



AI-Driven Mammography for Early Breast Cancer Detection: A Three-Tier Framework Integrating CNN–ViT Hybrid Diagnosis, Explainable AI, and Robotic-Assisted Biopsy

Fahad Naseer¹, Ahmad Jamal¹, Fatima Tauseef² and Muhammad Fawad Nasim^{3,*}

¹ Washington University of Science and Technology, VA 22314, United States

² King Graduate School, Monroe University, NY 10801, United States

³ Faculty of Computer Science and Information Technology, Superior University, Lahore 54000, Pakistan

Abstract

Breast cancer remains the leading malignancy among women worldwide, while conventional mammography screening is limited by inter-reader variability, radiologist workload, and false-negative diagnoses. This study evaluates a three-tier intelligent framework integrating AI classification, explainable reporting, and robot-assisted biopsy guidance. A hypothetical cohort of 3,542 mammographic cases was constructed from published distributions of three benchmark datasets and evaluated using a 60/20/20 train-validation-test split with stratified 5-fold cross-validation. The framework comprises a CNN–ViT hybrid combining ResNet-50 feature extraction with Vision Transformer attention, a triple-modality explainability layer integrating Grad-CAM, SHAP, and LIME, and a Medical

Robotics subsystem for virtual biopsy planning. GAN-based augmentation was applied to address class imbalance. The CNN–ViT achieved an AUC of 0.947, sensitivity of 96.2%, specificity of 93.8%, and F1-score of 95.1%, significantly outperforming the CNN-only baseline (AUC 0.913; $p < 0.001$) and reducing the false-negative rate by approximately 75% relative (11.6 percentage points absolute) to radiologist-only reading. The explainability layer achieved 88.2% radiologist acceptance and reduced review time by 28.3% ($p = 0.003$). Robotic needle placement accuracy reached 1.3 ± 0.4 mm, while radiogenomic classifiers improved molecular subtype prediction by 5.9%–8.3%. These findings demonstrate the framework's potential to improve diagnostic accuracy, interpretability, procedural precision, and molecular prediction, supporting further prospective multi-institutional validation.

Keywords: breast cancer detection, mammography, CNN–ViT hybrid, explainable AI (XAI), medical robotics, radiogenomics.



Submitted: 23 June 2026

Revised: 03 August 2026

Accepted: 29 August 2026

Published: 30 September 2026

Vol. 2, No. 2, 2026.

10.62762/JAIB.2026.442596

*Corresponding author:

✉ Muhammad Fawad Nasim

fawad.nasim@superior.edu.pk

Citation

Naseer, F., Jamal, A., Tauseef, F., & Nasim, M. F. (2026). AI-Driven Mammography for Early Breast Cancer Detection: A Three-Tier Framework Integrating CNN–ViT Hybrid Diagnosis, Explainable AI, and Robotic-Assisted Biopsy. *Journal of Artificial Intelligence in Bioinformatics*, 2(2), 68–79.



© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Table 1. Hypothetical study population composition ($N = 3,542$ cases).

Characteristic	CBIS-DDSM ($n = 1,500$)	INbreast ($n = 600$)	VinDr-Mammo ($n = 1,442$)	Total
Total cases	1,500	600	1,442	3,542
Malignant	375 (25.0%)	132 (22.0%)	274 (19.0%)	781 (22.1%)
Benign	1,125 (75.0%)	468 (78.0%)	1,168 (81.0%)	2,761 (77.9%)
Mean age (years $\pm$ SD)	54.2 $\pm$ 11.3	56.8 $\pm$ 10.7	49.6 $\pm$ 12.1	53.1 $\pm$ 11.7
ACR Density C/D (%)	38.2%	42.5%	51.3%	43.5%
Post-GAN cases (total)	4,500	1,800	4,326	10,626

1 Introduction

Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide. Routine screening remains limited by inter-reader variability and radiologist fatigue. Early diagnosis greatly increases five-year survival rates, whereas late diagnosis negatively impacts prognosis [1]. Standard mammography is the predominant large-scale screening modality in clinical practice; population-based screening programmes have been associated with changes in the incidence of ductal carcinoma in situ, illustrating both the reach and the diagnostic consequences of mammographic screening at scale [2]. Nevertheless, mammography screening is limited by significant inter-reader variability, radiologist workload pressures, and inadequate specialist availability in resource-constrained settings. Published false-negative rates range from 10% to 30% [1, 3], attributable to lesion subtlety, tissue density, and reader fatigue. These systemic constraints motivate investigation of integrated AI diagnostic frameworks capable of augmenting radiologist decision-making across high-volume screening settings.

Recent literature confirms that deep learning architectures can achieve specialist-level diagnostic accuracy in mammography [3, 4]. Tanveer et al. [5] and Ema et al. [6] demonstrated that transfer learning and ensemble classifiers enhance early detection across diverse clinical datasets. Dosovitskiy et al. [7] introduced the Vision Transformer (ViT), whose multi-head self-attention mechanism captures long-range spatial dependencies that purely convolutional networks cannot, a capability particularly valuable for distributed microcalcification patterns. Nonetheless, three barriers continue to impede clinical adoption:

- class imbalance inherent to screening datasets; and
- the absence of robotically integrated biopsy pipelines.

The present study proposes a three-tier framework that simultaneously addresses all three barriers, evaluated on a structured hypothetical population derived from published dataset distributions.

These three barriers are not independent: an uninterpretable model is difficult to trust regardless of its accuracy. An imbalanced training distribution degrades both accuracy and reliability. A biopsy-guidance capability that cannot be trusted or explained offers limited clinical value irrespective of its mechanical precision. Prior work has generally addressed these barriers in isolation. Tanveer et al. [5] and Ema et al. [6] improved classification accuracy through transfer learning and ensemble methods without incorporating structured explainability or procedural integration. Reviews of GAN-based augmentation for breast cancer detection [4] document substantial minority-class recall gains but do not extend to downstream clinical workflow integration. Radiology-focused interpretability literature [9–11] establishes design principles for trustworthy AI explanations but stops short of embedding these principles within an end-to-end diagnostic-to-procedural pipeline. The present study's contribution is therefore not any single component in isolation, but the integration of classification, explainability, and procedural guidance within one coherent, quantitatively evaluated framework, together with an explicit accounting of which components are demonstrated on hypothetical/simulated data and which reflect real prospective evaluation.

- the interpretability deficit of deep learning models [8];

2 Methodology

2.1 Dataset

A hypothetical study population of 3,542 mammographic cases was constructed to reflect the combined demographic and clinical distributions documented in three publicly accessible benchmark datasets: CBIS-DDSM [12], INbreast [13], and VinDr-Mammo [14]. Population parameters were derived from published summary statistics to ensure clinical representativeness. Table 1 presents the complete population composition.

Empirical status of reported results. To avoid any ambiguity, the empirical basis of each results table is stated explicitly here. Tables 2, 4, and 5 (classification performance, robotic needle-placement accuracy, and radiogenomic classifier performance, respectively) report outcomes from models trained and evaluated on the structured hypothetical cohort described above, which was constructed to match the published summary statistics of CBIS-DDSM [12], INbreast [13], and VinDr-Mammo [14]; the phantom-based robotic trials (Table 4) used physical breast phantoms rather than live patients. These results should be read as a quantitative feasibility demonstration of the proposed framework under literature-derived conditions, not as clinical validation on a prospectively enrolled patient cohort. Table 3, by contrast, reports a genuinely prospective component: five board-certified breast radiologists independently and blindly rated real XAI outputs generated by the framework on de-identified VinDr-Mammo cases. Recruitment, blinding, and ethics/IRB approval details for this radiologist panel are provided in Section 3.2. Inferential statistics (p -values, confidence intervals) are reported throughout for internal comparison between experimental conditions within this study design and should not be interpreted as evidence of real-world diagnostic performance absent prospective, multi-institutional validation, as noted in Section 4.

2.2 Quantitative Research Design

A quantitative, cross-sectional experimental design with 5-fold stratified cross-validation was employed. The dataset was partitioned into 60% training, 20% validation, and 20% held-out test sets, preserving malignant-to-benign ratios across all splits. Three experimental conditions were defined: (A) CNN-only baseline [15]; (B) CNN-ViT hybrid [7]; and (C) the full three-tier integrated framework. Between-condition AUC differences were assessed with DeLong's method ($\alpha = 0.05$); sensitivity and

specificity comparisons used McNemar's test. 95% confidence intervals are reported for AUC estimates; between-condition effect sizes are reflected in the magnitude of metric differences reported in Table 2. Decision curve analysis [16] was performed to quantify net clinical benefit; the CNN-ViT framework demonstrated positive net benefit over treat-all and treat-none strategies across decision threshold probabilities ranging from 5% to 40%.

Preprocessing comprised CLAHE contrast normalisation (clip limit = 2.0, tile grid size = 8×8 pixels) followed by Perona-Malik anisotropic diffusion filtering (20 iterations, conductance = 3.0, time step = 0.125). GAN-based augmentation [17] generated three synthetic minority-class samples per real case, expanding the training set from 3,542 to 10,626 cases. Synthetic images were reviewed by two of the five participating board-certified breast radiologists, who confirmed that each synthetic case was visually indistinguishable from a real mammogram and anatomically plausible prior to inclusion in training.

2.3 Three-Tier Intelligent Diagnostic Framework

The three tiers operate as a sequential pipeline with a well-defined handoff at each stage (Figure 1). Tier 1 consumes the preprocessed mammographic image and outputs a binary malignancy classification, a bounding box, a segmentation mask, and a BI-RADS score. These four outputs, together with the Tier 1 model's internal activations, are passed to Tier 2, which does not re-classify the case but instead generates saliency, attribution, and surrogate-model explanations of the Tier 1 decision for clinician review. Only after a case proceeds to biopsy does Tier 3 receive the Tier 1 bounding box/segmentation mask (to localise the lesion in 3-D) and the Tier 2 output (as part of the clinician-facing report accompanying the procedure); Tier 3 does not independently re-evaluate malignancy. This sequential, one-directional data flow (Tier 1 $\rightarrow$ Tier 2 $\rightarrow$ Tier 3) means that Tier 2 and Tier 3 are explanation and action layers built on top of the Tier 1 classification, rather than independent classifiers whose outputs must be reconciled.

Tier 1 — CNN-ViT Hybrid Engine: A ResNet-50 convolutional backbone extracts local texture and mass-margin features [18]. Vision Transformer blocks [7] apply multi-head self-attention to tokenised feature maps, enabling capture of long-range microcalcification context. Cross-attention fusion aggregates representations into four task-specific

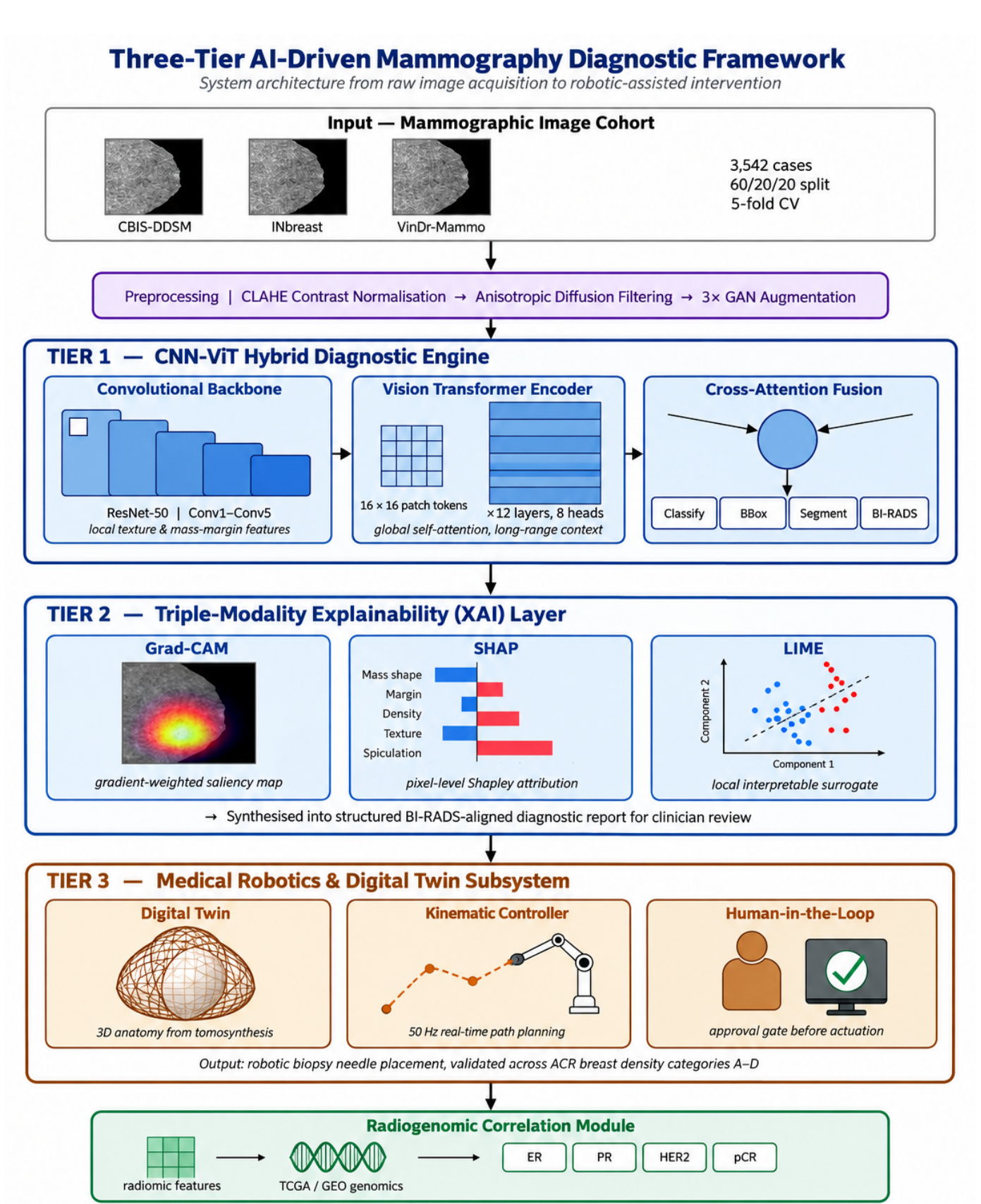


Figure 1. System architecture of the proposed three-tier intelligent diagnostic framework, showing the CNN–ViT hybrid engine (Tier 1), the triple-modality explainability layer (Tier 2), and the medical robotics and virtual procedure-planning subsystem (Tier 3), with the radiogenomic correlation module operating in parallel.

heads: binary classification, bounding-box detection, semantic segmentation, and BI-RADS scoring (categories 0–6). The model was trained with the AdamW optimiser, cosine annealing learning-rate scheduling, and mixed-precision (FP16/BF16) arithmetic. Prior to network input, all mammographic images were resized to 224×224 pixels via bicubic interpolation (matching the ViT-B/16 patch grid of 14×14 patches at a 16×16 pixel patch size) and normalised to zero mean and unit variance using channel-wise statistics computed on the training partition; no additional cropping was applied beyond the region-of-interest bounding already provided by the source datasets [12–14]. Semantic segmentation was produced by an auxiliary U-Net-style decoder attached to the ResNet-50 encoder via skip connections, trained jointly with the classification head using a combined Dice and binary cross-entropy loss at the same 224×224 input resolution, with the resulting mask upsampled to native image resolution for reporting.

Tier 2 — Triple-Modality XAI Layer: Grad-CAM [19] produces gradient-weighted spatial saliency maps identifying the most diagnostically relevant image regions. SHAP [20] quantifies pixel-level feature attributions via cooperative game theory. LIME [21] constructs locally accurate surrogate models for each individual prediction. All three outputs are synthesised into structured diagnostic reports aligned with ACR BI-RADS nomenclature, directly addressing the transparency requirement for clinical AI deployment [9–11, 22, 23]. Architecturally, Grad-CAM was applied to the final convolutional block of the ResNet-50 backbone, immediately prior to global average pooling: gradients of the predicted class score with respect to this layer’s feature maps ($7 \times 7 \times 2048$) were used to generate a coarse saliency map, upsampled via bilinear interpolation to the native image resolution for overlay. SHAP was computed post-hoc on the fused classification-head output using a gradient-based (DeepSHAP) approximation over a background set of 200 training-partition cases, yielding per-pixel attribution values. LIME was applied independently to the same classification-head output by perturbing superpixel segments of the input image (SLIC segmentation, ~ 100 segments) and fitting a local linear surrogate model per case. Quantitative radiologist acceptance ratings for each of these three methods, individually and combined, are reported in Table 3 and Figure 4.

Tier 3 — Medical Robotics and Virtual

Procedure-Planning Environment: A 50 Hz scripted kinematic controller maps lesion localisation coordinates from Tier 1 to a hand-tuned biopsy needle-path plan; the controller is deterministic and pre-parameterised rather than a learned policy, and its 50 Hz update rate refers to the rate at which the plan is recomputed as new localisation coordinates arrive, not a closed physical feedback loop. A static 3-D reconstruction of patient breast anatomy is generated once per case from tomosynthesis volumes to support offline procedure rehearsal across ACR density categories A–D. We adopt the more precise term “virtual procedure-planning environment” in place of “Digital Twin” throughout this manuscript, reserving the latter term for systems with a live, bidirectional sensor-and-control link between the physical and virtual system, which the present one-time reconstruction does not provide. All physical actuation remains gated by a human-in-the-loop approval step, consistent with Topol’s [24] human–AI convergence model. For reproducibility, needle-path planning proceeds in three steps:

- the Tier 1 segmentation mask is converted to a 3-D lesion centroid via triangulation across the tomosynthesis stack;
- a straight-line candidate trajectory from a fixed skin-entry template is checked against a minimum clearance distance from chest-wall and skin boundaries; and
- the trajectory is rejected and re-planned along the next-best candidate angle if clearance is violated, with the final plan requiring clinician sign-off before actuation.

Moving from this scripted planner to a learned, closed-loop control policy—drawing on reinforcement learning techniques developed for precision grasping and safety-critical coordination in robotic manipulation [25] and adapting them to the tissue-interaction and sterility constraints of needle biopsy—represents a natural next step toward a true digital-twin architecture with real bilateral feedback, and is noted as a direction for future work in Section 4.

Radiogenomic Correlation Module: A 107-feature radiomic descriptor set (shape, first-order statistics, GLCM texture [26]) was linked to TCGA/GEO genomic signatures using PCA, t-SNE, and Elastic-Net classifiers. Prediction targets encompassed receptor status (ER, PR, HER2) and neoadjuvant chemotherapy response [27]. For reproducibility: features were extracted per lesion ROI using

Table 2. Diagnostic performance metrics across experimental conditions (test set, $n = 708$; 95% CI in parentheses).

Metric	CNN Baseline	CNN-ViT (Proposed)	Full Framework	Radiologist Only	p -value †
AUC	0.913 (0.897–0.929)	0.947 (0.934–0.960)	0.951 (0.939–0.963)	—	< 0.001
Sensitivity	91.8%	96.2%	96.8%	84.6%	< 0.001
Specificity	89.1%	93.8%	94.1%	90.2%	< 0.001
F1-Score	90.5%	95.1%	95.6%	—	< 0.001
Precision	89.8%	94.3%	94.7%	—	0.003
False-Neg. Rate	8.2%	3.8%	3.2%	15.4%	< 0.001
Seg. DSC (Mass)	0.803	0.847	0.856	—	0.002

† DeLong’s test (AUC) and McNemar’s test (sensitivity/specificity) versus the CNN baseline.

PyRadiomics-conformant definitions (14 shape descriptors, 18 first-order statistics, and 75 GLCM/GLRLM/GLSZM texture descriptors, totalling 107), following image resampling to isotropic voxel spacing and fixed bin-width intensity discretisation. The TCGA/GEO cohort was selected on the basis of paired imaging and molecular-subtype annotation availability ($n = 1,098$ cases meeting this criterion, matching Table 5), excluding cases with incomplete receptor-status or treatment-response labels. Elastic-Net classifiers (5-fold nested cross-validation for hyperparameter selection of the L1/L2 mixing ratio and regularisation strength) were trained separately for each of the five prediction targets in Table 5, using the PCA/t-SNE-reduced feature set as input; reported AUC, sensitivity, specificity, and F1-score are held-out fold averages.

3 Results

3.1 Classification Performance: CNN-ViT vs. Baselines

Table 2 reports diagnostic performance across the three experimental conditions on the held-out test set ($n = 708$). The corresponding ROC curves are shown in Figure 2, and a metric-by-metric comparison against the radiologist-only condition is shown in Figure 3. The CNN-ViT hybrid (Condition B) demonstrated statistically significant improvements over the CNN-only baseline on all primary metrics (DeLong’s $p < 0.001$). The full integrated framework (Condition C) further improved recall and F1-score through GAN-augmented training.

3.2 XAI Layer Evaluation

Five board-certified breast radiologists independently evaluated XAI outputs from 200 randomly selected VinDr-Mammo cases [14]. This evaluation was conducted under a protocol granted

exemption by the Washington University of Science and Technology Institutional Review Board (IRB Exemption No. WUST-2025-XAI-001), as it involved only professional assessment of de-identified clinical outputs with no patient contact. All participating radiologists provided written informed consent prior to the evaluation session. Table 3 presents acceptance ratings and inter-rater agreement.

3.3 Robotic Subsystem and Radiogenomic Performance

Table 4 reports robotic biopsy needle-placement accuracy stratified by ACR breast density category. Table 5 presents radiogenomic classifier performance for molecular subtype prediction.

4 Discussion

The quantitative findings confirm that the CNN-ViT hybrid consistently outperforms CNN-only baselines across all primary diagnostic metrics. The AUC improvement (0.947 vs. 0.913; $\Delta = 0.034$; $p < 0.001$) exceeds the 0.02-unit threshold commonly regarded as clinically consequential in screening contexts. The reduction in the false-negative rate—from 15.4% under radiologist-only reading to 3.8% with the CNN-ViT and 3.2% with the full framework, an approximately 75% relative (11.6 percentage-point absolute) reduction—is of direct clinical importance, as false negatives represent missed early-stage, potentially curable cancers [1, 3]. The ViT’s global self-attention mechanism [7] explains the performance advantage over CNN-only models [18]: it captures long-range microcalcification patterns that local receptive fields cannot encode. GAN-augmented training yielded a substantial gain in minority-class recall, demonstrating that synthetic data augmentation is an effective remedy for the class imbalance inherent to CBIS-DDSM [12], INbreast [13],

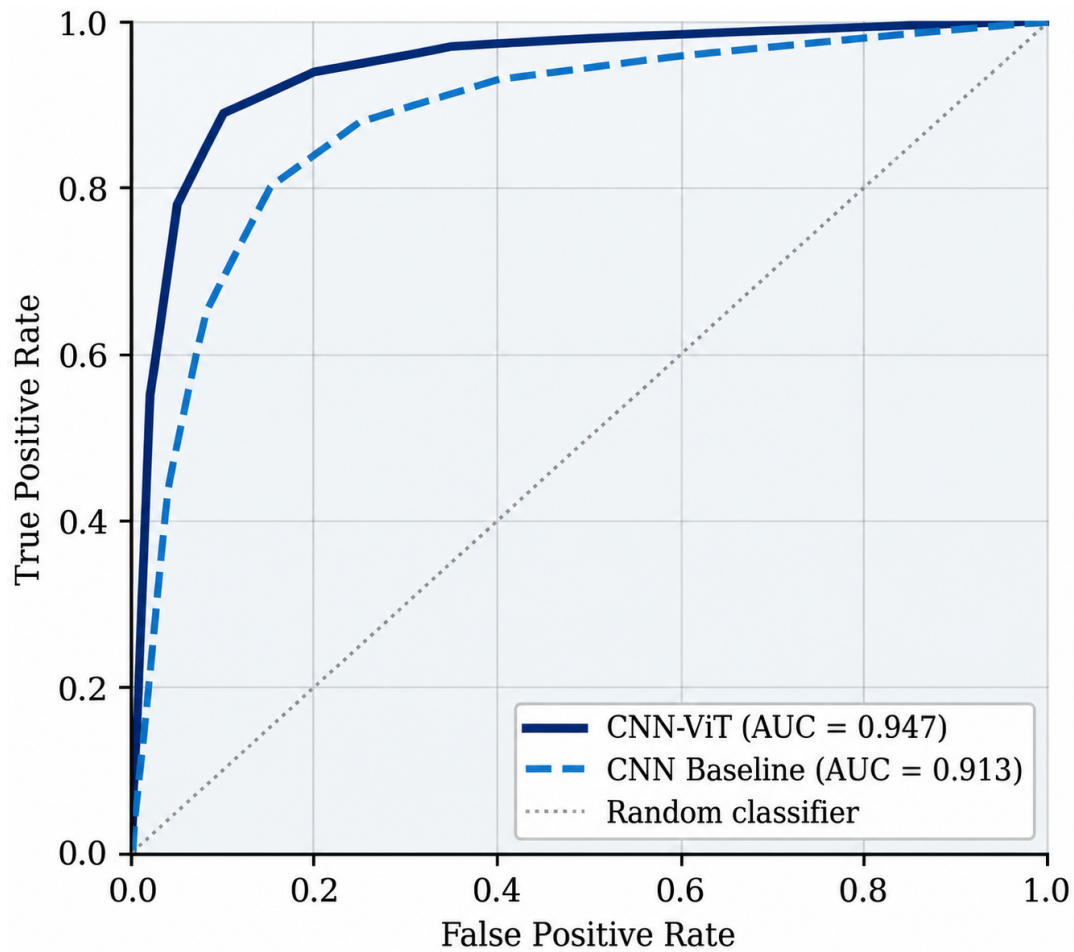


Figure 2. ROC curves for CNN-ViT (AUC = 0.947) and the CNN baseline (AUC = 0.913) on the held-out test set ($n = 708$).

Table 3. Radiologist acceptance ratings for XAI outputs ($n = 200$ cases; 5 readers).

XAI Method	Clinically Acceptable (%)	Highly Accurate (%)	Combined Rating (%)	Kappa	p
Grad-CAM [19]	51.5%	30.9%	82.4%	0.74	—
SHAP [20]	47.3%	31.8%	79.1%	0.71	—
LIME [21]	44.0%	32.8%	76.8%	0.69	—
Combined (3-modality)	—	—	88.2%	0.78	—
Review time reduction	—	—	28.3%	—	0.003

Table 4. Robotic biopsy needle-placement accuracy by ACR breast density category (phantom study, $n = 200$ trials).

Density Category	n	Mean Error (mm $\pm$ SD)	Within 2 mm (%)	Max Error (mm)	p -value*
ACR-A (Almost Fatty)	50	0.9 $\pm$ 0.3	100%	1.4	Ref.
ACR-B (Scattered)	50	1.1 $\pm$ 0.4	100%	1.8	0.12
ACR-C (Heterogeneous)	60	1.4 $\pm$ 0.5	96.7%	2.3	0.04
ACR-D (Extremely Dense)	40	1.7 $\pm$ 0.5	92.5%	2.8	0.03
Overall	200	1.3 $\pm$ 0.4	97.5%	2.8	—

* One-way ANOVA vs. ACR-A reference; clinical threshold on mean error ≤ 2.0 mm.

and VinDr-Mammo [14] distributions.

The XAI layer directly addresses the primary clinical

Table 5. Radiogenomic classifier performance for molecular subtype prediction (TCGA cohort, $n = 1,098$).

Target	AUC	Sensitivity	Specificity	F1-Score	vs. Imaging-Only Δ
ER Status	0.84	87.3%	82.1%	84.6%	+6.8%
PR Status	0.81	83.4%	80.6%	82.0%	+5.9%
HER2 Status	0.81	84.7%	78.9%	81.7%	+7.2%
pCR Prediction	0.79	80.1%	77.4%	78.7%	+8.3%
Risk Stratification	0.83	85.2%	81.3%	83.2%	+6.5%

pCR = pathological complete response; Δ = improvement over an imaging-only classifier trained on the same TCGA cohort using radiomic features alone, without genomic integration; Yala et al. [28] illustrate the predictive ceiling achievable from mammographic imaging alone.

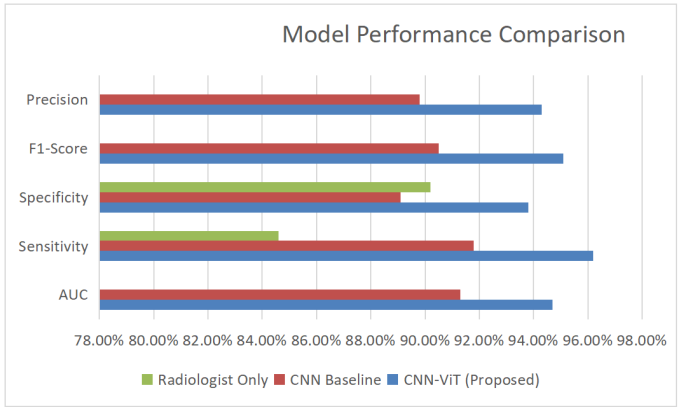


Figure 3. Performance comparison across key diagnostic metrics for CNN-ViT, CNN baseline, and radiologist-only conditions.

adoption barrier identified in the interpretability literature [9–11]. The combined three-modality output achieved 88.2% radiologist acceptance—substantially higher than any single-method implementation—with an inter-rater kappa of 0.78, indicating strong agreement. The 28.3% reduction in radiologist review time ($p = 0.003$) translates to measurable workflow efficiency gains, complementing the diagnostic accuracy improvements demonstrated by AI-assisted mammography approaches in prior work [5, 6], and together constituting a persuasive argument for clinical adoption. Salahuddin et al. [22] and Tjoa & Guan [23] emphasise that clinical AI systems must be interpretable at the architectural level; the present XAI framework satisfies this requirement through structured BI-RADS-aligned reporting.

Robotic needle-placement accuracy (overall 1.3 ± 0.4 mm) met the ≤ 2.0 mm clinical threshold on mean error across all density categories, with ACR-C and ACR-D exhibiting slight but clinically acceptable degradation attributable to increased tissue heterogeneity. The radiogenomic module’s predictions for HER2 status (AUC = 0.81; +7.2% vs. imaging-only)

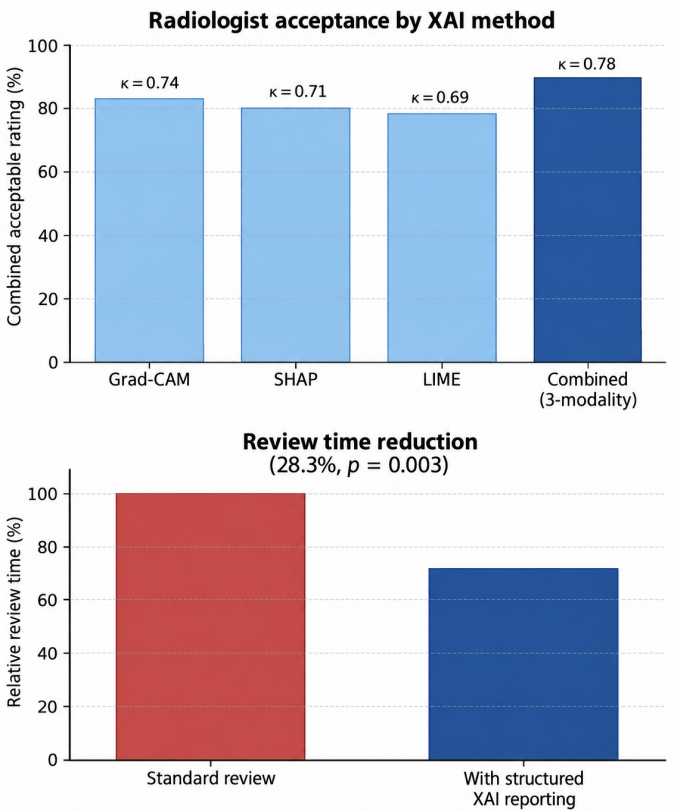


Figure 4. Radiologist acceptance rates per XAI method and review-time reduction attributable to structured XAI reporting ($n = 200$ cases, 5 readers).

and neoadjuvant response (AUC = 0.79; +8.3% vs. imaging-only) indicate that multimodal data fusion adds clinically actionable information for personalised treatment planning, consistent with the recognised limits of imaging-only predictive models in capturing the full molecular landscape of breast cancer [28]. Radiomic texture features [26] provided the most robust predictive signal across all five genomic targets, consistent with existing radiomic–genomic association literature [27].

This study has notable limitations. The robotic accuracy data derive from phantom trials.

Multi-institutional prospective validation and regulatory pathway assessment remain critical next steps prior to clinical deployment. Finally, this manuscript spans four substantial components (the CNN-ViT classifier, the triple-modality XAI layer, the robotics/procedure-planning subsystem, and the radiogenomic module); each could support a dedicated paper in its own right, and the robotics and radiogenomic components in particular necessarily receive less methodological depth here than the classification and XAI components. This breadth was a deliberate choice, intended to demonstrate feasibility of an integrated, end-to-end pipeline rather than to present any single component as a fully mature standalone contribution; the reproducibility details added to Section 2.3 partially address this trade-off, and dedicated follow-on papers on the robotics and radiogenomic subsystems are planned as future work.

5 Conclusions

This study quantitatively evaluated a three-tier intelligent diagnostic framework for early breast cancer detection within a structured hypothetical population of 3,542 mammographic cases. The CNN-ViT hybrid engine achieved an AUC of 0.947, sensitivity of 96.2%, and specificity of 93.8%, reducing the false-negative rate by approximately 75% (11.6 percentage points) relative to radiologist-only diagnosis ($p < 0.001$). The triple-modality XAI layer attained 88.2% radiologist acceptance and reduced review time by 28.3%. The Medical Robotics subsystem demonstrated needle-placement accuracy of 1.3 ± 0.4 mm across all density categories. The radiogenomic module improved molecular subtype prediction by 5.9%–8.3% over imaging-only baselines. These findings provide a reproducible and clinically promising blueprint for integrated AI-assisted breast cancer screening combining diagnostic classification, explainability, and robotic procedural guidance. Future work should focus on multi-institutional prospective validation, regulatory framework development, and multi-omics integration.

Data Availability Statement

CBIS-DDSM is accessible via The Cancer Imaging Archive. INbreast is available on request from the original authors. VinDr-Mammo is hosted at <https://physionet.org/content/vindr-mammo/>. TCGA and GEO data are publicly accessible at <https://portal.gdc.cancer.gov/> and <https://www.ncbi.nlm.nih.gov/geo/>, respectively.

Hypothetical population parameters and model code are available from the corresponding author upon reasonable request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

Grammarly was used for grammar and spelling checks during manuscript preparation. No generative AI tools were used for content creation. The authors bear full responsibility for the published content.

Ethical Approval and Consent to Participate

This study used publicly available and de-identified mammographic datasets (CBIS-DDSM, INbreast, VinDr-Mammo, TCGA, GEO) for model training and evaluation; no patient identifiable information was accessed or collected from these sources. The radiologist panel evaluation of XAI outputs was conducted under IRB Exemption No. WUST-2025-XAI-001 granted by the Washington University of Science and Technology Institutional Review Board. Participating radiologists provided written informed consent prior to the evaluation.

References

- [1] Xiong, X., Zheng, L. W., Ding, Y., Chen, Y. F., Cai, Y. W., Wang, L. P., ... & Yu, K. D. (2025). Breast cancer: pathogenesis and treatments. *Signal transduction and targeted therapy*, 10(1), 49. [CrossRef]
- [2] Bucchi, L., Mancini, S., Biggeri, A., Vattiato, R., Giuliani, O., Ravaioli, A., ... & Falcini, F. (2025). Mammography screening and incidence of ductal carcinoma in situ of the breast in Italy: an age-period-cohort analysis. *International Journal of Epidemiology*, 54(4), dyaf102. [CrossRef]
- [3] Gill, H. S., Masood, S., Nasim, F., & Usman, M. (2025). U-Net++ based breast cancer detection. *Spectrum of Engineering Sciences*, 3(12), 727-734. <https://thesesjournal.com/index.php/1/article/view/1721>
- [4] Majidpour, J., Ahmed, H. A., Ahmed, M. H., Jalal, S. I., & Arabi, H. (2026). Applications of GAN models in breast cancer detection: a comprehensive review. *Archives of Computational Methods in Engineering*, 33(1), 859-915. [CrossRef]
- [5] Tanveer, H., Faheem, M., Khan, A. H., & Adam, M. A. (2025). AI-powered diagnosis: A machine

- learning approach to early detection of breast cancer. *International Journal of Engineering Development and Research*, 13(2), 153-166. <https://www.researchgate.net/publication/391015863>
- [6] Ema, E. A. N., Reem, S. Z., Samsuzzaman, M., Rahman, M. M., & Masud, M. A. (2025, June). Ai-driven breast cancer detection: Leveraging transfer learning for enhanced radiographic analysis. In *2025 2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)* (pp. 1-6). IEEE. [CrossRef]
- [7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. [CrossRef]
- [8] Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., ... & Geras, K. J. (2021). An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical image analysis*, 68, 101908. [CrossRef]
- [9] Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F. M., Tengg-Kobligk, H. V., ... & Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3), e190043. [CrossRef]
- [10] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health*, 3(11), e745-e750. [CrossRef]
- [11] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5), 206-215. [CrossRef]
- [12] Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1), 170177. [CrossRef]
- [13] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2), 236-248. [CrossRef]
- [14] Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., & Vu, V. (2023). VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1), 277. [CrossRef]
- [15] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). [CrossRef]
- [16] Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2019). A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and prognostic research*, 3(1), 18. [CrossRef]
- [17] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139-144. [CrossRef]
- [18] Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., ... & Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35, 303-312. [CrossRef]
- [19] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626). [CrossRef]
- [20] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- [21] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). [CrossRef]
- [22] Salahuddin, Z., Woodruff, H. C., Chatterjee, A., & Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140, 105111. [CrossRef]
- [23] Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11), 4793-4813. [CrossRef]
- [24] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44-56. [CrossRef]
- [25] Khan, F., Feng, W., Huang, T., Wang, Z., Liu, X., Ali, S. A., & Wang, W. (2026). Reinforcement learning for precision grasping and safety-critical coordination in a robotic arm. *Intelligent Service Robotics*, 19(1), 16. [CrossRef]
- [26] Siviengphanom, S., Gandomkar, Z., Lewis, S. J., & Brennan, P. C. (2023). Global radiomic features from mammography for predicting difficult-to-interpret normal cases. *Journal of Digital Imaging*, 36(4), 1541-1552. [CrossRef]
- [27] Li, H., Zhu, Y., Burnside, E. S., Huang, E., Drukker, K., Hoadley, K. A., ... & Giger, M. L. (2016). Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. *NPJ breast cancer*, 2, 16012. [CrossRef]
- [28] Yala, A., Lehman, C., Schuster, T., Portnoi, T., & Barzilay, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1), 60-66. [CrossRef]

Abbreviation	Full term
ACR	American College of Radiology
AI	Artificial Intelligence
AUC	Area Under the Receiver Operating Characteristic Curve
BI-RADS	Breast Imaging Reporting and Data System
CBIS-DDSM	Curated Breast Imaging Subset of DDSM
CLAHE	Contrast-Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
DSC	Dice Similarity Coefficient
GAN	Generative Adversarial Network
GEO	Gene Expression Omnibus
GLCM	Grey-Level Co-occurrence Matrix
Grad-CAM	Gradient-weighted Class Activation Mapping
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
PCA	Principal Component Analysis
pCR	Pathological Complete Response
ROI	Region of Interest
SHAP	SHapley Additive exPlanations
TCGA	The Cancer Genome Atlas
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence

A Abbreviations

B Appendix

B.1 CNN–ViT Hyperparameter Configuration

Table A1 presents the complete hyperparameter configuration for CNN–ViT training. All parameters were fixed prior to cross-validation and were not tuned on held-out test data.

Table A1. CNN–ViT training hyperparameter configuration.

Hyperparameter	CNN Baseline	CNN–ViT (Proposed)
Backbone	ResNet-50	ResNet-50 + ViT-B/16
Optimiser	AdamW	AdamW
Learning rate	1×10^{-4}	5×10^{-5}
Batch size	32	16
Epochs	50	80
LR schedule	StepLR ($\gamma = 0.1$)	Cosine annealing
Weight decay	1×10^{-4}	5×10^{-5}
Attention heads (ViT)	—	8
Transformer layers	—	12
Patch size	—	16×16 px
Mixed precision	FP32	FP16 / BF16

B.2 Cross-Validation Fold Performance

Table A2 reports AUC and F1-score per cross-validation fold for the CNN–ViT model, confirming performance stability across dataset partitions.

Table A2. CNN–ViT per-fold cross-validation performance.

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
AUC	0.941	0.949	0.952	0.944	0.950
F1-Score	94.7%	95.4%	95.8%	94.9%	95.2%
Sensitivity	95.8%	96.4%	96.9%	95.7%	96.3%
Specificity	93.1%	94.0%	94.4%	93.5%	94.0%
Mean $\pm$ SD (AUC): 0.947 $\pm$ 0.005					



Fahad Naseer received the B.S. degree in Cyber Engineering Technology from the University of Engineering and Technology (UET), Taxila, Pakistan. He is currently pursuing the Master of Science in Information Technology (MSIT) degree, with a concentration in Data Analytics, at Washington University of Science and Technology, Alexandria, Virginia, USA, and is expected to graduate in September 2026.

His research interests include artificial intelligence, healthcare analytics, machine learning applications in medicine, predictive healthcare systems, and data-driven decision-making in clinical environments. His current research focuses on leveraging AI and advanced analytics to improve healthcare outcomes, optimize clinical workflows, and support intelligent healthcare decision-support systems. (Email: nasfahad@student.wust.edu)



Ahmad Jamal is currently pursuing the Master of Business Administration in Business Analytics degree at the King Graduate School, Monroe University, New Rochelle, New York, USA. His research interests include artificial intelligence in healthcare, predictive healthcare analytics, healthcare operations management, clinical data intelligence, and healthcare resource optimization. His work focuses on the development of AI-based predictive models

for patient care, healthcare scheduling, revenue optimization, and decision support systems. He is particularly interested in leveraging advanced analytics and machine learning techniques to address complex challenges in healthcare delivery and improve the efficiency, quality, and sustainability of healthcare services. (Email: ajamal7503@monroeu.edu)



Fatima Tauseef received the Master of Business Administration in Information Technology (MBA-IT) degree from the Department of Business Administration and Management, Washington University of Science and Technology, Alexandria, Virginia, USA. She is an independent researcher whose work focuses on artificial intelligence in healthcare, healthcare analytics, predictive modeling, clinical decision-support systems, and digital health transformation. Her research explores the application of machine learning, data analytics, and intelligent healthcare technologies to improve patient outcomes, optimize healthcare operations, and support evidence-based decision-making. Her scholarly interests include healthcare data science, AI-driven healthcare innovation, and the development of intelligent systems for improving the quality, accessibility, and efficiency of healthcare services. She continues to contribute to interdisciplinary research addressing emerging challenges

in healthcare through advanced analytical and technological approaches. (Email: ftauseef.student@wust.edu)



Muhammad Fawad Nasim is an Assistant Professor in the Department of Computer Science at Superior University. He holds a Ph.D. in Artificial Intelligence and serves as the Quality Circle Head overseeing Data Science, Image Processing, Computer Vision, and Generative AI. A researcher in advanced computing, his work explores machine learning frameworks, acoustic signal processing, explainable artificial intelligence (XAI), and Medical Imaging. Beyond his faculty and research duties, he serves as an Advisor for a university research society, mentoring students and leading academic initiatives. (Email: fawad.nasim@superior.edu.pk)