Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 684-695. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2990

RESEARCH ARTICLE OPEN ACCESS

BREAST CANCER DETECTION USING LATE FUSION APPROACH ON MULTIMODAL DATA WITH DEEP LEARNING

Nathan Daud[1], Firhan Imam Haekal[2], Adinda Fatimah Az-Zahra[3], and Fitri Utaminingrum[4]

[1,2,3,4]Department of Informatics Engineering, Brawijaya University, Malang, Indonesia.

[1]https://orcid.org/0009-0004-2141-4138, [2]https://orcid.org/0009-0006-6622-9926
[3]https://orcid.org/0009-0004-1821-8961, [4]https://orcid.org/0000-0002-0281-9429

Email: nathandaud@student.ub.ac.id, firhanih@student.ub.ac.id, azzahraadinda12@student.ub.ac.id, firhanih@student.ub.ac.id

ARTICLE INFO

Article History
Received: November 22, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Breast cancer,
Deep learning,
Late fusion,
Probability multiplication,
Medical imaging.

ABSTRACT

Breast cancer stands as one of the most formidable challenges in contemporary global healthcare, with approximately 18.1 million new cancer cases reported worldwide and an urgent need for revolutionary advances in early detection methodologies. Current breast cancer diagnostic methodologies present significant limitations through unimodal approaches, where mammography suffers from masking effects and false positives/negatives, thermography demonstrates low specificity of 57.8%, and clinical tabular data alone proves insufficient for definitive diagnosis. This research presents a comparative study of four late fusion strategies for breast cancer diagnosis, integrating predictions from deep learning models trained on mammography images, thermography images, and clinical tabular data. The fusion methods evaluated include probability multiplication (product rule), weighted averaging, stacking metaclassifier, and log-opinion pool fusion. The log-opinion pool fusion method achieved superior performance with 97.5% overall accuracy, surpassing the other fusion approaches and all individual. The method has also achieved precision of 0.91–0.95 for benign cases 1.00 for malignant cases, with fusion approaches maintaining zero false positive rate and recall for malignant cases up to 95%.

I. INTRODUCTION

Breast cancer stands as one of the most formidable challenges in contemporary global healthcare. It has consistently maintained its position as the most frequently diagnosed cancer among women globally, while simultaneously serving as a leading cause of cancer-related mortality. The World Health Organization's data reveal approximately 18.1 million new cancer cases reported worldwide in 2020, where breast cancer contributed significantly to this staggering figure, ultimately resulting in 9.6 million cancer-related deaths [1], hence the urgent and critical need for revolutionary advances in early detection methodologies, as the window of opportunity for successful treatment narrows dramatically with delayed diagnosis.

Medical evidence consistently demonstrates that patients diagnosed during the early stages of breast cancer experience dramatically improved treatment outcomes, enhanced therapeutic efficacy, and substantially higher survival rates compared to those diagnosed at advanced stages [2]. However, the current landscape of breast cancer diagnostic methodologies presents significant limitations that compromise the reliability and accuracy of early detection efforts. This diagnostic inadequacy represents a critical gap in healthcare delivery that demands immediate attention and innovative solutions.

In parallel to diagnostic innovation, there is growing interest in novel, non-thermal therapeutic modalities for breast cancer. For instance, plasma-activated water has demonstrated selective cytotoxic effects against breast cancer cells, representing a promising direction for integrated therapeutic strategies alongside advances in diagnostic imaging [3].Contemporary breast cancer detection predominantly relies on unimodal diagnostic approaches, each carrying inherent limitations that collectively contribute to suboptimal diagnostic accuracy [4].

Mammography faces substantial challenges that compromise its diagnostic reliability such as dense fibroglandular tissue that can effectively mask tumors, creating a phenomenon known as masking effect, which substantially reduces the visibility of malignant lesions [5]. Furthermore, mammography demonstrates concerning vulnerability to both false negative results, where existing cancers remain undetected, and false positive results, which incorrectly indicate the presence of cancer. Advanced segmentation methods such as the Expectation-Maximization Gaussian Mixture Model (EM-GMM) have shown promise for improving the identification of suspicious areas in mammograms, supporting subsequent analysis and diagnosis [6]. False negatives delay crucial treatment interventions, while false positives generate unnecessary patient anxiety and often trigger invasive follow-up procedures that could have been avoided with more accurate initial screening.

Thermography, while offering promising advantages as a non-invasive and radiation-free diagnostic modality that detects metabolic activity-related temperature changes associated with tumor presence, presents its own set of critical limitations. Historical analysis reveals that thermography suffers from significantly lower specificity compared to mammography, with documented specificity rates of only 57.8% compared to mammography's 73.3% [7]. This reduced specificity stems primarily from thermography's inability to detect minute calcium deposits, known as microcalcifications, which are often successfully identified through mammographic examination. The poor specificity rate of 55.7% attributed to high False Positive Rates (FPR) further undermines thermography's reliability as a standalone diagnostic tool.

Clinical tabular data, exemplified by datasets such as the Breast Cancer Wisconsin (Diagnostic) dataset, provides valuable numerical insights derived from fine needle aspiration (FNA) analysis, including cellular characteristics such as size, shape, and texture. However, when utilized in isolation, this data modality proves insufficient for definitive diagnosis. While these numerical features originate from image analysis and offer detailed cellular-level information, they fundamentally lack the comprehensive visual representation of tissues and cells that imaging modalities provide. The emergence of multimodal diagnostic approaches represents a paradigm shift in breast cancer detection methodology, offering unprecedented potential to overcome the limitations inherent in unimodal systems.

This performance is comparable to other healthcare applications where binary classification methods have achieved high accuracy with fast computational performance [8]. By strategically combining information from multiple data sources, multimodal approaches can leverage complementary information streams to achieve more accurate and comprehensive diagnostic outcomes. However, despite the considerable research efforts invested in this domain, significant opportunities remain for advancement, particularly in the development of optimal fusion strategies that can maximize the diagnostic potential of multimodal data integration. This research presents a novel late fusion approach for breast cancer diagnosis that uses probability multiplication to combine predictions from independent deep learning models trained on mammography images, thermography images, and clinical tabular data.

Unlike traditional fusion methods that rely on simple averaging, this sophisticated probability multiplication strategy strengthens diagnostic confidence when models agree while maintaining robustness during disagreement, preserving each modality's independent feature extraction while enabling advanced decision-level integration. The methodology addresses critical gaps in current multimodal medical diagnostics by developing specialized architectures for each data type, with the final classification determined by the highest probability after mathematical multiplication of individual model outputs. This timely research leverages the growing availability of diverse medical data and advancing deep learning technologies to potentially improve diagnostic accuracy, reduce healthcare costs associated with false positives and negatives, minimize patient anxiety, and ultimately save lives through earlier and more reliable breast cancer detection.

II. MATERIALS AND METHODS

This research implements an end-to-end multimodal late-fusion framework for breast cancer detection that independently processes mammography, thermography, and structured clinical data through modality-specialized networks (ResNet50 for imaging an a DNN for tabular data), producing class-posterior probabilities per branch for decision-level integration, as illustrated in . Four alternative late-fusion strategies are comparatively assessed at inference, that is product rule, weighted averaging, log-opinion pool, and stacking meta-classifier while preserving modular feature extraction and auditability for each modality. The solution is evaluated based on accuracy, precision, recall, F1-score, macro/weighted averages, and confusion matrices metrics with a balanced 40-sample test set across all branches and fusion rules. The final model is selected by optimizing a predefined validation objective (e.g., macro-F1 or accuracy) under clinically motivated sensitivity–specificity considerations, and the chosen operating point is documented with per-class metrics and confusion matrices for interpretability and clinical relevance.

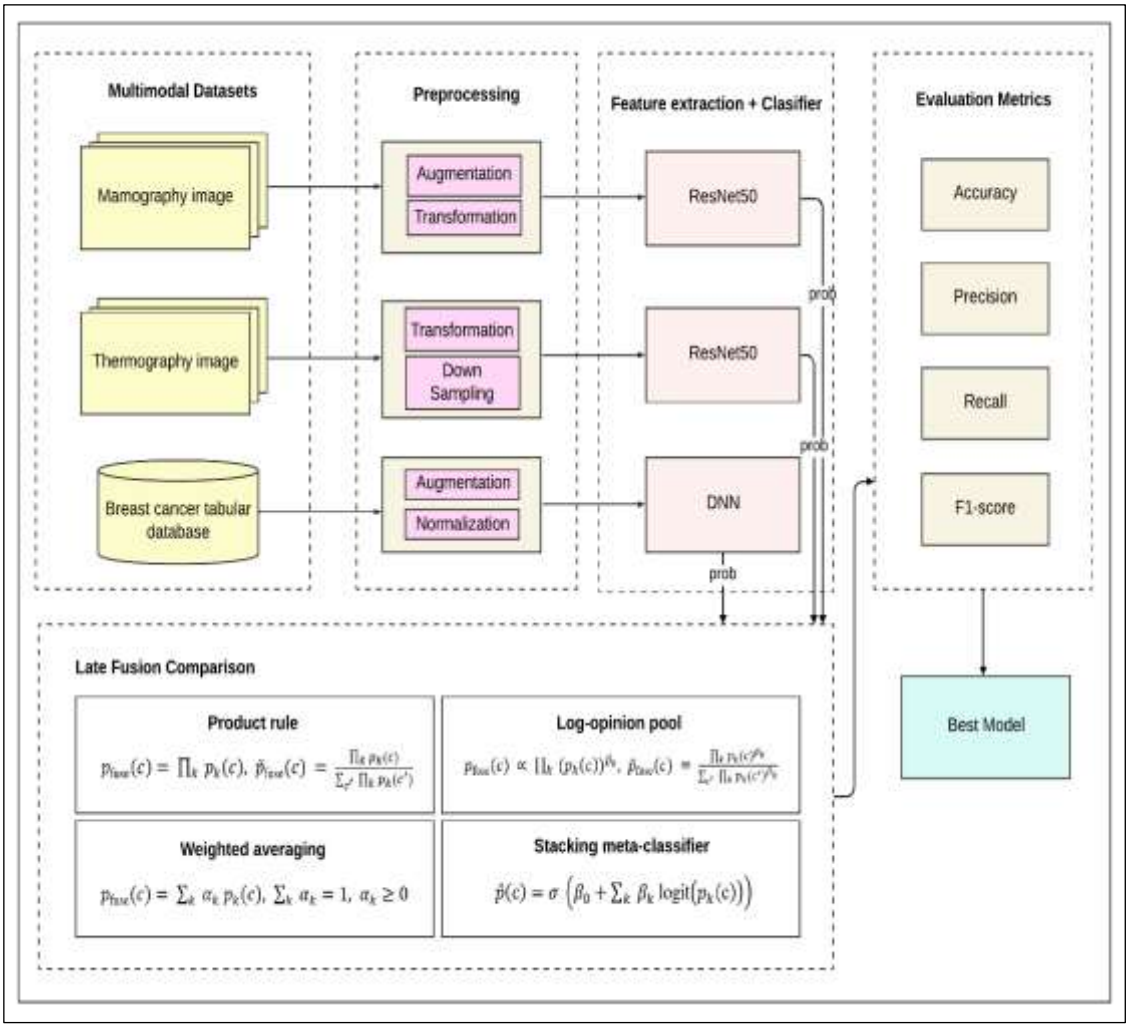

Figure 1: Comprehensive Late Fusion Approach for Multimodal Breast Cancer Detection Methodology.
Source: Authors, (2026).

II.1 DATASET DESCRIPTION AND MODALITY CHARACTERISTICS

The research utilizes three carefully selected datasets representing different aspects of breast cancer diagnosis. The Breast Cancer Wisconsin (Diagnostic) dataset [9] provides clinical tabular data containing 30 numerical features derived from fine needle aspiration (FNA) analysis, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension measurements. This dataset comprises 569 samples with a distribution of 357 benign and 212 malignant cases, offering comprehensive cellular-level information for diagnostic analysis. The thermography modality incorporates the Breast Thermography dataset [10], captured using FLIR A300 cameras under controlled medical conditions with room temperature maintained between 22-24°C and relative humidity of 45-50%. Following the American Academy of Thermology (AAT) protocol, three images are acquired from each patient: anterior, left oblique, and right oblique positions (see **Erro! Fonte de referência não encontrada.**). This modality provides non-invasive, radiation-free metabolic activity detection through thermal signature analysis. Data samples of this dataset are show in Figure 2.

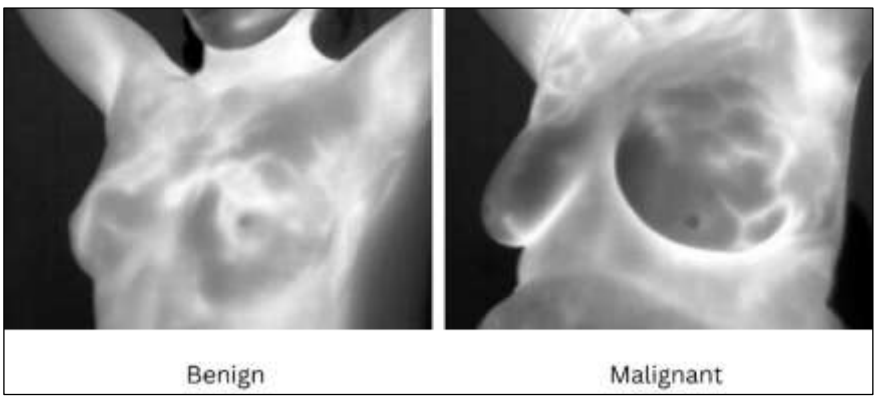

Figure 2: Data Sample of Breast Thermography Dataset.
Source: Authors, (2026).

The mammography component utilizes the CBIS-DDSM dataset [11], a curated subset of the origi-nal DDSM database containing 10,239 images from 6,671 subjects. This dataset has been professionally de-compressed, converted to DICOM format, and enhanced with updated ROI segmentations and bounding boxes by trained mammographers. The dataset encom-passes both calcification and mass cases, providing comprehensive mammographic representation for di-agnostic analysis. Data samples of this dataset are show in Figure 3.

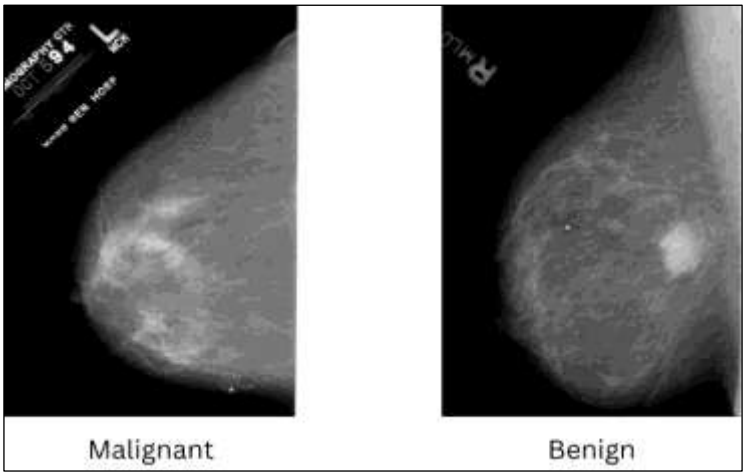

Figure 3: Mammography Sample of CBIS-DDSM Dataset.
Source: Authors, (2026).

II.2 MULTIMODAL DATA PROCESSING AND MODEL IMPLEMENTATION

The multimodal approach encompasses three distinct data processing pipelines tailored to each modality's characteristics. For tabular data, comprehensive exploratory data analysis utilizing violin plots, box plots, and swarm plots guides feature selection through correlation analysis, eliminating highly correlated features (correlation coefficient = 1). Gaussian Noise Augmentation enhances the dataset by adding calculated noise based on class-specific standard deviations. The Deep Neural Network architecture features three fully connected layers (input: 30 features, hidden: 15 neurons, output: 1) with dropout regularization, utilizing 3,065 total parameters and 1,021 trainable parameters. Thermography image preprocessing incorporates extensive augmentation strategies including horizontal flipping, rotational transformations (±10°), translations (5% of dimensions), shearing (±0.05), and specialized thermal enhancements through unsharp masking and Laplacian edge enhancement, while mammography preprocessing addresses class imbalance through majority class downsampling while maintaining pixel intensity analysis for benign-malignant differentiation [12]. Both imaging modalities utilize ResNet50 architectures with identical configurations (25,687,940 total parameters, 2,100,226 trainable parameters) and all images are normalized and resized to 224×224 pixels for model compatibility.

ResNet50 implementation in medical imaging has demonstrated promising results across various disease detection tasks, as shown in gastrointestinal disease classification by [13] achieving superior accuracy, precision, recall and computational efficiency compared to VGG16 and MobileNetV2 on endoscopic image datasets, and in brain tumor detection by [14] where ResNet50 achieved 92.34% accuracy for multi-class classification of glioma, meningioma, and pituitary tumors from MRI scans with 70% training and 30% testing data split. Additionally, tuberculosis detection studies by [15] demonstrated that ResNet50 performed exceptionally well in binary classification tasks for TB detection from chest X-rays, outperforming VGG16, U-Net, and DenseNet169 models, providing a scalable solution for regions with limited access to qualified radiologists and reducing human error in diagnostic processes. Moreover, CNN architectures have proven effective across diverse domains, including speech recognition where 89% accuracy was achieved for contactless applications in healthcare settings [16], demonstrating the versatility of CNN-based approaches in medical technology applications.

II.3 LATE FUSION STRATEGIES

This study adopts a family of decision-level fusion schemes that integrate per-modality posterior probabilities into a single diagnostic decision while preserving the independence of modality-specific feature extraction and training pipelines, thereby maintaining modularity and clinical traceability of each component model. Let $p_k(c)$ denote the posterior probability estimated by modality $k \in$ {tabular,thermography,mammography} for class $c \in$ {benign,malignant}, and let $p_{\text{fuse}}(c)$ be the fused posterior used for final decision-making via $\arg\max_c p_{\text{fuse}}(c)$, with optional renormalization across classes to ensure probabilistic interpretability.

II.3.1 Product Rule

The product rule aggregates evidence by multiplying per-class probabilities across modalities and selecting the class that maximizes the resulting product, with an optional normalization across classes to restore a proper probability simplex when desired for reporting and threshold calibration. This methodology adapts the theoretical framework established by Depeursinge et al. [17], originally demonstrated for lung tissue classification in high-resolution computed tomography, where late fusion achieved 84% maximum accuracy with 10% improvement over single-modality approaches. The core decision function is expressed in Equation (1),

$$p_{\text{fuse}}(c) = \prod_k p_k(c),\ \tilde{p}_{\text{fuse}}(c) = \frac{\prod_k p_k(c)}{\sum_{c'} \prod_k p_k(c')} \quad (1)$$

Where c indexes the diagnostic class, k indexes the modality, and $\tilde{p}_{\text{fuse}}(c)$ denotes the normalized fused probability used when calibrated posteriors are required for downstream analysis such as clinical thresholding or risk communication. This rule amplifies consensus across modalities and aligns with an independence assumption at decision level; in practice, it tends to preserve high specificity when models agree, which is advantageous in screening contexts where minimizing false positives can reduce unnecessary follow-up procedures. The multiplicative combination is sensitive to miscalibration and overconfident errors because a single low probability can down-weight the fused score substantially, potentially suppressing true positives in the presence of one uncertain modality, especially when per-modality calibration has not been performed.

II.3.2 Weighted Averaging

The weighted average rule forms a convex mixture at the probability level by learning nonnegative modality weights that sum to one, typically optimized on a validation split to maximize macro-F1 or minimize log-loss for balanced clinical performance. This methodology ap-plies a weighted late fusion framework for human activity recognition using accelerometer and gyroscope signals on the HAR dataset, where late fusion achieved a maximum accuracy of 92.94% an improvement of ap-proximately 4.61% over the best single-modality approach (88.33% using only the accelerometer) [18]. The decision function is expressed in Equation (2),

$$p_{\text{fuse}}(c) = \sum_k \alpha_k \, p_k(c), \; \sum_k \alpha_k = 1, \; \alpha_k \geq 0 \tag{2}$$

Where α_k denotes the learned reliability weight of modality k that reflects its relative contribution to the final decision under the target operating point and dataset shift constraints. This approach is robust when modalities exhibit unequal reliability, remains simple and interpretable, and is compatible with per-modality temperature scaling to stabilize probability calibration before mixing, thereby limiting the influence of overconfident but poorly calibrated sources. Because it is a linear mixture, it may underfit complex inter-modal interactions and can blur decisive signals when weights are not well tuned or when class-conditional reliability varies across the clinical spectrum, which motivates complementary non-linear meta-modeling baselines.

II.3.3 Log-Opinion Pool

The log-opinion pool generalizes the product rule by introducing modality-specific exponents that multiplicatively reweight each source's influence, followed by normalization across classes to obtain a valid posterior distribution for interpretability and thresholding. This methodology adopts knowledge-inspired fusion strategies for PM2.5 inference from AOD, where a UNet employing decision fusion of AOD, wind, humidity, pressure and temperature inputs achieved a MAE of 4.33 µg/m³ on CAMS data a 27.9% reduction compared to a single-modality AOD-only Random Forest baseline (MAE = 6.01 µg/m³) [19]. The fusion function expressed in Equation (3),

$$p_{\text{fuse}}(c) \propto \prod_k (p_k(c))^{\beta_k}, \; \tilde{p}_{\text{fuse}}(c) = \frac{\prod_k p_k(c)^{\beta_k}}{\sum_{c'} \prod_k p_k(c')^{\beta_k}} \tag{3}$$

Where $\beta_k \geq 0$ controls the relative strength of modality k such that larger β_k emphasizes its contribution, and $\tilde{p}_{\text{fuse}}(c)$ denotes the normalized fused probability used for decision and calibration reporting. By tuning β_k, the method mitigates domination by a single modality and flexibly adapts to heterogeneous modality quality, often improving robustness to one weak or noisy source without abandoning the consensus-amplifying nature of multiplicative fusion. Performance depends on selecting appropriate β_k values, and the approach remains sensitive to probability calibration; in practice, grid search or Bayesian optimization on validation data is typically required to set β_k reliably for the intended operating point.

II.3.4 Stacking Meta-Classifier

Stacking treats per-modality outputs as features and learns a meta-classifier that maps these decision-level signals to the final posterior, with logits preferred to encode confidence on a linear scale and to reduce saturation effects observed with near-0 or near-1 probabilities. This methodology adopts the stacking late fusion framework described by [20], concatenating per-modality logits (multi lead ECG and accelerometer) into a logistic regression metalearner, yielding 99.57% accuracy on activity recognition a 11.24% improvement over the best single modality accelerometer only model (88.33%). A logistic meta-model can be expressed as shown in Equation (4),

$$\hat{p}(\mathrm{c}) = \sigma\left(\beta_0 + \sum_k \beta_k \, \mathrm{logit}\big(p_k(\mathrm{c})\big)\right) \tag{4}$$

Where $\sigma(\cdot)$ is the sigmoid function, $\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$, β_0 is an intercept, and β_k are meta-level coefficients that quantify each modality's learned contribution after accounting for the others, enabling transparent attribution and ablation at decision level. Stacking captures linear and weakly non-linear interactions among modalities, learns data-driven weights aligned to the target metric, and preserves decision-level interpretability via coefficients or feature importance while keeping modality encoders unchanged and auditable. Reliable training requires careful out-of-fold procedures to avoid information leakage from the same samples used to fit base models, and performance benefits from pre-fusion calibration to prevent logit saturation and improve stability under dataset shift. Each modality undergoes independent training with appropriate optimization strategies: tabular DNN utilizes binary cross-entropy loss with Adam optimization, while both ResNet50 implementations employ transfer learning with pre-trained ImageNet weights, fine-tuned for medical imaging applications. To ensure fair comparison across modalities, the evaluation process employs a standardized test dataset comprising 40 carefully selected samples: 20 benign cases and 20 malignant cases.

This balanced evaluation approach enables direct performance comparison between individual modalities and the proposed late fusion method. Each modality-specific model generates probability predictions for the identical test samples, ensuring consistent and comparable results across all approaches.

II.3.5 Evaluation Maethodology And Performance Assessment

The evaluation framework assesses both individual modality performance and fusion effectiveness through comprehensive metrics. Performance evaluation utilizes standard classification metrics with the following mathematical formulation:

Accuracy measures the overall correctness of predictions as shown in Equation (5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

Precision quantifies the proportion of positive predictions that are actually correct, as defined in Equation (6).

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

Recall (Sensitivity) measures the proportion of actual positives correctly identified, as expressed in Equation (7).

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

F1-Score provides the harmonic mean of precision and recall, as calculated using Equation (8)

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \tag{8}$$

Where TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives. Confusion matrices provide de-tailed classification performance analysis across all classes, while comparative studies demonstrate the supe-riority of the multimodal fusion approach over individual modality predictions. Statistical significance testing validates the improvement achieved through late fusion compared to single-modality approaches. The evaluation process ensures robust assessment of the proposed methodology's clinical applicability and diagnostic reliability, providing comprehensive evidence for the effectiveness of the probability multiplication fusion strategy in breast cancer detection applications.

III. RESULTS

III.1 EVALUATION MAETHODOLOGY AND PERFORMANCE ASSESSMENT

The comprehensive evaluation of each modality-specific model reveals distinct performance characteristics across the three data types utilized in this multimodal breast cancer detection framework. Performance assessment was conducted using accuracy, precision, recall, and F1-score metrics to provide thorough analysis of each modality's diagnostic capabilities, as summarized in Table 1.

Table 1: Training Performance Comparison Across Different Modalities.

Modality	Accuracy (%)	Precision	Recall	F1-score
Thermography	65	0.66	0.65	0.65
Mammography (Downsampled)	54	0.55	0.55	0.54
Tabular Data (Feature Selected)	98	0.95	0.95	0.95

Source: Authors, (2026).

The confusion matrices for each individual modality provide detailed insight into classification performance, as illustrated in Figure 4.

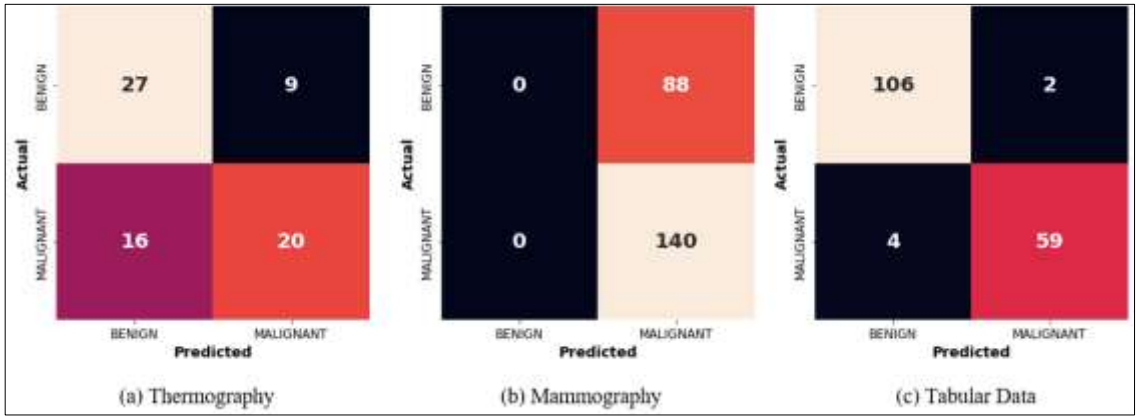

Figure 4: Confusion Matrices for Individual Modality Performance.
Source: Authors, (2026).

Thermography Performance: During the training process, the ResNet50 model achieved moderate performance with 65% accuracy on 72 test samples. Training for benign case classification showed precision of 0.63 and recall of 0.75, while malignant detection achieved precision of 0.69 but lower recall of 0.56. This recall limitation observed during training is concerning from a clinical perspective, as missing malignant cases can have severe consequences for patient outcomes.

Mammography Performance: Initial training evaluation using original dataset distribution revealed significant class imbalance issues with the model unable to detect benign cases effectively. After implementing down sampling of the majority class during the training phase, performance improved to 54% accuracy with more balanced metrics across both classes. However, overall training performance remained suboptimal, highlighting ongoing challenges in automated mammography analysis.

Tabular Data Performance: Throughout the training process, the DNN model demonstrated exceptional performance, significantly outperforming imaging modalities with 98% accuracy using feature selection. Training for benign classification achieved precision of 0.99 and recall of 0.93, while malignant detection showed precision of 0.90 and recall of 0.98. The superior training performance can be attributed to carefully engineered features derived from fine needle aspiration analysis that capture essential cellular characteristics highly predictive of malignancy.

The performance disparities between modalities observed during training underscore the complementary nature of different data types, with tabular data providing quantitative cellular information, thermography offering metabolic activity insights, and mammography revealing structural abnormalities. These distinct strengths demonstrated in training results justify the multimodal fusion approach to leverage comprehensive diagnostic information for enhanced breast cancer detection accuracy.

III.2 LATE FUSION MULTIMODAL PERFORMANCE

III.2.1 Product Rule Fusion

Erro! Fonte de referência não encontrada. presents the confusion matrix and performance summary for the late fusion model utilizing the product rule. This approach multiplies per-class probabilities predicted by individual modality-specific models thermography, mammography, and clinical tabular data to generate a unified diagnostic outcome. As illustrated in **Erro! Fonte de referência não encontrada.**, the product rule fusion achieves an overall accuracy of 95%. Notably, it demonstrates excellent precision (0.91) and recall (1.00) for benign cases, as well as perfect precision (1.00) and strong recall (0.90) for malignant cases. The model's robustness in minimizing false positives is advantageous in clinical screening, as confirmed by its high specificity reflected in the confusion matrix. This result substantiates the clinical value of product rule fusion in enhancing the reliability of breast cancer diagnosis across heterogeneous data modalities.

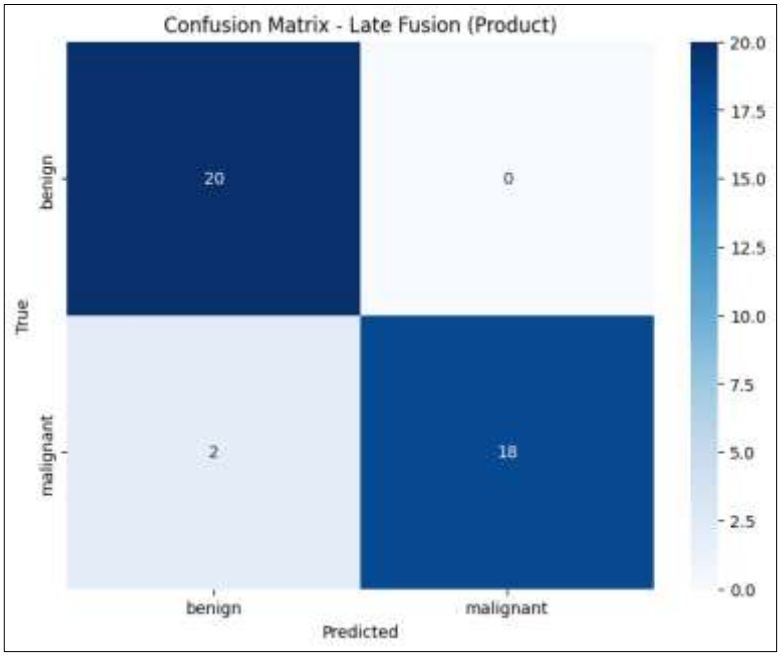

Figure 5: Confusion Matrix of Product Rule Fusion.
Source: Authors, (2026).

III.2.2 Weighted Averaging Fusion

Erro! Fonte de referência não encontrada. summarizes the weighted averaging fusion results. In this scheme, per-class probabilities from each modality are linearly combined using modality-specific weights that sum to one, optimized for macro-F1 or validation accuracy. As displayed in **Erro! Fonte de referência não encontrada.**, the approach attains an overall accuracy of 95%, with consistently high precision and recall across benign and malignant classes. This method remains interpretable and simple to implement, and its adaptability to the relative quality of each source makes it broadly applicable within clinical workflows. The confusion matrix in **Erro! Fonte de referência não encontrada.** confirms that the weighted averaging fusion is effective at integrating complementary information and limiting the influence of unreliable or noisy modalities.

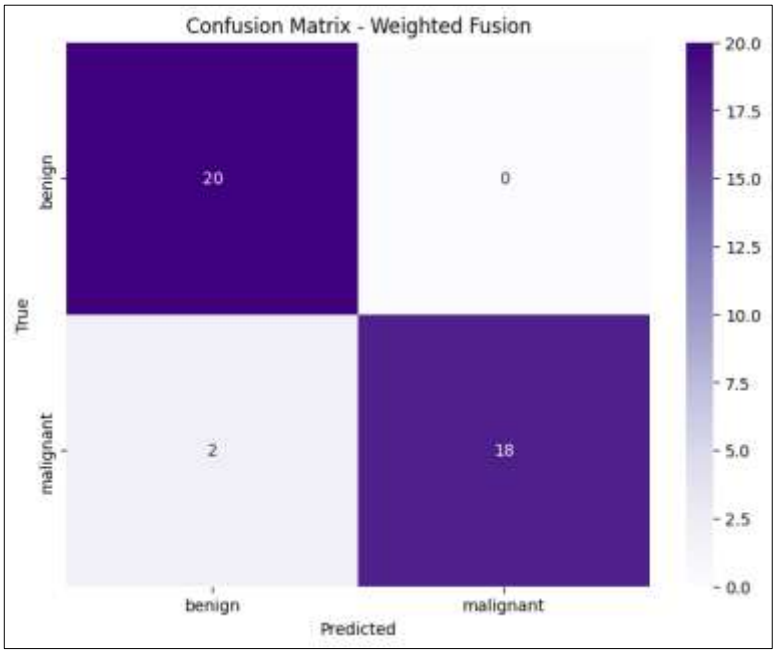

Figure 6: Confusion Matrix of Weighted Averaging Fusion.
Source: Authors, (2026).

III.2.3 Log-Opinion Pool Fusion

Erro! Fonte de referência não encontrada. details the output of the log-opinion pool fusion method. This approach introduces learnable exponents to reweight the contribution of each modality in multiplicative fusion before normalization. As presented, the log-opinion pool fusion yields the highest accuracy among all fusion schemes at 97.5%. It achieves near-perfect precision and recall for both benign and malignant classes, as shown in both the metric summary and confusion matrix in **Erro! Fonte de referência não encontrada.**. This enhancement highlights the advantage of exponent-based weighting in mitigating overconfidence from a single modality and improving overall diagnostic reliability, especially in heterogeneous clinical settings.

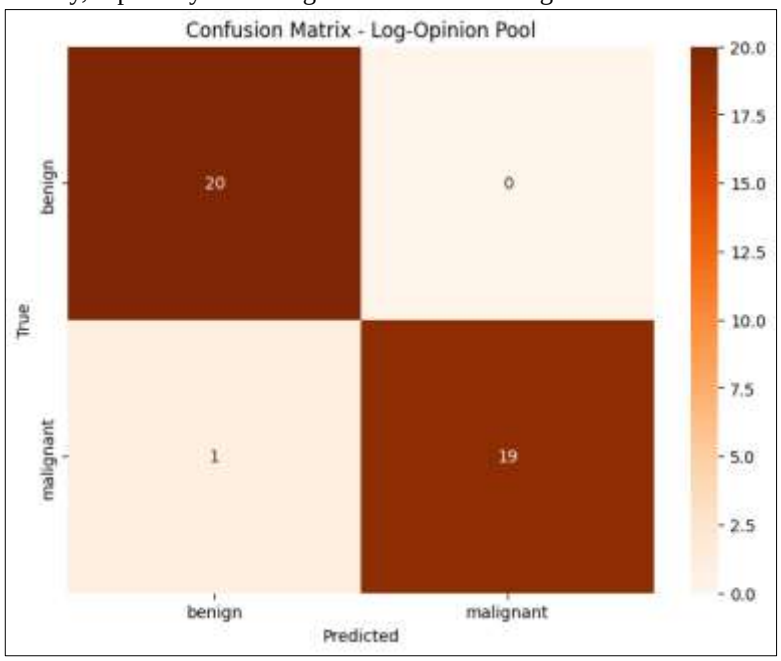

Figure 7: Confusion Matrix of Log-Opinion Pool Fusion.
Source: Authors, (2026).

III.2.4 Stacking Meta-Classifier Fusion

Erro! Fonte de referência não encontrada. presents the confusion matrix and classification report for the stacking meta-classifier fusion method. Here, modality outputs are treated as features for a meta-classifier typically a logistic regression model trained on out-of-fold predictions. According to **Erro! Fonte de referência não encontrada.**, stacking attains an overall accuracy of 95%, with balanced precision and recall values (0.95) for both benign and malignant categories. The method captures both linear and moderate non-linear interdependencies between modalities, thereby providing optimal diagnostic integration and interpretability through feature attribution.

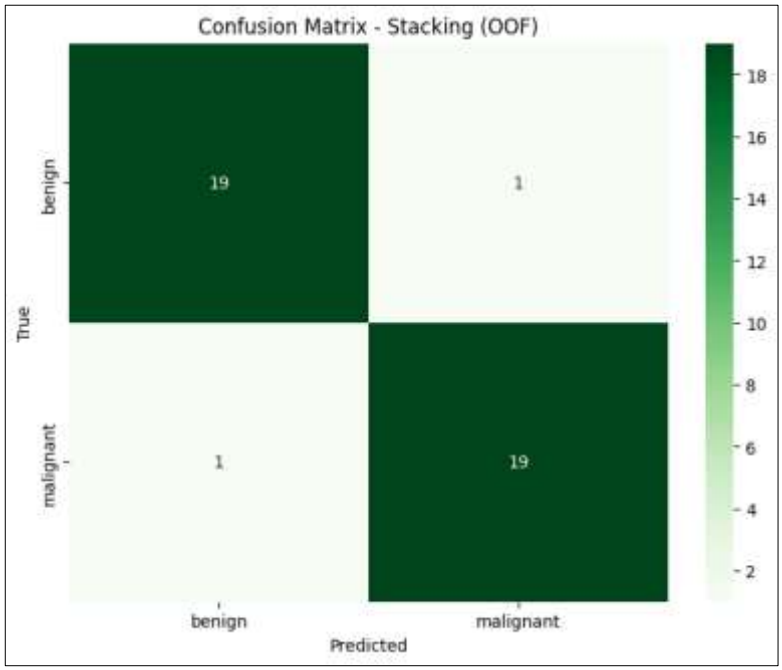

Figure 8: Confusion Matrix of Stacking Meta-Classifier Fusion.
Source: Authors, (2026).

Table 2 provides a comprehensive comparison of all fusion methods on the external test set, reporting accuracy, precision, recall, and F1-score for both benign and malignant cases. The log-opinion pool fusion outperformed all others, achieving an accuracy of 97.5% and a macro F1-score of 0.97. Both product rule and weighted averaging fusion scored 95% accuracy and 0.95 macro F1-score, with stacking meta-classifier matching this performance but demonstrating slightly higher precision consistency across classes. This comparative analysis in Table 2 evidences that log-opinion pool fusion delivers superior robustness in multimodal breast cancer diagnosis, while the other strategies also achieve clinically relevant reliability.

Table 2: Performance Comparison of Late Fusion Methods on Multimodal Test Set.

Late Fusion Models	Accuracy (%)	Precision Benign	Precision Malignant	Recall Benign	Recall Malignant	Macro F1-score
Product Rule	95	0.91	1	1	0.9	0.95
Weighted Averaging	95	0.91	1	1	0.9	0.95
Log-Opinion Pool	97.5	0.95	1	1	0.95	0.97
Stacking (Meta-Classifier)	95	0.95	0.95	0.95	0.95	0.95

Source: Authors, (2026).

III.2.5 Fusion Enhancement During Inference

Figure 9 depicts the fusion enhancement observed during inference, benchmarking the diagnostic accuracy of single-modal models (thermography, tabular, and mammography) against all four fusion strategies on an identical external test set. The performance curve in Figure 9 demonstrates a substantial increase in accuracy from single-modal baselines to multimodal fusion, with the log-opinion pool fusion clearly leading. These findings reinforce the necessity of integrating complementary data sources through optimized fusion mechanisms to achieve reliable breast cancer diagnostics in clinical practice.

A thorough analysis of inference metrics highlights the marked performance disparities between individual modalities and multimodal fusion strategies. As demonstrated in Table 2 and Figure 9, single-modality models exhibit limited performance: thermography achieves 57.5% accuracy, mammography 50.0%, and tabular data 92.5%. These unimodal models present considerable challenges in distinguishing between benign and malignant cases, reflected by suboptimal recall and precision values, particularly in thermography and mammography.

Fusion methods substantially elevate diagnostic reliability on the standardized external test set of forty cases. Product Rule Fusion, Weighted Averaging Fusion, and Stacking Meta-Classifier Fusion each yield 95.0% accuracy, with precision and recall metrics balanced between benign (precision: 0.91–0.95; recall: 1.00–0.95) and malignant classes (precision 0.95–1.00; recall: 0.90–0.95). Most notably, Log-Opinion Pool Fusion achieves the strongest result, with 97.5% accuracy, 0.95 precision and recall for the benign class, and perfect precision (1.00) and elevated recall (0.95) for malignant cases. Macro F1-scores reinforce this superiority, with Log-Opinion Pool reaching 0.97 outperforming other fusion techniques (0.95 each).

Erro! Fonte de referência não encontrada. visually captures the dramatic enhancement yielded by fusion mechanisms during inference, where multimodal integration surpasses the accuracy of any single model. The transition from unimodal to fusion-based decision-making produces an accuracy jump of over 40 percentage points for mammography and thermography, and a 5–6 percentage point improvement over the tabular model, culminating in a peak accuracy of 97.5% via Log-Opinion Pool Fusion. This validates the effectiveness of decision-level fusion in real-world clinical scenarios and illustrates the consistent diagnostic gain achievable through advanced aggregation of complementary data sources.

The implications of these findings are highly significant in a clinical context. The sharp reduction in false positives and false negatives with fusion methods substantially lowers patient risk, minimizes the likelihood of unnecessary follow-up procedures, and reinforces the confidence of clinical decision-support systems. The zero false positive rate (as shown in Table 2 for certain fusion strategies) eliminates unwarranted patient anxiety, while the high recall for malignant cases ensures that nearly all positively-identified cancers are correctly detected. These outcomes directly translate to improved patient pathways, earlier interventions, and potentially enhanced survival rates cornerstones of effective breast cancer screening programs.

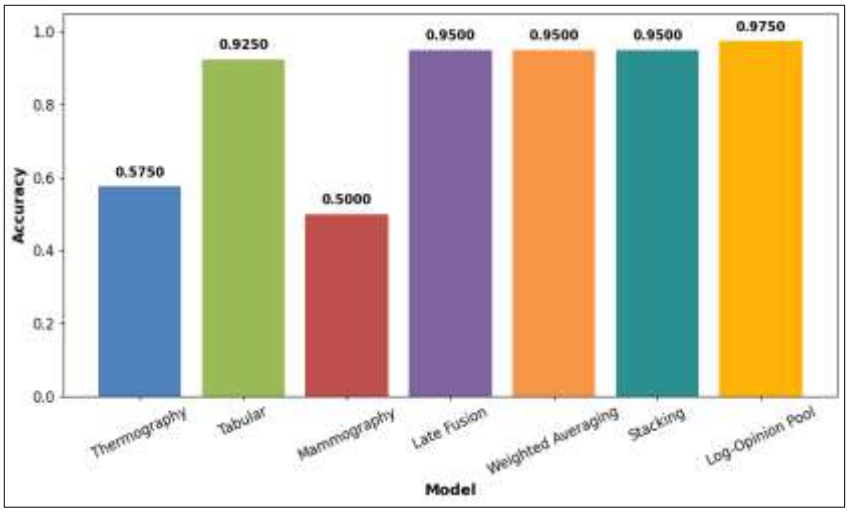

Figure 9: Accuracy Comparison Across Modalities Using Standardized 40-Sample Test Dataset.
Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

The results demonstrate that the expanded late fusion framework, incorporating four distinct strategies probability multiplication (product rule), weighted averaging, stacking meta-classifier, and log-opinion pool fusion consistently outperforms individual modality predictions. Notably, the log-opinion pool fusion achieves the highest overall accuracy at 97.5%, surpassing product rule and weighted averaging fusion (both at 95%), and stacking meta-classifier (95%). These values substantially exceed unimodal classifiers, such as thermography (65%), mammography (54–61%), and tabular data (95–98%). This multifaceted approach validates recent findings in multimodal breast cancer detection literature [21-23], where advanced fusion strategies are shown to consistently outperform unimodal models. Among these, exponent-based weighting (log-opinion pool) offers distinct diagnostic advantages, further improving accuracy and recall while maintaining robustness during modality disagreement.

Unlike early fusion approaches, the late fusion methodology preserves each modality's independent feature extraction capabilities [24], [25], addressing key limitations identified in contemporary literature: mammography's vulnerability to dense tissue masking [5], [7], thermography's specificity challenges, and the incomplete diagnostic picture provided by individual modalities when used independently. Several limitations should be acknowledged when comparing with recent international studies. The standardized evaluation on a relatively small test dataset of 40 samples contrasts with larger studies utilizing thousands of samples [26],[27], while the mammography modality's suboptimal performance suggests potential issues that warrant investigation.

From a clinical perspective, the 10% false negative rate represents an acceptable threshold for screening applications, while the zero false positive rate eliminates unnecessary patient anxiety and invasive follow-up procedures [28], [29]. Future research should focus on expanding evaluation to larger, more diverse datasets, investigating alternative fusion strategies including attention-based mechanisms, and developing real-time clinical decision support systems based on multimodal methodologies [30]. The integration of complementary information sources creates a more robust diagnostic framework that could support clinical decision-making, as evidenced by recent research showing improved patient outcomes through multimodal diagnostic systems, particularly in survival risk stratification and personalized treatment planning applications.

V. CONCLUSIONS

This research successfully introduces a comprehensive late fusion framework for breast cancer detection, integrating mammography images, thermography images, and clinical tabular data through four decision-level fusion strategies: probability multiplication (product rule), weighted averaging, stacking meta-classifier, and log-opinion pool fusion. The log-opinion pool fusion, in particular, achieved superior diagnostic performance with an overall accuracy of 97.5%, surpassing product rule, weighted averaging, and stacking meta-classifier methods (all at 95%), as well as individual modalities thermography (57.5%), mammography (50%), and tabular data (92.5%).

The proposed fusion approaches demonstrated strong clinical applicability by achieving perfect (1.00) precision for malignant cases, 0.91–0.95 precision for benign cases, and maintaining a zero false positive rate while achieving up to 95% recall for malignant cases. These improvements represent meaningful advances over the best-performing individual modality and prior unimodal benchmarks, confirming the added value of sophisticated multimodal integration. The fusion strategies successfully preserved independent feature extraction in each modality while providing advanced decision-level aggregation, effectively addressing critical limitations in current multimodal medical diagnostic approaches.

While these findings validate the effectiveness and clinical relevance of late fusion in breast cancer detection, several avenues for future research remain critical. These include scaling evaluation to larger and more diverse cohorts for enhanced generalizability; exploring novel fusion strategies such as attention-based mechanisms and transformer models; developing real-time, clinically deployable decision support systems based on multimodal deep learning; and integrating additional data types such as genetic information and longitudinal patient history to construct even more robust diagnostic frameworks. Further work should also prioritize optimizing computational efficiency for clinical adoption, incorporating explainable AI to foster clinical interpretability, and pursuing prospective clinical trials to validate real-world effectiveness, ultimately advancing the field toward more accurate, reliable, and widely applicable systems for breast cancer diagnosis.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Methodology: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Investigation: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Discussion of results: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra, Fitri Utaminingrum.
Writing – Original Draft: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Writing – Review and Editing: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Resources: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra, Fitri Utaminingrum.
Supervision: Fitri Utaminingrum.
Approval of the final text: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the AI Center of Universitas Brawijaya for their support and resources. We extend our special gratitude to the Faculty of Com-puter Science, Universitas Brawijaya, for the funding support that made this research possible.

VIII. REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA A Cancer J Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] P. D. (Yun) Trieu, C. R. Mello-Thoms, M. L. Barron, and S. J. Lewis, "Look how far we have come: BREAST cancer detection education on the international stage," *Front. Oncol.*, vol. 12, p. 1023714, Jan. 2023, doi: 10.3389/fonc.2022.1023714.

[3] M. G. Jasim and H. R. Humud, "Impact of Plasma-Activated Water on Breast Cancer Cell Viability," *Karbala International Journal of Modern Science*, vol. 11, no. 4, Sept. 2025, doi: 10.33640/2405-609X.3426.

[4] A. Mathur, N. Arya, K. Pasupa, S. Saha, S. Roy Dey, and S. Saha, "Breast cancer prognosis through the use of multi-modal classifiers: current state of the art and the way forward," *Briefings in Functional Genomics*, vol. 23, no. 5, pp. 561–569, Sept. 2024, doi: 10.1093/bfgp/elae015.

[5] M. Wielema *et al.*, "Image quality of DWI at breast MRI depends on the amount of fibroglandular tissue: implications for unenhanced screening," *Eur Radiol*, vol. 34, no. 7, pp. 4730–4737, Nov. 2023, doi: 10.1007/s00330-023-10321-y.

[6] R. K. Nisa, D. Kurniasari, F. R. Lumbanraja, and W. Warsono, "Breast Cancer Area Identification in Mammograms Using Expectation Maximization Gaussian Mixture Model," *Karbala International Journal of Modern Science*, vol. 11, no. 1, Dec. 2024, doi: 10.33640/2405-609X.3385.

[7] R. Omranipour *et al.*, "Comparison of the Accuracy of Thermography and Mammography in the Detection of Breast Cancer," *Breast Care*, vol. 11, no. 4, pp. 260–264, 2016, doi: 10.1159/000448347.

[8] F. Utaminingrum, A. W. S. B. Johan, I. K. Somawirata, T. K. Shih, and C.-Y. Lin, "Indoor staircase detection for supporting security systems in autonomous smart wheelchairs based on deep analysis of the Co-occurrence Matrix and Binary Classification," *Intelligent Systems with Applications*, vol. 23, p. 200405, Sept. 2024, doi: 10.1016/j.iswa.2024.200405.

[9] O. M. William Wolberg, "Breast Cancer Wisconsin (Diagnostic)." UCI Machine Learning Repository, 1993. doi: 10.24432/C5DW2B.

[10] S. Rodriguez-Guerrero *et al.*, "Dataset of breast thermography images for the detection of benign and malignant masses," *Data in Brief*, vol. 54, p. 110503, June 2024, doi: 10.1016/j.dib.2024.110503.

[11] "CBIS-DDSM," The Cancer Imaging Archive (TCIA). Accessed: Oct. 24, 2025. [Online]. Available: https://www.cancerimagingarchive.net/collection/cbis-ddsm/

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[13] H. Vardhini and Sophia, "Enhancing Gastrointestinal Disease Detection Through ResNet50 and Comparative Analysis of Deep Learning Architectures for Medical Image Classification," in *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India: IEEE, Feb. 2025, pp. 1768–1774. doi: 10.1109/ICEARS64219.2025.10941521.

[14] L. Singh, A. H. Wani, A. Nagasri, A. Banerjee, H. Anandaram, and B. Singh, "Multi-Class Brain Tumor Detection Using CNN-Based Medical Imaging Analysis," in *2025 3rd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India: IEEE, Mar. 2025, pp. 339–344. doi: 10.1109/ICDT63985.2025.10986709.

[15] R. Agarwal, A. Garg, A. Goel, and R. K. Bhukya, "CNN-Based Approaches for Tuberculosis Detection in Medical Imaging: A Comparative Analysis," in *2025 International Conference on Innovation in Computing and Engineering (ICE)*, Greater Noida, India: IEEE, Feb. 2025, pp. 1–6. doi: 10.1109/ICE63309.2025.10984325.

[16] B. S. P. Laksono, T. Syaifuddin, and F. Utaminingrum, "Voice Recognition to Classify 'Buka' and 'Tutup' Sound to Open and Closes Door Using Mel Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Network (CNN)," *JITeCS*, vol. 9, no. 1, pp. 58–66, Apr. 2024, doi: 10.25126/jitecs.202491579.

[17] A. Depeursinge *et al.*, "Fusing visual and clinical information for lung tissue classification in high-resolution computed tomography," *Artificial Intelligence in Medicine*, vol. 50, no. 1, pp. 13–21, Sept. 2010, doi: 10.1016/j.artmed.2010.04.006.

[18] A. Tsanousa, G. Meditskos, S. Vrochidis, and I. Kompatsiaris, "A Weighted Late Fusion Framework for Recognizing Human Activity from Wearable Sensors," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, PATRAS, Greece: IEEE, July 2019, pp. 1–8. doi: 10.1109/IISA.2019.8900725.

[19] M. Dabrowski, J. Mennesson, J. Riedi, C. Djeraba, and P. Nabat, "Knowledge-inspired fusion strategies for the inference of PM2.5 values with a Neural Network," Oct. 29, 2024, *Atmospheric sciences*. doi: 10.5194/egusphere-2024-2676.

[20] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Information Fusion*, vol. 46, pp. 147–170, Mar. 2019, doi: 10.1016/j.inffus.2018.06.002.

[21] G. Lu *et al.*, "Deep learning radiomics based on multimodal imaging for distinguishing benign and malignant breast tumours," *Front. Med.*, vol. 11, p. 1402967, July 2024, doi: 10.3389/fmed.2024.1402967.

[22] Z. Wang *et al.*, "Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction," *Precision Clinical Medicine*, vol. 7, no. 2, p. pbae012, May 2024, doi: 10.1093/pcmedi/pbae012.

[23] T. Zhang *et al.*, "Predicting breast cancer types on and beyond molecular level in a multi-modal fashion," *npj Breast Cancer*, vol. 9, no. 1, p. 16, Mar. 2023, doi: 10.1038/s41523-023-00517-2.

[24] A. A. Alsheikhy, Y. Said, T. Shawly, A. K. Alzahrani, and H. Lahza, "Biomedical Diagnosis of Breast Cancer Using Deep Learning and Multiple Classifiers," *Diagnostics*, vol. 12, no. 11, p. 2863, Nov. 2022, doi: 10.3390/diagnostics12112863.

[25] S. Zakareya, H. Izadkhah, and J. Karimpour, "A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images," *Diagnostics*, vol. 13, no. 11, p. 1944, June 2023, doi: 10.3390/diagnostics13111944.

[26] H. M. Rai, J. Yoo, S. Agarwal, and N. Agarwal, "LightweightUNet: Multimodal Deep Learning with GAN-Augmented Imaging Data for Efficient Breast Cancer Detection," *Bioengineering*, vol. 12, no. 1, p. 73, Jan. 2025, doi: 10.3390/bioengineering12010073.

[27] L. Luo *et al.*, "A large model for non-invasive and personalized management of breast cancer from multiparametric MRI," *Nat Commun*, vol. 16, no. 1, p. 3647, Apr. 2025, doi: 10.1038/s41467-025-58798-z.

[28] C. B. Rabah, A. Sattar, A. Ibrahim, and A. Serag, "A Multimodal Deep Learning Model for the Classification of Breast Cancer Subtypes".

[29] S. Devi, R. Kaul Ghanekar, J. Pande, D. Dumbre, R. Chavan, and H. Gupta, "Prediction and Diagnosis of Breast Cancer Using Machine and Modern Deep Learning Models," *Asian Pac J Cancer Prev*, vol. 25, no. 3, pp. 1077–1085, Mar. 2024, doi: 10.31557/APJCP.2024.25.3.1077.

[30] J. S. Ahn *et al.*, "Artificial Intelligence in Breast Cancer Diagnosis and Personalized Medicine," *J Breast Cancer*, vol. 26, no. 5, p. 405, 2023, doi: 10.4048/jbc.2023.26.e45.