A Computational Study on Classification of Benign and Malignant Tissue

A predictive to prescriptive approach

Siju K.S[1] Dr. Vipin.V[2]

[1]Roll No. CB.AI.R4CEN24003
Amrita School of Artificial Intelligence

[2]Thesis Supervisor
Amrita School of Artificial Intelligence

November 2024

Introduction

- **Objective:** Enhance breast cancer diagnosis using statistical and machine learning models.
- **Dataset:** UCI Breast Cancer Wisconsin Dataset (569 instances, 32 attributes) [5].
- **Approach:**
  - Supervised models: Logistic Regression, SVM, Decision Tree [2, 3].
  - Unsupervised model: K-means clustering.
- **Outcomes:**
  - Predictive accuracy for diagnosis.
  - Prescriptive insights for medical procedures.

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT, 1956

Methodology

- **Data Preprocessing:**
  - IQR-based outlier detection and square root transformation.
  - Feature scaling and normalization.
- **Feature Selection:**
  - Key features: Concave Points Worst, Area Worst, Perimeter Worst.
  - Correlation coefficient > 0.75 with malignancy.
- **Algorithms Applied [1]:**
  - Logistic Regression (Baseline).
  - SVM (Linear & Non-linear Kernels)[4].
  - Decision Trees.
  - K-means Clustering.

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT, 1956

Key Results-1

Performance Metrics (Accuracy):

- Logistic Regression: 94.7%.
- SVM (RBF Kernel): 97.2%.
- Decision Tree: 94.5%.

Comparison Table:

Model	Accuracy	Sensitivity	Specificity	F1-Score
Logistic Regression	94.7%	92.5%	96.3%	94.4%
SVM (RBF Kernel)	97.2%	95.8%	98.0%	97.0%
Decision Tree	94.5%	91.8%	95.9%	94.2%

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT, 1956

Key Results (Clustering)

Table: Distribution of Benign and Malignant cases over clusters.

Cluster	Diagnosis				Total
	Benign		Malignant		
	Count	%	Count	%	
C1	355	83.53	70	16.47	425
C2	2	1.63	121	98.37	123
C3	0	0.00	21	100.0	21
Total cleaned samples					569

Model Insights

- **Explainability:**
  - SVM: High accuracy but less interpretable.
  - Logistic Regression & Decision Tree: Balanced accuracy and interpretability.
- **Clustering:**
  - Revealed misclassified cases.
  - Supports additional diagnosis.
- **Prescriptive Model:**
  - Guides clinicians on severity and subsequent procedures.

Conclusion and Future Work

- **Conclusion:**
  - Effective use of machine learning enhances diagnostic accuracy.
  - Prescriptive models assist in comprehensive clinical decision-making.
- **Future Directions:**
  - Integration of additional datasets for validation.
  - Development of real-time diagnostic tools.

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT, 1956

References I


Stephen Boyd and Lieven Vandenberghe.
Convex optimization.
Cambridge university press, 2004.


Arturo Gago, Jean Marko Aguirre, and Lenis Wong.
Machine learning system for the effective diagnosis and survival prediction of breast cancer patients.
International Journal of Online & Biomedical Engineering, 20(2), 2024.


Zakia Salod and Yashik Singh.
A five-year (2015 to 2019) analysis of studies focused on breast cancer prediction using machine learning: A systematic review and bibliometric analysis.
Journal of Public Health Research, 9(1):jphr.2020.1772, 2020.
PMID: 32642458.

References II

K.P. Soman, R. LOGANATHAN, and V. AJAY.
Machine Learning with SVM and Other Kernel Methods.
PHI Learning, 2009.

WIlliam Wolberg.
Breast Cancer Wisconsin (Original).
UCI Machine Learning Repository, 1990.
DOI: https://doi.org/10.24432/C5HP4Z.