femington

Int. J. Clin. Res. Med. Sci. (2026) 1(4):1–14
Article ID: IJCRMS-2026-013

BI-RADS-Net: Multi-Task Learning for Structured Breast Lesion Characterization and Malignancy Prediction in Ultrasound Images

Dr. Latha Kiran Krishna Rajendran
meetlathakiran@gmail.com
Independent Researcher

Abstract

Breast ultrasound imaging is widely used for breast lesion assessment; however, automated diagnostic systems often provide limited insight into the semantic characteristics underlying their predictions. This study presents BI-RADS-Net, a multi-task deep learning framework designed to jointly predict five ultrasound BI-RADS descriptors, a continuous malignancy score, and binary benign/malignant classification. The proposed architecture employs a shared VGG-16 encoder with seven task-specific prediction heads, enabling joint optimization of descriptor classification, malignancy regression, and binary classification objectives. The framework incorporates a consistency objective intended to encourage agreement between the regression and binary classification outputs. An internal evaluation was conducted using a curated collection of 1,192 breast ultrasound images aggregated from the BUSI and BUSIS repositories. In the reported internal comparison, BI-RADS-Net achieved an accuracy of 91.1%, sensitivity of 88.7%, specificity of 92.8%, and an area under the receiver operating characteristic curve (AUC) of 0.941. These results represent descriptive comparisons against two internal model variants and do not establish statistical significance. The study is limited by incomplete dataset-level records, the absence of verified patient-level data partitioning, and the lack of independent external validation. The proposed framework provides a basis for investigating structured BI-RADS descriptor prediction alongside malignancy classification, although its clinical interpretability, probability calibration, and generalizability require further validation.

Keywords

• Explainable AI • Breast Ultrasound • Breast Imaging-Reporting and Data System (BI-RADS) • Deep Learning • Multi-task Learning • Malignancy Prediction

1. Introduction and Motivation

Breast cancer remains one of the leading causes of mortality among women worldwide. Early detection plays a pivotal role in reducing the disease burden, with survival rates improving drastically when malignant tumors are identified and treated at an early stage [1]. Among various imaging modalities, ultrasound is widely used for breast assessment because it is non-ionizing, relatively accessible, and useful for characterizing breast lesions [2–5].

Despite its clinical utility, interpreting breast ultrasound (BUS) images presents significant challenges. Variations in tissue appearance, operator dependency, and subtle visual features make accurate and reproducible diagnosis difficult [6]. Automated computer-aided diagnostic (CAD) systems offer promise

© 2026 Femington Academic Publishing. All rights reserved.

in standardizing interpretation and supporting radiologists in decision-making. However, the reliability and clinical adoption of these systems heavily depend on their accuracy and explainability.

Conventional CAD frameworks have predominantly focused on binary classification or lesion segmentation tasks. These approaches often rely on black-box neural networks, providing limited insight into the reasoning behind predictions [7]. Such opacity is particularly problematic in medical applications where accountability and interpretability are essential for clinician trust and regulatory compliance.

To bridge this gap, explainable AI (XAI) methods have been explored. Visual attention maps, gradient-based saliency, and feature attribution methods have shown potential in highlighting image regions influential to predictions [8, 9]. Nevertheless, these methods often fail to align with the structured, semantic reasoning process used by clinicians in real-world diagnostic workflows.

The Breast Imaging Reporting and Data System (BI-RADS) was introduced to standardize breast imaging interpretation by defining a set of descriptors (e.g., shape, margin, orientation) and risk categories for malignancy [10, 11]. Integrating BI-RADS-based reasoning into AI models could enhance both interpretability and clinical relevance, allowing outputs to mirror radiologist decision logic.

In this context, we propose BI-RADS-Net, a multi-task deep learning framework designed to predict structured lesion characteristics alongside malignancy status. Rather than offering a single malignancy label, the model predicts five semantic BI-RADS descriptors, a continuous malignancy score, and a binary tumor classification from a shared visual encoder.

The central motivation behind BI-RADS-Net is that jointly learning multiple clinically relevant prediction tasks may encourage the shared encoder to develop representations that capture complementary characteristics of breast lesions. For example, shape and orientation descriptors may provide complementary information for malignancy classification. By optimizing these related objectives through a common feature representation, the proposed architecture investigates whether descriptor-level supervision can contribute to binary classification performance. However, the extent to which these relationships are learned and whether they correspond to clinically meaningful reasoning require further empirical evaluation.

Furthermore, to address ambiguity in diagnostic labels caused by radiologist disagreement, particularly in intermediate risk categories (e.g., BI-RADS 4A–4C), we integrate both classification and regression objectives. This hybrid modeling approach allows the system to express malignancy likelihood more flexibly than fixed-category predictions.

The framework provides structured descriptor predictions and an additional malignancy score alongside binary classification. Agreement between the regression and classification outputs may offer a quantitative measure of prediction consistency. However, output agreement should not be interpreted as a measure of model explainability or clinician confidence without independent validation. The clinical usefulness of these outputs remains to be investigated.

Overall, this study investigates a multi-task learning approach for structured breast ultrasound analysis. By combining BI-RADS descriptor prediction with malignancy regression and binary classification, BI-RADS-Net explores a unified approach to generating multiple complementary diagnostic outputs. The present work provides an internal evaluation of this architecture and identifies the additional validation required before its potential use in clinical decision-support applications can be established.

2. Literature Review and Contextual Background

2.1 Explainability in Medical Imaging AI

The rise of deep learning has transformed medical image analysis, enabling the development of data-driven approaches across multiple imaging modalities [1, 12]. However, the inherent black-box nature of deep neural networks presents challenges in clinical settings where explainability, accountability, and decision traceability are essential. To address this gap, several explainable AI (XAI) methods have been proposed.

Among the earliest strategies were saliency-based methods such as Grad-CAM [7] and Guided Backpropagation, which aim to highlight spatial regions that most influence a model's prediction. While useful for visualization, these methods often lack semantic granularity and cannot be directly mapped to clinical decision-making processes.

To bring interpretability closer to human reasoning, methods have emerged that align model outputs with structured medical reporting. For example, in thyroid imaging, the TIRADS lexicon has been used to frame model predictions around semantic descriptors like shape, echogenicity, and calcification [13]. Such lexicon-guided approaches enable models to speak the same language as radiologists, thus improving trust and integration into diagnostic workflows.

Attention-based networks have also gained traction, enabling models to focus on specific anatomical or pathological regions; transformer-based approaches have also been investigated for breast ultrasound image classification [14]. While these architectures provide insights into model focus, they do not inherently produce clinically interpretable outputs unless linked with medical ontologies or descriptors.

Another growing trend is multi-task learning (MTL), where related tasks such as classification, segmentation, and regression are learned jointly [15]. MTL not only improves generalization through shared representations but also offers an avenue to enforce consistency between model outputs and clinical logic.

It is important to distinguish between visual explanation techniques and structured semantic prediction. Saliency-based approaches identify image regions that influence model predictions, whereas descriptor-based approaches generate outputs corresponding to predefined clinical characteristics. Multi-task learning provides a mechanism for jointly optimizing these outputs but does not inherently guarantee interpretability. Consequently, the interpretability of a medical imaging model must be evaluated through the accuracy, consistency, and clinical relevance of its predictions rather than inferred solely from its architecture.

2.2 Explainable Models for Breast Ultrasound and BI-RADS

In the domain of breast imaging, prior work has explored various levels of explainability. DeepMiner [16] integrated visual explanations with shape and margin descriptors for mammographic classification, generating narrative-style justifications for predictions. However, such models were often constrained to one or two descriptors and limited to mammography rather than ultrasound.

Ultrasound-specific models have made substantial progress in recent years. For instance, automated systems have achieved high segmentation and classification accuracy on BUS datasets using convolutional and transformer-based networks [17]. Yet, few of these models provide structured, interpretable outputs aligned with the BI-RADS lexicon, which is the standard framework for radiological breast assessment.

Some recent efforts have partially addressed this. For example, Liu et al. [6] proposed a multi-task model that jointly predicts tumor class and a subset of BI-RADS features. However, their architecture was limited in scope, supporting only three descriptors and lacking a malignancy probability regression

component.

Additionally, few studies have accounted for inter-observer variability, a common challenge in breast ultrasound interpretation. Label ambiguity, especially in intermediate BI-RADS categories like 4A, 4B, and 4C, can mislead binary classifiers. Regression-based outputs that estimate malignancy probability offer a finer-grained approach to handle such variability.

Interpretability can also be enriched by assessing consistency across outputs. Some models explore intra-branch agreement or entropy to quantify output reliability. However, quantitative interpretability metrics remain underexplored in the context of multi-output breast ultrasound CAD.

Existing research has investigated breast ultrasound classification, BI-RADS descriptor prediction, and multi-task learning independently and in combination. Nevertheless, the joint formulation explored in this study, comprising five ultrasound BI-RADS descriptor predictions, continuous malignancy score estimation, and binary classification through a shared encoder, provides a specific architectural configuration for further investigation. The contribution of this configuration must be assessed through controlled comparisons with relevant existing approaches.

The principal methodological contribution of BI-RADS-Net is its joint prediction of five ultrasound BI-RADS descriptors, a continuous malignancy score, and a binary classification label using seven task-specific heads attached to a shared VGG-16 encoder. A consistency objective (Eq. (1)) is also incorporated to encourage agreement between the regression and classification outputs. The experiments presented in this study compare the proposed architecture with two internal variants and do not include a controlled empirical comparison against Liu et al. [6] or other external methods. Therefore, the findings should be interpreted as an internal evaluation of the proposed architecture rather than evidence of state-of-the-art superiority.

3. Methodology and BI-RADS-Net Architecture

3.1 BI-RADS Lexicon and Clinical Semantics

The Breast Imaging Reporting and Data System (BI-RADS) offers a standardized approach for reporting breast imaging findings [10, 11]. For ultrasound, it includes five key descriptors: *shape*, *orientation*, *margin*, *echo pattern*, and *posterior acoustic features*. These attributes provide a semantic foundation for diagnostic decision-making and structured communication of breast imaging findings [10, 18]. Integrating these descriptors into a deep learning model provides structured outputs aligned with standard radiological terminology.

In the curated dataset, each lesion is represented across these five semantic dimensions; however, the exact discrete class taxonomies, numerical label encodings, and per-class sample distributions used during model training were not retained in the available experimental records. Recovering and reporting the complete category-to-integer mappings from the original training code and annotation tables remains necessary for full reproducibility.

3.2 Model Architecture Overview

We propose BI-RADS-Net, a multi-task deep neural network that concurrently predicts:

- The five BI-RADS descriptors (as multi-class labels),
- A continuous malignancy score (as regression),
- A binary class label (benign/malignant).

The framework performs multi-task classification and regression; it does not contain a segmentation decoder, does not predict pixel-level masks, and is not trained or evaluated with a segmentation loss.

The network uses a shared visual encoder based on the VGG-16 architecture pretrained on ImageNet and fine-tuned end-to-end on three-channel 256×256 ultrasound inputs. Convolutional feature learning is widely used in medical image analysis, while shared representations provide a basis for jointly learning related prediction tasks [1, 12, 15]. From this encoder, seven task-specific fully connected prediction heads branch out: five for the BI-RADS descriptors, one for malignancy score prediction, and one for binary tumor classification. Specific internal layer widths, intermediate activations, dropout rates, and exact tensor dimensions entering the prediction heads are not documented in the available source records and require verification against the original model definition.

Task-Specific Prediction Heads

The shared VGG-16 encoder generates a feature representation that is passed to seven task-specific prediction heads. Five heads are responsible for predicting the BI-RADS descriptors, with each head producing a probability distribution over the corresponding descriptor categories. The sixth head estimates a continuous malignancy score, while the seventh head generates a binary malignancy prediction probability.

The descriptor heads are optimized using categorical classification objectives, whereas the malignancy score and binary classification heads employ regression and binary classification objectives, respectively. All seven heads receive the shared encoder representation, allowing the model to jointly optimize the associated prediction tasks.

3.3 Mathematical Objective

The training objective combines classification and regression losses for each output, following the general multi-task learning principle of optimizing related tasks through shared representations [6, 15]. Let y_i and $\hat{y}_i$ denote ground truth and predicted values for task i, and $\mathcal{L}_i$ the loss function. The reported total loss formulation is defined as:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{7} \lambda_i \cdot \mathcal{L}_i(\hat{y}_i, y_i) + \lambda_a \cdot \mathcal{L}_a\left(|\hat{y}_6 - \hat{y}_7|, |y_6 - y_7|\right) \tag{1}$$

where the task indices are: $i = 1, \ldots, 5$ correspond to the five BI-RADS descriptor heads (shape, orientation, margin, echo pattern, posterior features), $i = 6$ to the malignancy regression head, and $i = 7$ to the binary classification head. The quantities $\hat{y}_6$ and $\hat{y}_7$ in the consistency term are intended to represent continuous probabilities in the interval $[0, 1]$. In particular, $\hat{y}_7$ denotes the sigmoid probability produced by the binary classification head before any decision threshold is applied, whereas y_6 and y_7 denote the mapped malignancy score and binary target, respectively.

Each loss component in Eq. (1) is defined as follows. For the five descriptor heads ($i = 1, \ldots, 5$), $\mathcal{L}_i$ is the categorical cross-entropy between the predicted descriptor distribution $\hat{y}_i$ and the ground-truth descriptor label y_i. For the malignancy regression head ($i = 6$), $\mathcal{L}_6$ is the mean squared error between the predicted malignancy score $\hat{y}_6$ and the mapped ground-truth target y_6 defined in Section 4.3. For the binary classification head ($i = 7$), $\mathcal{L}_7$ is the binary cross-entropy between $\hat{y}_7$ and the ground-truth binary label y_7.

Note that comparing the predicted absolute difference $|\hat{y}_6 - \hat{y}_7|$ against the target difference $|y_6 - y_7|$ penalizes deviation from the discrepancy between the heuristic BI-RADS target and the binary label

rather than directly enforcing identical predictions $\hat{y}_6 \approx \hat{y}_7$. Because the original training script was not supplied with this manuscript, the exact functional form of $\mathcal{L}_a$ and the use of pre-threshold sigmoid outputs in the implemented loss calculation require verification against the original code.

3.4 BI-RADS-Net Architecture Diagram

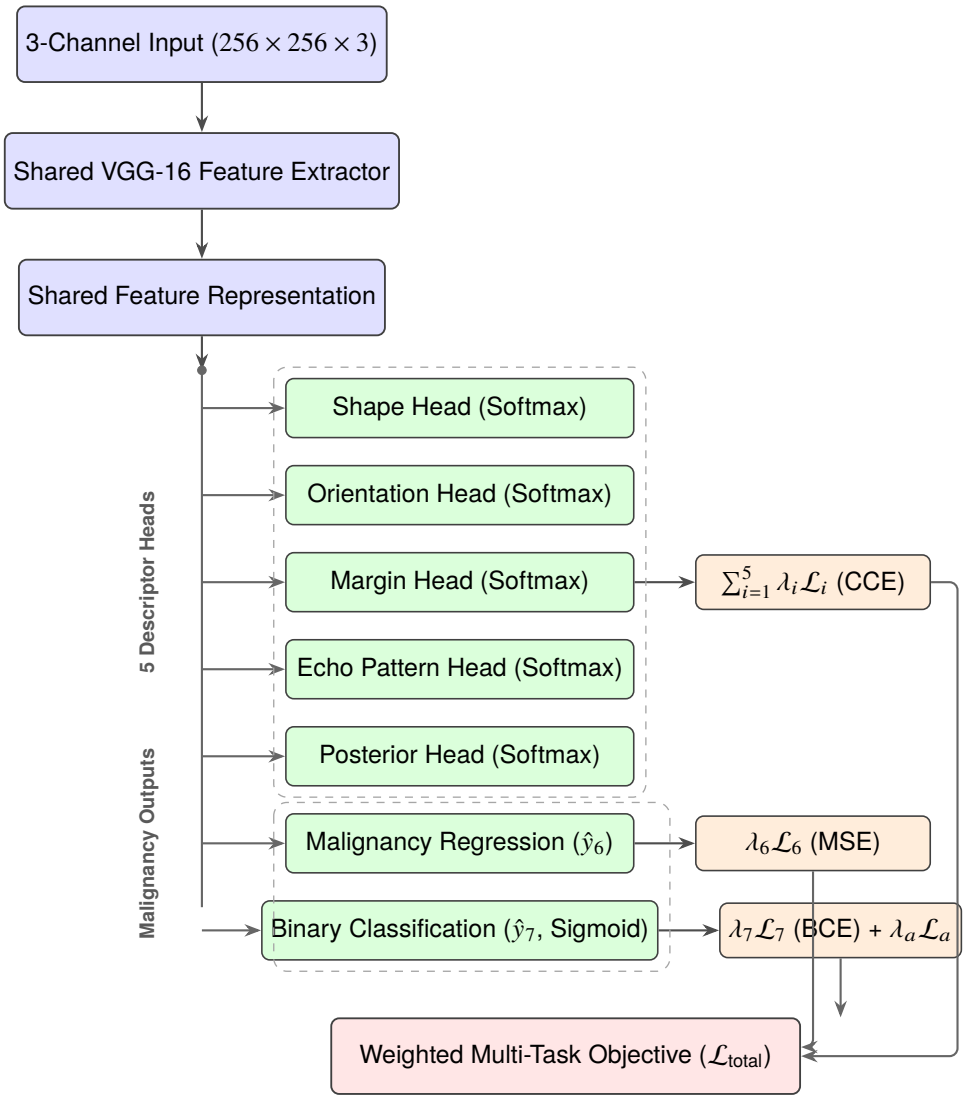


Figure 1: Schematic of BI-RADS-Net showing the three-channel 256×256 input, shared VGG-16 feature extractor, five semantic descriptor heads, distinct malignancy regression ($\hat{y}_6$) and binary classification ($\hat{y}_7$) heads, and weighted multi-task training objective.

The network implementation corresponding to Figure 1 uses a VGG-16 backbone pretrained on ImageNet and fine-tuned end-to-end. Each of the seven task-specific heads is a fully connected branch attached to the shared encoder representation: five descriptor heads produce multi-class probability distributions over the BI-RADS descriptor categories, one regression head outputs a continuous malignancy score ($\hat{y}_6$), and one binary classification head outputs a sigmoid malignancy probability ($\hat{y}_7$). Training uses the Adam optimizer, batch size 16, and the multi-task loss defined in Eq. (1) with the weights in Eq. (2).

4. Experimental Design and Dataset Overview

4.1 Datasets and Annotations

To evaluate the internal performance of the proposed BI-RADS-Net framework, we utilized a curated dataset of 1,192 breast ultrasound (BUS) images aggregated from two publicly available repositories: the Breast Ultrasound Dataset (BUSI) and BUSIS. Reviews of breast ultrasound and automated breast ultrasound describe the modality's clinical use and the variability encountered across imaging settings [2–5]. External generalizability was not evaluated in this study. Each image in the dataset is annotated with a binary malignancy label (benign or malignant), corresponding BI-RADS assessment category, and five semantic descriptors: shape, margin, orientation, echo pattern, and posterior features.

The BUSI dataset includes ultrasound images labeled by expert radiologists, capturing a range of tumor morphologies. BUSIS adds images collected from multiple sources. Although this may introduce variation in acquisition settings, the present study does not include repository-specific or independent external evaluation and therefore does not establish robustness to domain shift or real-world deployment.

Table 1 summarizes the reported dataset composition. The BUSI benign and malignant counts sum to 647 rather than the reported 780 images, leaving 133 images unaccounted for by the supplied class counts (such as normal scans or excluded cases in the original repository). This arithmetic remainder is not evidence of a verified third class or exclusion category without source verification. The BUSIS class distribution, duplicate screening counts, and combined benign/malignant totals cannot be determined from the available dataset records.

Table 1: Composition of the curated breast ultrasound dataset (1,192 images).

Source	Total Images	Benign	Malignant	Unaccounted Remainder
BUSI	780	437	210	133
BUSIS	412	Not reported	Not reported	Not determinable
Total	1,192	Not determinable	Not determinable	Not determinable

4.2 Annotation Protocol

The binary malignancy labels and BI-RADS classifications originate from the public BUSI and BUSIS repositories. However, the complete annotation protocol for the five semantic BI-RADS descriptors—including annotator qualifications, number of independent readers, annotation guidelines, disagreement-resolution rules, and inter-observer agreement statistics—is not documented in the supplied records. Distinguishing repository-provided labels from study-specific descriptor annotations and verifying image-level exclusions remain necessary in future dataset documentation.

Furthermore, images were split at the image level because patient identifiers were unavailable in the curated dataset. Consequently, multiple images acquired from the same patient may appear across different partitions, introducing a potential risk of data leakage. Recovering patient identifiers and performing strict patient-grouped cross-validation are required before claiming patient-independent generalization.

4.3 Data Preprocessing Pipeline

To standardize inputs across sources, each BUS image was center-cropped and resized to 256×256 pixels. Standardized preprocessing is commonly used in medical image analysis to support consistent model

inputs [1, 17]. Rather than relying solely on grayscale intensity, we constructed three input channels:

- Channel 1: Raw grayscale image.
- Channel 2: Histogram-equalized image to enhance local contrast.
- Channel 3: Gaussian-smoothed image to reduce high-frequency speckle variations.

Because fixed center-cropping is applied without automated lesion detection or ground-truth bounding boxes, retention of the full lesion margin cannot be guaranteed for eccentric tumors. Additionally, exact source image dimensions, resizing interpolation kernels, Gaussian filter parameters, and intensity normalization ranges are not recorded in the available implementation notes.

The three-channel representation provides complementary intensity transformations of the original ultrasound image. The raw grayscale channel preserves the original intensity information, histogram equalization enhances local contrast, and Gaussian smoothing reduces high-frequency image variations. The contribution of these transformations to classification performance should be assessed through dedicated preprocessing ablation experiments.

4.4 Label Transformation and Loss Supervision

The malignancy regression head was supervised using study-specific continuous malignancy scores derived from BI-RADS assessment categories [10, 18]. Table 2 summarizes the numerical scores assigned to each BI-RADS category based on American College of Radiology (ACR) risk reference ranges and heuristic midpoints.

Table 2: Study-specific BI-RADS malignancy score mapping.

BI-RADS Category	Assigned Score
3	0.010
4A	0.060
4B	0.300
4C	0.725
5	0.975

These point values represent heuristic approximations rather than empirically calibrated patient-level probabilities. Handling of non-included BI-RADS categories (e.g., categories 0, 1, 2, or 6), category sample counts across BUSI and BUSIS, and potential discrepancies between BI-RADS-derived regression targets and biopsy-confirmed binary labels are not documented in the supplied records.

Losses for descriptor classification heads used categorical cross-entropy, the binary classification head used binary cross-entropy, and the malignancy regression head employed mean squared error (MSE). To encourage alignment between outputs, we included the consistency term in Eq. (1).

4.5 Data Augmentation Strategy

To introduce spatial variation during training, online data augmentation was applied using random perturbations [1, 12]:

- Zooming (up to 20%).
- Horizontal flipping.

- Width shifting (up to 10%).
- Rotation (up to 5 degrees).

These transformations introduce controlled spatial variations during training. Their suitability for breast ultrasound imaging should be assessed to ensure that clinically relevant lesion characteristics are preserved, particularly for orientation-sensitive descriptors. Exact augmentation implementation details, including confirmation that transformations were restricted strictly to the training partition and interpolation fill modes, are not retained in the available experimental logs.

4.6 Training Procedure and Hyperparameters

The model was evaluated using a 5-fold cross-validation scheme in which each iteration allocated 70% of samples to training, 15% to internal validation, and 15% to testing. The use of shared representations for related prediction objectives is consistent with established multi-task learning formulations [15]. Because exact fold-level sample indices, random seeds, stratification rules, and test-set overlap records were not retained, non-overlapping test coverage across all five folds cannot be independently verified from the available records.

The Adam optimizer was used with an initial learning rate of 1×10^{-5}, which decayed to 1×10^{-6} upon stagnation of validation performance. Training was conducted for up to 150 epochs with a batch size of 16, using early stopping triggered after 30 consecutive epochs without validation improvement. Task-specific loss weights were heuristically set as:

$$\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 0.2, \quad \lambda_6 = 0.2, \quad \lambda_7 = 0.5, \quad \lambda_a = 0.2 \tag{2}$$

where λ_1–λ_5 weight the five descriptor heads, λ_6 weights the malignancy regression head, λ_7 weights the binary classification head, and λ_a weights the consistency term in Eq. (1).

4.7 Implementation Details

All experiments were implemented in Python using TensorFlow on an NVIDIA RTX A6000 GPU. The VGG-16 backbone was initialized with ImageNet weights and fine-tuned during training with a batch size of 16. Exact software package versions (Python, TensorFlow, CUDA, and image processing libraries), checkpoint selection criteria, and runtime measurements were not logged in the supplied records.

The implementation configuration establishes the computational environment used for the internal evaluation. However, robustness to differences in acquisition devices and imaging environments requires additional evaluation using independent datasets.

5. Results, Evaluation, and Interpretation

5.1 Evaluation Metrics

Evaluation was conducted separately for the three prediction components of BI-RADS-Net. Binary malignancy classification was assessed using threshold-dependent classification metrics (accuracy, sensitivity, specificity, and F1-score) and ROC-AUC [1]. Descriptor prediction was evaluated independently for each BI-RADS attribute to account for differences in class distribution. The malignancy regression output was assessed using mean squared error (MSE) and Pearson correlation with the mapped reference scores. Consistent with Section 3.2, all evaluation is at the image level; no pixel-level segmentation metric or

probability calibration metric (such as Brier score or reliability diagrams) is reported. These evaluation procedures distinguish classification performance from regression agreement and avoid interpreting correlation alone as evidence of probability calibration.

5.2 Ablation Study: Contribution of Multi-Task Learning

To evaluate the impact of multi-task learning, we conducted an internal ablation study comparing three configurations:

- **Baseline-CNN**: VGG-16 encoder with only the binary classification head.
- **BI-RADS-Net w/o Regression**: Shared VGG-16 encoder with the five descriptor heads and binary classification head, excluding the malignancy regression head.
- **BI-RADS-Net Full**: Full seven-head multi-task architecture incorporating descriptor classification, malignancy regression, binary classification, and the consistency term.

Table 3: Reported performance comparison on the internal BUS test set (descriptive metrics).

Model	Accuracy	Sensitivity	Specificity	AUC
Baseline-CNN	0.849	0.806	0.891	0.883
BI-RADS-Net w/o Regression	0.879	0.843	0.905	0.907
BI-RADS-Net Full	**0.911**	**0.887**	**0.928**	**0.941**

The fold-level predictions and metric values needed to calculate five-fold standard deviations and 95% confidence intervals are unavailable. Accordingly, Table 3 reports only the descriptive aggregate performance values available in the supplied records; no standard deviations, confidence intervals, or confusion matrices are claimed. The reported values in Table 3 show higher accuracy, sensitivity, specificity, and AUC for the full BI-RADS-Net than for the two internal comparison variants [6, 15], but should be interpreted descriptively rather than as evidence of statistically significant differences.

5.3 Statistical Analysis

Pairwise statistical significance testing between Baseline-CNN, BI-RADS-Net w/o Regression, and BI-RADS-Net Full (such as DeLong's test for correlated ROC curves, McNemar's test, or paired bootstrap resampling accounting for cross-validation dependence) was not performed because per-sample fold predictions were not retained. Consequently, no inferential statistical claims are made regarding the observed performance differences. External validation on an independent dataset and controlled empirical comparisons against external breast ultrasound architectures on identical splits were likewise not conducted.

5.4 Descriptor-wise Impact on Prediction

To examine the contribution of individual BI-RADS descriptors, one descriptor head at a time was removed during training and the resulting reduction in AUC relative to the full-model baseline (0.941) was recorded. Figure 2 summarizes the reported reductions.

In the reported descriptor-removal comparison, removing Shape and Margin corresponded to the largest observed AUC reductions (5.4 and 4.1 percentage points, respectively), followed by Echo Pattern

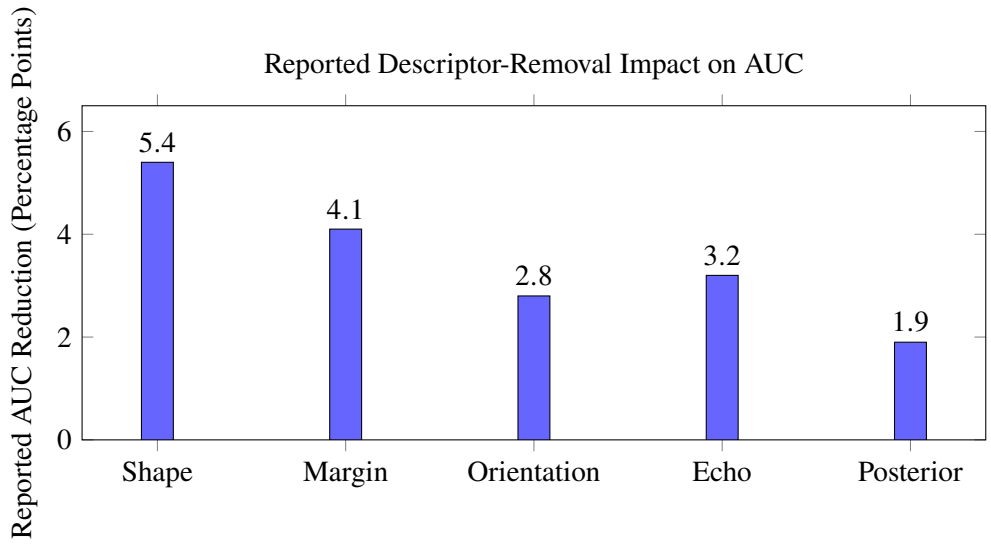


Figure 2: Reported reduction in AUC relative to the full model (0.941) when individual BI-RADS descriptors are removed from training.

(3.2), Orientation (2.8), and Posterior Features (1.9). The underlying fold-level logs, exact ablated AUC values, and loss re-weighting details required to independently verify these values are not included in the available experimental records.

5.5 Malignancy Regression Alignment

The regression head estimates a continuous malignancy score based on the study-specific mapping of BI-RADS assessment categories. Its relationship with the mapped reference scores is evaluated using Pearson correlation. The reported correlation of $r = 0.81$ indicates a positive linear association between the predicted and reference scores, while reported agreement between regression and binary classification outputs exceeded 92% in the low-risk (BI-RADS 3) and high-risk (BI-RADS 5) categories.

However, correlation and category-extreme agreement do not establish probability calibration or clinical accuracy of the estimated scores. Category-specific sample counts, exact agreement definitions, and fold-level dispersion metrics were not retained in the available records. Further calibration analysis and validation against biopsy-confirmed outcomes are required to establish the clinical reliability of the regression predictions.

5.6 Structured Prediction Outputs and Visual Analysis

No representative ultrasound images, ground-truth lesion annotations, model prediction visualizations, or saliency heatmaps generated by the implemented model are available in the supplied materials. Consequently, this study does not provide qualitative visual evidence of the model's spatial focus, and no claim that saliency maps identify tumor boundaries or acoustic shadow features is made. The BI-RADS descriptor outputs provide structured predictions, but their clinical interpretability and usefulness have not been evaluated through a radiologist study.

6. Discussion, Limitations, and Future Work

The internal evaluation indicates that the full BI-RADS-Net configuration achieved higher reported classification metrics than the single-task baseline and the reduced multi-task variant. Prior work has explored deep learning and AI-assisted analysis in breast ultrasound, while reviews also emphasize the need for careful validation before clinical translation [2, 4, 5]. These observations suggest that jointly optimizing descriptor prediction, malignancy regression, and binary classification may be beneficial within the evaluated experimental setting. However, the absence of fold-level uncertainty estimates and statistical comparisons prevents conclusions regarding the significance of these differences. Further experiments are required to establish whether the observed improvements are reproducible.

BI-RADS-Net generates multiple structured outputs corresponding to breast ultrasound descriptors, malignancy score estimation, and binary classification [10, 11, 19]. These outputs provide more detailed model predictions than a single binary label. Nevertheless, structured outputs alone do not establish clinical interpretability. Their accuracy, consistency with radiologist assessments, and usefulness in clinical decision-making require evaluation through independent annotation studies, calibration analyses, and radiologist-based assessments.

Although the curated dataset combines images from two public repositories, repository-specific evaluation and independent external validation were not performed. Consequently, the reported findings cannot establish robustness to differences in imaging equipment, acquisition protocols, patient populations, or annotation practices. Future evaluation should include training on one repository and testing on an independent repository, together with additional validation using institutionally independent data.

Several methodological and evidentiary limitations constrain the interpretation of this study:

1. **Static 2D imaging and missing clinical metadata:** The model operates exclusively on static 2D ultrasound images without dynamic cine loops, multi-plane context, breast composition, or patient metadata (e.g., age, clinical history, or genetic risk).

2. **Incomplete dataset and reproducibility records:** Complete class breakdowns for BUSIS, verification of the 133-image remainder in BUSI, descriptor annotation protocols, inter-observer agreement metrics, and full software/architecture logs were not retained in the available records.

3. **Heuristic BI-RADS probability mappings:** The continuous regression targets rely on study-specific point approximations from BI-RADS categories rather than biopsy-confirmed empirical malignancy rates [10, 18], and formal probability calibration was not performed.

4. **Potential patient-level data leakage and missing statistical uncertainty:** Because patient identifiers were unavailable, images were split at the image level, creating a risk of patient overlap across folds. Moreover, fold-level dispersion metrics, confidence intervals, and paired statistical significance tests were not available.

5. **Absence of external validation, visual analysis, and clinical evaluation:** Repository-specific testing, independent external cohort validation, model-generated visual explanations, prospective radiologist reader studies, and real-time PACS workflow integration remain to be conducted.

Future extensions of this work will explore:

- Patient-level cross-validation with complete dataset reconciliation, fold-level statistical uncertainty reporting, and reproducible open-source documentation.

- External validation across independent breast ultrasound repositories and clinical institutions.
- Probability calibration and regression training against biopsy-confirmed malignancy outcomes.
- Prospective assessment of structured descriptor and risk predictions with qualified breast radiologists.
- Integration of temporal cine-loop ultrasound data, 3D imaging modalities, and patient clinical metadata, as well as federated training across institutions to expand generalizability while preserving data privacy.

References

[1] G. Litjens, T. Kooi, and B.E. et al. Bejnordi. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[2] I. Boca, A. I. Ciurea, C. A. Ciortea, and S. M. Dudea. Pros and cons for automated breast ultrasound (abus): A narrative review. *Journal of Personalized Medicine*, 11(8):703, 2021.

[3] K. Malherbe and D. Tafti. Breast ultrasound. In *StatPearls [Internet]*. StatPearls Publishing, 2024.

[4] H. F. Brot and V. L. Mango. Artificial intelligence in breast ultrasound: Application in clinical practice. *Ultrasonography*, 43(1):3–14, 2024.

[5] R. Iacob, E. R. Iacob, E. R. Stoicescu, D. M. Ghenciu, D. M. Cocolea, A. Constantinescu, et al. Evaluating the role of breast ultrasound in early detection of breast cancer in low-and middle-income countries: A comprehensive narrative review. *Bioengineering*, 11(3):262, 2024.

[6] Y. Liu, Y. Zhang, and W. et al. Zhang. Deep multi-task learning for joint prediction of breast cancer diagnosis and bi-rads descriptors in ultrasound images. *In Proc. MICCAI*, 2020.

[7] R.R. Selvaraju, M. Cogswell, and A. et al. Das. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE ICCV*, pages 618–626, 2017.

[8] M.D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV*, pages 818–833, 2014.

[9] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proc. ICML*, 2017.

[10] M. Eghtedari, A. Chong, R. Rakow-Penner, and H. Ojeda-Fournier. Current status and future of BI-RADS in multimodality imaging, from the AJR special series on radiology reporting and data systems. *American Journal of Roentgenology*, 216(4):860–873, 2021.

[11] H. Barazi and M. Gunduru. Mammography BI-RADS grading. In *StatPearls*. StatPearls Publishing, 2023.

[12] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[13] Y. Pan, H. Liu, and L. Yang. Ti-rads guided deep learning for thyroid ultrasound diagnosis. In *Proceedings of MICCAI*, 2021.

[14] B. Gheflati and H. Rivaz. Vision transformers for classification of breast ultrasound images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 480–483. IEEE, July 2022.

[15] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[16] E. Tsai, J. Wu, S. Lin, and W. Hsu. Deepminer: Discovering interpretable representations for mammogram classification and explanation. In *MICCAI*, 2019.

[17] Z. Yu, Y. Zhang, and Y. et al. Wang. Automated breast ultrasound image segmentation using deep learning. *Physics in Medicine and Biology*, 64(13), 2019.

[18] S. Fazeli, J. Stepenosky, M. S. Guirguis, B. Adrada, R. Rakow-Penner, and H. Ojeda-Fournier. Understanding BI-RADS category 3. *RadioGraphics*, 45(1):e240169, 2024.

[19] A. Cozzi, K. Pinker, A. Hidber, T. Zhang, L. Bonomo, R. Lo Gullo, et al. BI-RADS category assignments by GPT-3.5, GPT-4, and Google Bard: A multilanguage study. *Radiology*, 311(1): e232133, 2024.