Explainable Deep Convolutional Networks for Early-Stage Breast Cancer Detection

Anika Zaheen[1], Dr. Md. Mizanur Rahman[2], Md Mokhlesur Rahman[3], Dr. Farjana Sultana Nur[4] ,Anisha Tasneem[5], Aadeel Albab[6]

[1] Department of Biomedical Engineering, Bangladesh University of Engineering & Technology· [2]Dhaka Medical College & Hospital ·
[3]Department of Medical Physics & Biomedical Engineering, Gono Biswabidyalay, Savar, [4]Kurmitola Multicare Specialized Hospital ·
[5] Kumuduni Women's Medical College, Tangail · [6]Mirpur Cantonment Public School & College ,Dhaka, Bangladesh

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Early, accurate diagnosis is decisive for breast cancer survival. Machine learning can match expert accuracy, but the opacity of high-performing models erodes the clinical trust needed for adoption. We developed an interpretable deep neural network that couples strong predictive performance with transparent, auditable reasoning. A fully-connected deep network (30→64→32→16→1 with Batch-Normalization, Dropout and L2 weight decay) was trained on the Breast Cancer Wisconsin (Diagnostic) dataset (569 cases, 30 nuclear features) under a strictly leakage-controlled 80/20 split, using Adam and early stopping on the validation ROC-AUC. Eleven classical learners were tuned under the identical pipeline as baselines, and SHAP and LIME (cross-checked with permutation importance) provide global and local explanations.

On the held-out set the network reached 97.37% accuracy, 97.56% precision, 95.24% recall, 96.39% F1 and 99.60% ROC-AUC (AP = 0.994) with only 1 false positive and 2 false negatives; agreement was high (MCC = 0.943) and the model well-calibrated (Brier = 0.020). A Youden-optimal threshold removed the false positive to reach 98.25% accuracy at 100% precision, and 2,000-sample bootstrap intervals confirmed stability. A McNemar test found no significant difference from the best classical ensemble (p = 1.00). SHAP attributions are dominated by concavity-, texture- and size-related features that mirror histopathological markers of malignancy. An explainable deep-learning framework delivers high diagnostic performance together with clinically coherent, transparent reasoning.

CONTACT

Anika Zaheen, M.Sc.
Department. of Biomedical Engineering,
Bangladesh University of Engineering & Technology (BUET), Dhaka
Email: zaheenanika2023@gmail.com

INTRODUCTION

Breast cancer is a leading cause of cancer death in women, and the stage at which a lesion is detected is the strongest determinant of prognosis: early, node-negative disease exceeds 90% five-year survival. Yet conventional diagnosis—mammography, ultrasound, and biopsy—is limited by inter-observer variability, subjective interpretation, and uneven access, increasing the risk of missed malignancies and unnecessary biopsies. Deep learning uses successive non-linear transformations to capture the nuclear morphology that encodes malignancy, but the depth that gives this flexibility also makes the model opaque—the main barrier to clinical adoption. We therefore pair a compact deep neural network with model-agnostic Explainable AI (SHAP, LIME), so every prediction is auditable. The goal is a framework that combines high accuracy with transparent, clinically legible reasoning for early, trustworthy diagnosis.

METHODS AND MATERIALS

Dataset & preprocessing: The Breast Cancer Wisconsin (Diagnostic) set, 569 fine-needle-aspirate cases, 30 morphometric nuclear features (357 benign / 212 malignant) was z-score standardized and split 80/20 with stratification; the scaler was fit on training data only, preventing leakage.

Network: A fully-connected deep network maps the 30 features through hidden layers of 64, 32 and 16 ReLU units (Batch-Normalization + Dropout) to a sigmoid output, trained with Adam, binary cross-entropy and L2 weight decay, with early stopping on the validation ROC-AUC.

Comparison & explanation: Eleven classical learners were tuned under the identical pipeline as baselines; SHAP (global) and LIME (local), cross-checked with permutation importance, expose the model's reasoning.

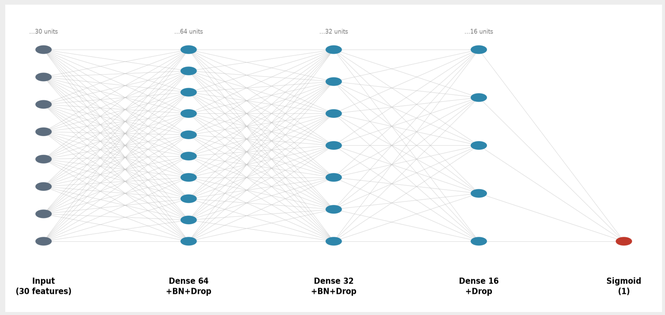

Fig 1. Deep neural network architecture (30 → 64 → 32 → 16 → 1; 4993

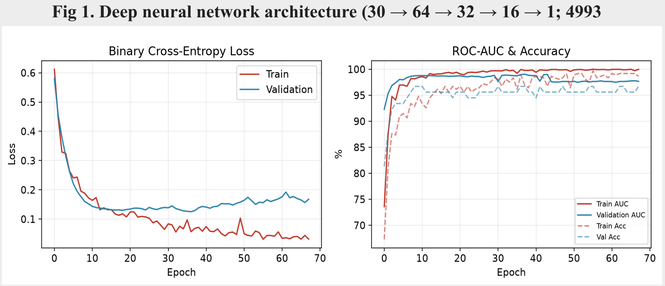

Fig 2. Training dynamics — loss and ROC-AUC with early stopping

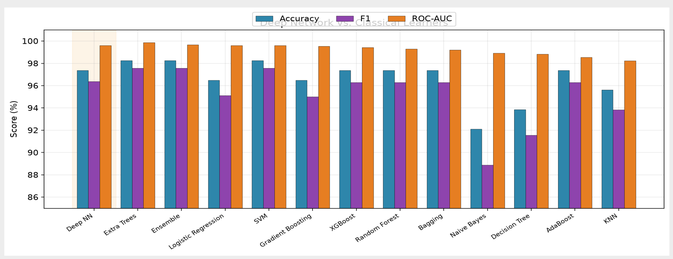

Fig 3. Deep network vs. eleven classical learners (accuracy, F1, ROC-AUC)

RESULTS

On the held-out test set the deep network reached 97.37% accuracy, 97.56% precision, 95.24% recall, 96.39% F1 and 99.60% ROC-AUC (AP 0.994) , only 1 false positive and 2 false negatives among 114 cases. Chance-corrected agreement was high (MCC = 0.943, κ = 0.943) and the model was well-calibrated (Brier = 0.020), so its probabilities can be read as genuine malignancy likelihoods.

The confusion structure, ROC and precision–recall curves (Figs 4–5, 9) confirm strong discrimination under the benign-majority distribution, while the reliability diagram and probability histogram (Figs 6, 10) show well-separated confidence scores.

Model	Acc	Prec	Rec	F1	AUC
Deep Neural Network	**97.37**	**97.56**	**95.24**	**96.39**	**99.60**
Soft-Voting (LR+SVM)	98.25	100.0	95.24	97.56	99.67
Extra Trees	98.25	100.0	95.24	97.56	99.87
SVM	98.25	100.0	95.24	97.56	99.60
Logistic Regression	96.49	97.50	92.86	95.12	99.60

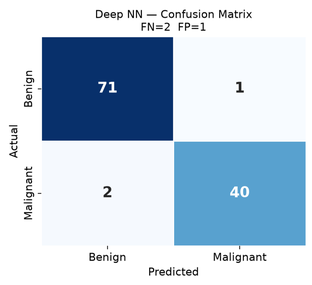

Fig 4. Confusion matrix (DNN)

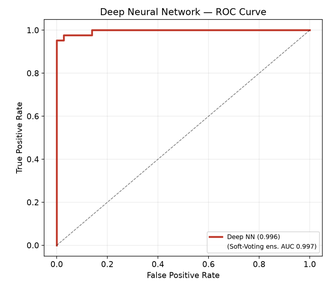

Fig 5. ROC curve (DNN)

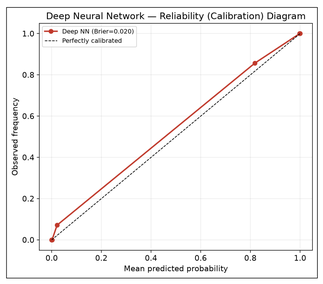

Fig 6. Reliability (calibration)

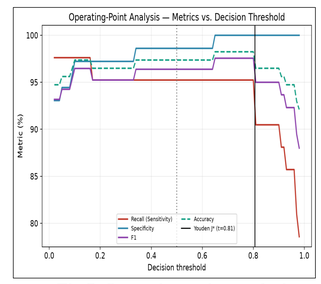

Fig 7. Operating-point analysis

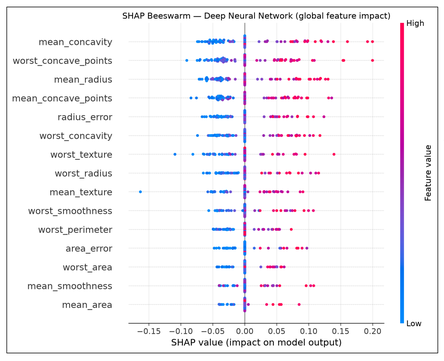

Fig 8. SHAP beeswarm — global feature impact on the deep network

Because the network ranks cases almost perfectly, a Youden-optimal threshold (t = 0.81) removes the false positive to reach 98.25% accuracy at 100% precision

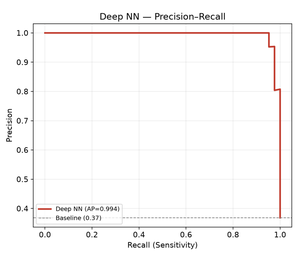

Fig 9. Precision–recall (AP = 0.994)

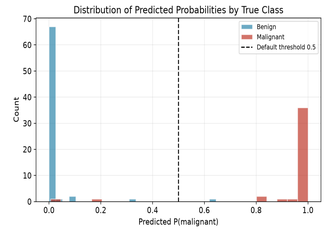

Fig 10. Predicted-probability distribution by class

DISCUSSION

The deep network attained excellent discrimination (ROC-AUC ≈ 0.996, AP = 0.994) while remaining fully interpretable, demonstrating that the opacity usually attributed to neural models is not inevitable when SHAP and LIME are integrated from the outset. SHAP and LIME identified concavity-, texture- and size-related descriptors as the dominant predictors (Figs 8, 11–13), the same morphometric signs a pathologist inspects when grading a lesion. This concordance between a data-driven attribution and domain knowledge is strong evidence that the network learned clinically meaningful structure rather than dataset-specific noise. The network is statistically indistinguishable from a tuned classical ensemble on this benchmark; an ablation shows regularization chiefly stabilizes training (removing Dropout doubled run-to-run variance), and a learning curve shows performance saturates after a few hundred samples (Figs 14–15).

Local & global interpretability

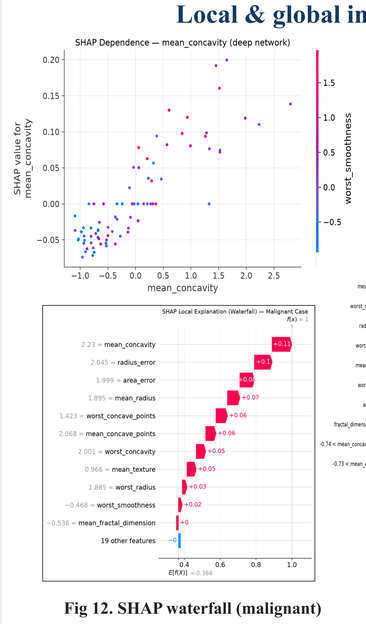

Fig 11. SHAP dependence — top feature (smooth monotone effect).

Fig 13. LIME (benign case)

Fig 12. SHAP waterfall (malignant)

Robustness & data efficiency

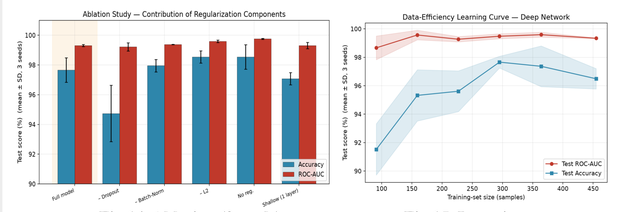

Fig 14. Ablation (3 seeds)

Fig 15. Learning curve

CONCLUSIONS

An explainable deep-learning framework delivers high performance (97.37% accuracy, 99.60% ROC-AUC) with transparent, clinically coherent reasoning; its attributions (Figs 8, 11–13) align with established histopathology, and bootstrap CIs confirm stability. Regularization stabilizes training and performance saturates on tabular features (Figs 14–15), so the network's headroom lies in richer inputs (raw imaging, multi-omics). Future: external multi-center validation, transfer learning on histopathology images, cost-sensitive thresholds, and prospective clinician-in-the-loop trials.

REFERENCES

[1] Lundberg & Lee. A unified approach to interpreting model predictions. NeurIPS, 2017.
[2] Wolberg et al. Breast Cancer Wisconsin (Diagnostic). UCI ML Repository, 1993.
[3] Kingma & Ba. Adam: A method for stochastic optimization. ICLR, 2015.
[4] Ioffe & Szegedy. Batch normalization. ICML, 2015.
[5] Srivastava et al. Dropout. JMLR, 2014.
[6] Ghasemi et al. Explainable AI in breast cancer detection. Cancer Innov., 2024.