

Breast Cancer Diagnosis Analysis: Employing MATLAB Pre-trained Models with Machine and Deep Learning

POSTER CODE
isqua2024.33f05b1

Hao Mai Xuan[1], Duong Cao Thi Thuy[1], Yankuba B. Manga[1,2,*]

[1]The Master Program in Smart Healthcare Management (SHM), International College of Sustainability Innovations, National Taipei University, New Taipei City, Taiwan

[2]The Bachelor Program in Smart Sustainable Development and Management (SSDM), International College of Sustainability Innovations, National Taipei University, New Taipei City, Taiwan

INTRODUCTION

Breast cancer is one of the major global health burdens and a leading cause of death in the female population [1]. We acknowledge the utmost importance of early screening, detection, diagnosis, and treatment of breast cancer [2].

Furthermore, despite various challenges, Artificial intelligence remains and will continue to be an effective solution that brings specific benefits [3].

AIM

This study aimed to
enhance breast cancer diagnosis by integrating feature and histopathological imaging data through advanced computational methods

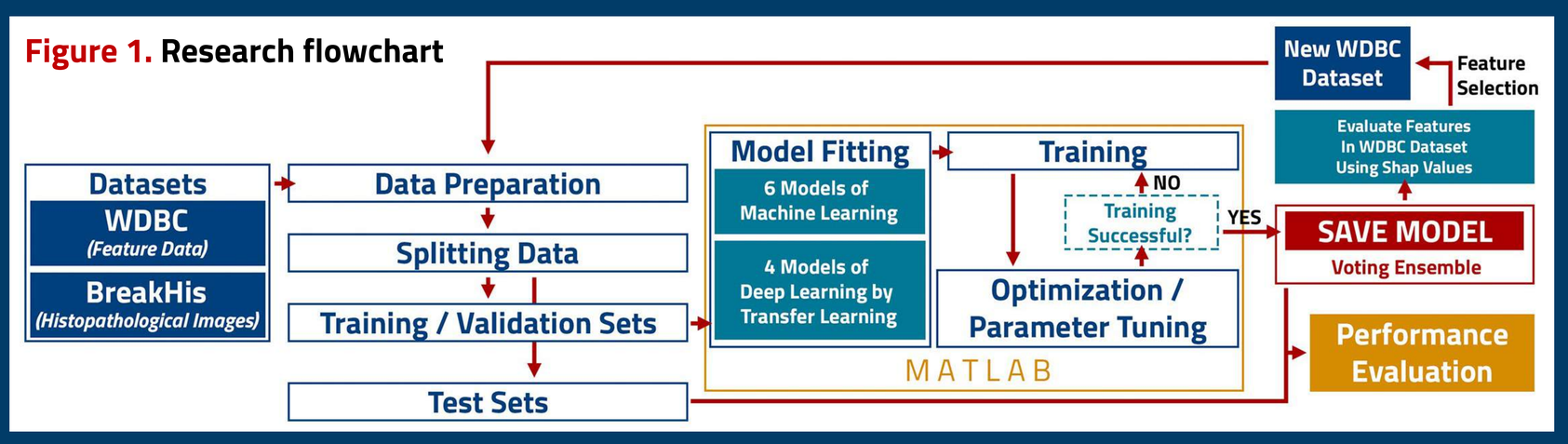
Keywords: Breast Cancer, Early Diagnosis, Data Mining, Computational Models, Machine and Deep Learning, Convolutional Neural Network

METHOD

This study compared six machine learning algorithms (Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Gaussian Naïve Bayes, and Linear Discriminant Analysis) using the Wisconsin Diagnostic Breast Cancer dataset (WDBC). Four convolutional neural network architectures (SqueezeNet, GoogLeNet, ResNet-50, and DenseNet-121) were applied to the BreakHis histopathological imaging dataset.

Transfer learning techniques, oversampling and SHAP value were utilized. Evaluation metrics included AUC, accuracy, precision, sensitivity, specificity, negative predictive value, and F1-score. MATLAB R2024a (The MathWorks, Inc.) was used for analysis (Figure 1).

Figure 1. Research flowchart



RESULT

SVM exhibited exceptional classification performance, achieving an accuracy of 0.972 across both datasets (Table 1). The 6 models performed similar results, demonstrating the evaluation of feature impact on class using SHAP values (Figure 2). Significantly, ResNet-50 and DenseNet-121 consistently maintain high accuracy and other performance metrics, making them promising candidates for accurate and reliable image classification tasks in the medical domain.

Figure 2. WDBC dataset SHAP values by 6 models

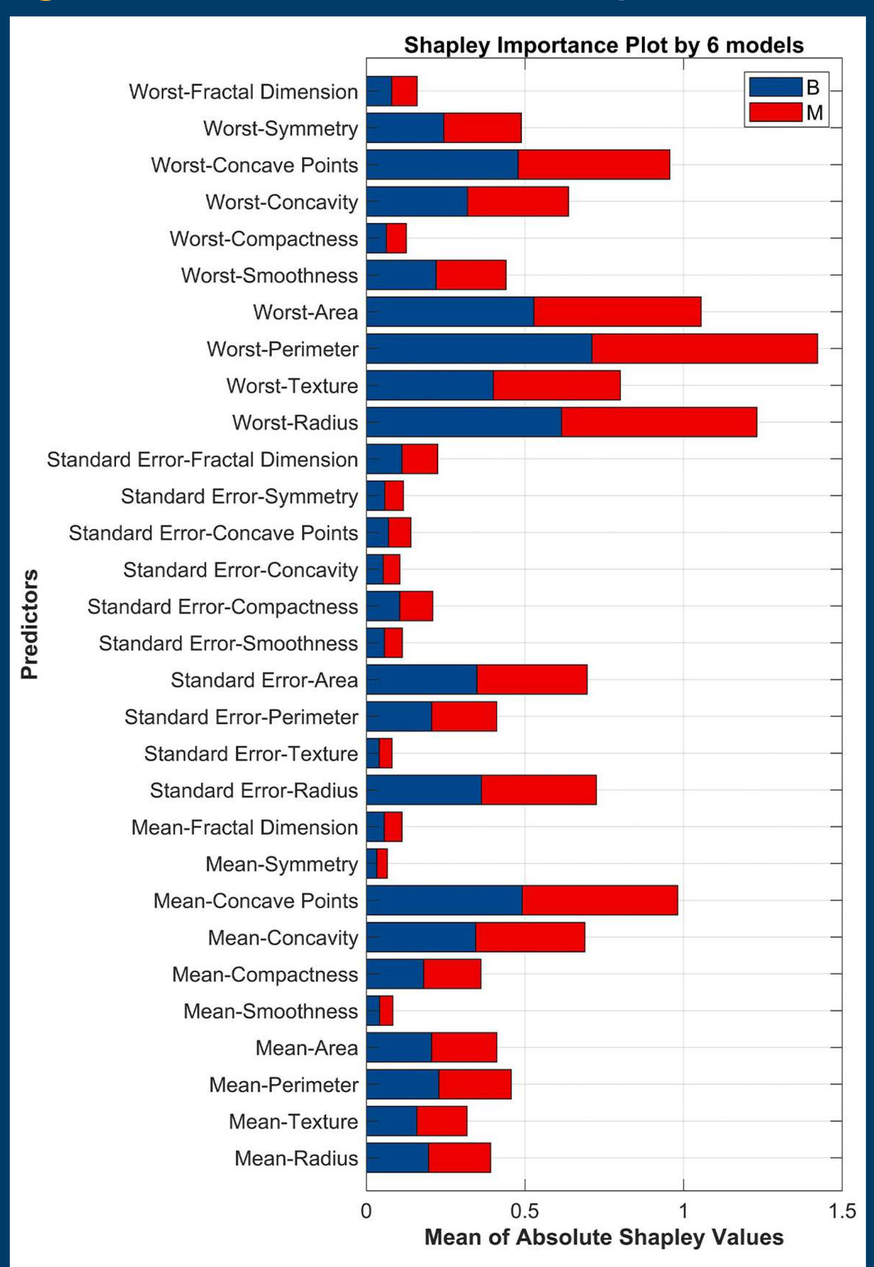
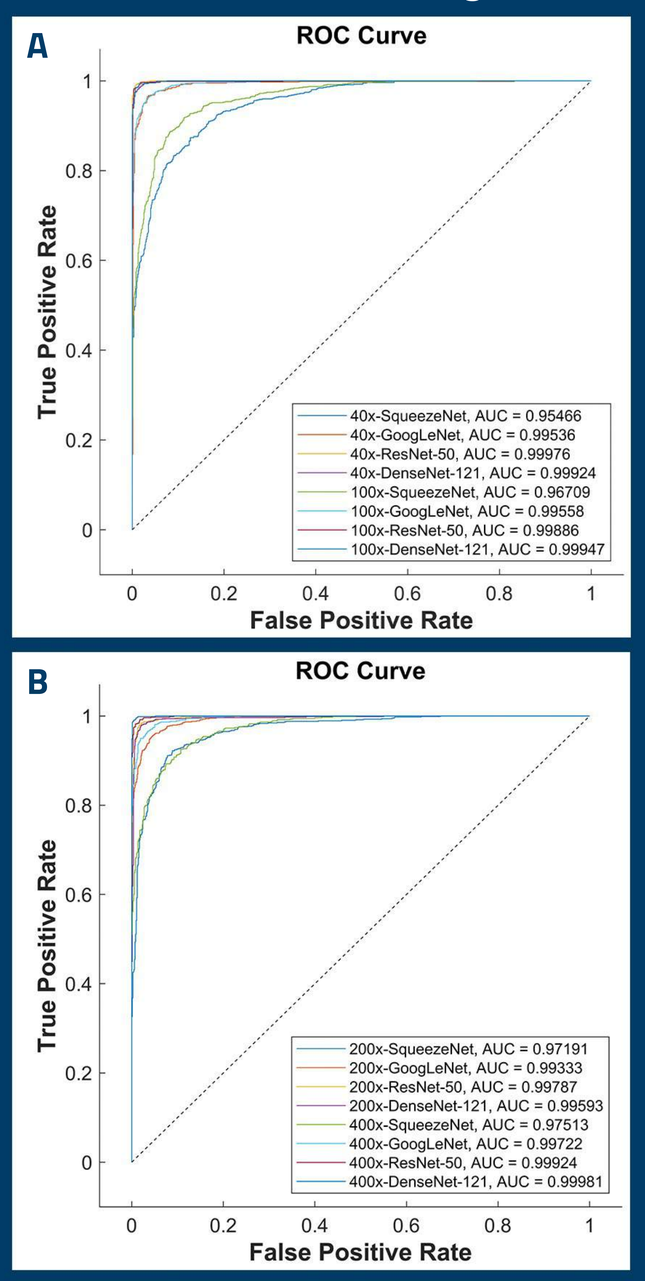


Table 1. Models performance analysis

Dataset	Models	Accuracy	Precision	Sensitivity	Specificity	NPV	F1-score
WDBC (30 Features)	LR	0.971	0.984	0.958	0.984	0.959	0.971
	SVM	0.972	0.986	0.959	0.986	0.960	0.972
	DT	0.941	0.954	0.927	0.955	0.928	0.940
	RF	0.970	0.975	0.965	0.975	0.965	0.970
	GNB	0.940	0.946	0.934	0.945	0.934	0.940
	LDA	0.964	0.985	0.942	0.986	0.943	0.963
WDBC (15 Features - SHAP Value)	LR	0.968	0.976	0.960	0.976	0.960	0.968
	SVM	0.972	0.987	0.958	0.987	0.959	0.972
	DT	0.942	0.946	0.938	0.945	0.937	0.942
	RF	0.965	0.966	0.965	0.965	0.964	0.965
	GNB	0.957	0.957	0.958	0.957	0.958	0.958
	LDA	0.963	0.986	0.939	0.987	0.941	0.962
BreakHis – 40x	SQ	0.873	0.826	0.946	0.800	0.937	0.882
	GG	0.965	0.940	0.993	0.937	0.992	0.966
	R50	0.991	0.995	0.988	0.995	0.988	0.991
	D121	0.984	0.985	0.983	0.985	0.983	0.984
BreakHis – 100x	SQ	0.898	0.855	0.958	0.838	0.953	0.904
	GG	0.963	0.981	0.944	0.981	0.946	0.962
	R50	0.986	0.979	0.993	0.979	0.993	0.986
	D121	0.986	0.977	0.995	0.977	0.995	0.986
BreakHis – 200x	SQ	0.912	0.915	0.909	0.916	0.910	0.912
	GG	0.954	0.938	0.974	0.935	0.973	0.955
	R50	0.980	0.990	0.969	0.990	0.969	0.979
	D121	0.978	0.983	0.974	0.983	0.974	0.978
BreakHis – 400x	SQ	0.908	0.872	0.957	0.859	0.952	0.912
	GG	0.965	0.997	0.932	0.997	0.936	0.964
	R50	0.984	0.986	0.981	0.986	0.981	0.984
	D121	0.992	0.992	0.992	0.992	0.992	0.992

LR: Logistic Regression, SVM: Support Vector Machin, DT: Decision Tree, RF: Random Forest, GNB: Gaussian Naïve Bayes, LDA: Linear Discriminant Analysis, NPV: Negative Prediction Value; SQ: SqueezeNet; GG: GoogLeNet; R50: ResNet-50; D121: DenseNet-121

Figure 3. AUC models classification on BreakHis dataset
a. 40x and 100x magnification
b. 200x and 400x magnification



CONCLUSION

This study highlights the effectiveness of AI models in MATLAB for diagnosing breast cancer. SVM shows promise for classification, while ResNet-50 show potential among CNNs. The findings suggest that future data training and model optimization efforts could benefit from leveraging SHAP values to eliminate less impactful features, thereby reducing noise. The research affirms the role of artificial intelligence in healthcare and poses challenges requiring large, accurate, diverse datasets for practical applications.

REFERENCES

1. *Arnold, M., et al., Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast, 2022. 66: p. 15-23.*
2. *American Cancer Society. Understanding a Breast Cancer Diagnosis. 2021; Available from: https://www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html.*
3. *Alowais, S.A., et al., Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ, 2023. 23(1): p. 689.*

Corresponding Author
Yankuba B. Manga, PhD.
orcid.org/0000-0002-8750-3721
Email: bydrmanga@gm.ntpu.edu.tw
Phone: +886-2-86741111 Ext. 68908