignant cases

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score** – Balance between Precision & Recall

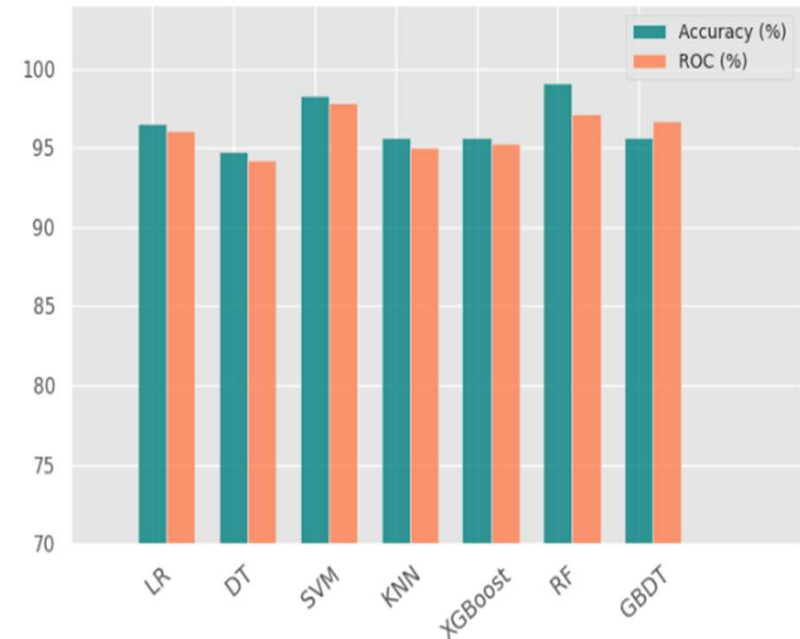$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- **ROC-AUC** – Discriminative ability across thresholds

Results and Discussion

Models	Precision	Recall	F1-Score
DT	0.96	0.96	0.96
SVM	0.98	0.98	0.98
KNN	0.96	0.96	0.96
RF	0.99	0.99	0.99
GBDT	0.96	0.96	0.96
LR	0.97	0.96	0.96
XG-Boost	0.96	0.96	0.96

Results and Discussion

- Compared **7 ML models**: LR, RF, SVM, KNN, XGBoost, DT, GBDT
- **Random Forest (RF):** Best performer → Precision, Recall, F1 = **0.99**
- **SVM**: High **ROC-AUC**, excellent at class separation
- **Other models**: Accuracy ≈ **0.96–0.97** (good but less consistent)
- **Conclusion:**
- RF → most dependable & consistent model
- SVM → strongest discriminative ability

Conclusion

- Machine learning enhances **accuracy & efficiency** in breast cancer detection
- **Random Forest (RF):** Best overall → Precision, Recall, F1 = **0.99**
- **SVM:** Strongest in discrimination (highest ROC-AUC)
- All models showed **high accuracy (≥ 0.96)**
- **Future Scope:**
  - Use of larger & more diverse datasets
  - Integration of multimodal data (genetic, imaging, clinical)
  - Clinical validation for **real-world deployment**

References

- Pang T., Wong J.H.D., Ng W.L., Chan C.S. Deep learning radiomics in breast cancer with different modalities: Overview and future. *Expert Syst. Appl.* 2020 ;158:113501.
- Zhou X., Li C., Rahaman M.M., Yao Y., Ai S., Sun C., Wang Q., Zhang Y., Li M., Li X., et al. A comprehensive review for breast histopathology image analysis using classical and deep neural networks. *IEEE Access.* 2020;8:9093190956.
- Burstein, Harold J., Kornelia Polyak, Julia S. Wong, Susan C. Lester, and Carolyn M. Kaelin. "Ductal carcinoma in situ of the breast." New England Journal of Medicine 350, no. 14 (2004): 1430-1441.
- Mahmood T., Li J., Pei Y., Akhtar F., Imran A., Rehman K.U. A Brief Survey on Breast Cancer Diagnostic with Deep Learning Schemes Using Multi-Image Modalities. *IEEE Access.* 2020;8:165779–165809.
- Billah, Mustain, and Sajjad Waheed. "Minimum redundancy maximum relevance (mRMR) based feature selection from endoscopic images for automatic gastrointestinal polyp detection." Multimedia Tools and Applications 79, no. 33 (2020): 23633-23643.
- Chen, Dehua, Guangjun Qian, Cheng Shi, and Qiao Pan. "Breast cancer malignancy prediction using incremental combination of multiple recurrent neural networks." In Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24, pp. 43-52. Springer International Publishing, 2017.

Thank You!