International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(17s) 2026 pp. 958-1002
https://doi.org/10.70917/ijcisim-2026-2518
Article

A Hybrid Multimodal AI Framework for BI-RADS Breast Cancer Risk Prediction and Report Generation: Radiologist-Mimetic Cognitive Knowledge Graph Injection and Cognitive Prompt Engineering

Ashok Kumar Sahoo[1*], Rajkumar Rajavel[2]

[1]School of Computer Science and Engineering, Christ University, Bangalore, India
[2]Department of AI, ML and Data Science, Christ University, Bangalore, India
*Corresponding author E-mail: ashok.sahoo123@gmail.com

Abstract: Automated BI-RADS breast cancer risk stratification from multimodal clinical data remains limited by two persistent gaps: purely discriminative architectures cannot resolve fine-grained BI-RADS 4 subcategory boundaries with calibrated uncertainty, and no existing mechanism translates multimodal evidence into reasoning-transparent structured radiology reports. This study presents a hybrid discriminative-generative framework integrating deep learning encoders for mammography, ultrasound, and DCE-MRI; XGBoost-Random Forest ensembles for structured clinical, pathological, and environmental data; and ClinicalBERT for free-text narratives, unified through Transformer cross-modal attention. Two novelties ground the framework in radiologist cognitive reasoning: the Radiologist-Mimetic Cognitive Knowledge Graph (RM-CKG), an OWL 2 DL, three-round expert-validated ontology encoded by a Graph Attention Network and injected as a patient-specific cognitive prior at the fusion layer; and Radiologist-Mimetic Cognitive Prompting (RMCP), which serializes the RM-CKG subgraph traversal into a five-layer reasoning narrative conditioning a LoRA fine-tuned vision-language model for structured report generation. Evaluated on 11,847 patients, the framework achieves a macro-averaged AUC of 0.93 and QWK of 0.84, with the RM-CKG producing its largest isolable gain at the BI-RADS 4A/4B subcategory boundary (ΔAUC = +0.117); RMCP-conditioned reports achieve a blinded radiologist Clinical Accuracy Score of 4.31/5.00 and a BI-RADS Consistency Rate of 0.96.
Keywords: BI-RADS classification, Multimodal deep learning, Knowledge graph, Radiology report generation, Uncertainty quantification, OWL 2 DL ontology, Graph attention network, Cognitive prompting, Large language model

1. Introduction

Breast cancer remains one of the most prevalent and life-threatening malignancies among women globally, with the WHO reporting over two million new cases annually [1]. Early detection is the strongest survival determinant: the American Cancer Society reports survival exceeding 90% at stage I versus markedly lower rates at advanced stages [2]. Despite established modalities — mammography, DBT, MRI, ultrasound, biopsy — clinical effectiveness remains constrained by subjective human interpretation, as per Aristokli et al., 2022 [3]. Elmore et al., 2015 [4] found pathologist disagreement against a consensus panel reaches 25% for complex, dense-tissue cases, carrying direct implications for outcomes where accurate classification governs intervention urgency.

Reliance on manual review carries systemic costs beyond individual errors. McKinney et al., 2020 [5] documented workload burdens and error rates in mammography screening, noting that rising volume and radiologist

shortages create bottlenecks delaying diagnosis and driving clinical burnout that further reduces throughput. These pressures, combined with the inter-observer variability Elmore et al. [4] quantified, elevate both false positives — unnecessary invasive procedures — and false negatives, where malignancy progresses undetected.

Two broad remedies address inter-observer variability and screening bottlenecks. The first is procedural: mandatory double-reading, as in the UK NHS Breast Screening Programme, where a second radiologist reviews every mammogram. This reduces false negatives, but Elmore et al., 2015 [4] found pathologist disagreement on the same specimens — especially atypia and ductal carcinoma in situ — remains substantial even under structured consensus, concordance falling precisely where accuracy matters most. The second, increasingly dominant remedy is AI-based image analysis. McKinney et al., 2020 [5] showed a deep learning system trained on UK and USA mammograms surpassed six radiologists' average performance by an 11.5% AUC-ROC margin, cutting false positives by 5.7% and false negatives by 9.4% in the USA cohort, and reduced second-reader workload by 88% when deployed in the UK's double-reading workflow. However, both remedies operate exclusively on imaging data, excluding non-image context — demographics, family history, hormonal factors, biopsy reports, prior screening records — that clinicians routinely integrate. Neither incorporates a structured BI-RADS risk framework synthesizing multimodal data or leverages LLMs for free-text narratives. This gap motivates the present research: a hybrid AI framework fusing imaging and non-imaging data through classical machine learning, deep learning, and large language models to deliver a more comprehensive, clinically grounded breast cancer risk assessment.

Researchers have pursued classical machine learning, deep learning, and generative AI to address diagnostic inconsistency and screening inefficiency. Classical methods — SVMs, decision trees, k-NN — relied on hand-engineered features and offered modest interpretability but limited generalization, as per Tan et al., 2023 [6]. Research subsequently shifted to convolutional architectures applied to mammographic, ultrasound, and histopathological imaging. BI-RADS, the ACR's standardized lexicon, categorizes findings into seven risk categories (0–6), with category 4 subdivided into 4A (2–10%), 4B (10–50%), and 4C (50–95%) — a clinically ambiguous spectrum prone to unnecessary biopsies or missed malignancies. Tan et al., 2023 [6] demonstrated a multi-view CNN on mammograms from 12,433 Asian women achieving AUC up to 0.995 and improving inter-reader agreement from 0.629 to 0.672, correctly reclassifying 83.1% of false-positive BI-RADS 4B/4C findings as benign, though the system used single-vendor, single-ethnicity imaging without clinical variables. Parallel work in LLMs has shown models such as GPT-4 can improve BI-RADS classification consistency from free-text reports and generate meaningful interpretations from structured records, yet no published framework fuses LLM-based clinical reasoning with image-derived deep learning features under a unified BI-RADS architecture. A 2025 systematic review of 49 multimodal AI studies (2015–2025) confirmed models integrating imaging with clinical, pathological, and genomic data consistently outperform unimodal counterparts, identifying standardized BI-RADS-aligned fusion as a critical unresolved gap — precisely where the present research intervenes, mitigating inter-observer variability by anchoring predictions to an objective risk scale and reducing screening bottlenecks through automated multimodal triage independent of any single channel.

Against this backdrop — fragmented AI approaches, absent unified calibration, opaque reasoning, and the gap between risk scoring and actionable reporting — this work proposes a hybrid AI framework integrating multimodal patient data through three complementary AI paradigms within a single end-to-end system. First, we present a unified discriminative pipeline assigning each modality to its architecturally best-suited paradigm: deep learning encoders for mammography, ultrasound, and DCE-MRI; gradient-boosted tree ensembles and random forests for structured clinical, pathological, and environmental data; and a domain-specialized LLM for free-text narratives, fused via Transformer cross-modal attention and calibrated against the ACR BI-RADS 5th edition schema including 4A/4B/4C subcategories. Second, we introduce the Radiologist-Mimetic Cognitive Knowledge Graph (RM-CKG), an ontology-grounded, expert-validated graph whose node taxonomy, edge vocabulary, and inference rules are formally defined by a multi-domain OWL 2 DL ontology and independently validated by a senior breast imaging radiologist prior to data-driven population; a patient-specific subgraph is retrieved via FAISS approximate nearest-neighbor search, encoded by a three-layer Graph Attention Network, and injected as a cognitive prior at the fusion layer, permeating the fusion process rather than acting as post-hoc correction. Third, we propose Radiologist-Mimetic Cognitive

Prompting (RMCP), which serializes the patient-specific RM-CKG subgraph traversal into a five-layer cognitive narrative — observational grounding, evidential interpretation, analogical recall, uncertainty flagging, decision synthesis — injected as inference context for a fine-tuned vision-language model, producing, for the first time in a unified framework, both a calibrated ordinal risk score and a clinically interpretable report grounded by the same knowledge graph. The framework further incorporates learnable missing-modality tokens with stochastic dropout for robustness under incomplete modality availability, and provides multi-level explainability through Grad-CAM, SHAP, and GAT attention visualization. The remainder of this paper proceeds as follows: Section 2 surveys related literature; Section 3 presents the framework methodology; Section 4 describes datasets and evaluation protocol; Section 5 reports results with ablation analysis; Section 6 discusses clinical implications and limitations; Section 7 concludes with future directions.

2. Related Work

2.1 Breast Cancer Screening, Inter-Observer Variability, and the BI-RADS Standardization Framework

Breast cancer accounts for the highest female malignancy incidence globally and the second-largest share of cancer mortality, with population-level reduction tied to screening sensitivity and specificity. Elmore et al., 2015 [4], in a study of 115 radiologists interpreting 240 mammograms, found recall rates spanning 3%-61% and a ten-year cumulative false-positive-recall probability of 61% for women without prior biopsies. Sprague et al., 2014 [7] showed this variability concentrates at boundary categories - particularly BI-RADS 3's narrow 2% threshold - with institutional volume and subspecialty training explaining only a fraction of inter-reader disagreement. The ACR BI-RADS 5th edition [8] defines seven categories with linked management recommendations, yet its shared vocabulary alone does not resolve the disagreement Elmore [4] and Sprague [7] documented; Berg et al., 2012 [9] further showed supplemental ultrasound increases cancer detection but also the false-positive biopsy rate, underscoring single-modality inadequacy in complex presentations.

2.2 Deep Learning for Breast Image Analysis

He et al., 2016 [10] introduced deep residual learning, resolving the training instability of earlier deep architectures and establishing the template adopted across breast imaging classification. McKinney et al., 2020 [5] applied a large CNN to 76,000 UK and 15,000 US mammograms, reducing false positives by 5.7% and false negatives by 9.4% relative to individual radiologists. Wu et al., 2019 [11] extended this via a dual-view cross-attention architecture whose inter-view consistency proved more predictive than either single view - motivating the dual-stream mammography branch adopted here. For ultrasound, Tan and Le, 2019 [12] introduced EfficientNet's compound scaling, and Shen et al., 2021 [13] applied it to BUSI/UDIAT lesion classification with sensitivity/specificity exceeding earlier CNNs and human readers. For DCE-MRI, Kuhl et al., 2005 [14] established 91% sensitivity for invasive cancers versus 33% (mammography) and 37% (ultrasound), identifying kinetic curve morphology as decisive; Truhn et al., 2019 [15] recovered these features via 3D CNNs at near-radiologist performance, and Antropova et al., 2017 [16] quantified washout curve morphology's independent prognostic contribution.

A limitation common to single-modality systems, per Litjens et al., 2017 [17], is distributional specificity: site- and scanner-specific artefacts transfer poorly across acquisition protocols. Mei et al., 2022 [18] addressed this through RadImageNet, a 1.35-million-image pre-training corpus improving cross-site transfer, adopted here as encoder initialization. Yala et al., 2019 [19] showed mammography-only models plateau below the accuracy achievable when imaging is combined with clinical risk factors - motivating the structured non-image branch.

2.3 Classical Machine Learning for Structured Clinical Risk Stratification

Structured risk modelling predates deep learning by decades. The Gail model [20] introduced logistic regression over age, reproductive and family history, and prior biopsy findings; the Tyrer-Cuzick model [21] added BRCA carrier probability and hormonal history within a Bayesian framework; BOADICEA [22] incorporated polygenic risk scores

via segregation analysis. Each remains fixed to its original risk-factor set and cannot integrate real-time laboratory, imaging-derived, or pathological data as dynamic inputs.

Chen and Guestrin, 2016 [23] introduced XGBoost, achieving top single-algorithm performance on structured clinical benchmarks; Breiman, 2001 [24] established Random Forest's variance-reduction properties, theoretically complementary to gradient boosting. Grinsztajn et al., 2022 [25], across 45 tabular datasets, showed tree-based ensembles consistently outperform deep networks - including FT-Transformer and TabNet - on medium-scale, heterogeneous tabular data; Shwartz-Ziv and Armon, 2022 [26] attributed this to neural networks' rotational invariance versus trees' axis-aligned splits respecting clinical variable semantics - justifying the exclusion of deep architectures from the non-image branch here.

2.4 Large Language Models for Clinical Text Understanding

Devlin et al., 2019 [27] introduced BERT's bidirectional pre-training, establishing the pre-train-then-fine-tune paradigm. Alsentzer et al., 2019 [28] showed domain-specific pre-training (PubMed, MIMIC-III) substantially outperforms general BERT on clinical NLP, with MIMIC-III-trained ClinicalBERT benefiting from exposure to clinical abbreviation, negation, and temporal patterns.

Gu et al., 2021 [29] showed PubMedBERT, pre-trained from scratch on biomedical text, outperforms domain-adaptive fine-tuning of general models. Singhal et al., 2023 [30] showed Med-PaLM-scale instruction-tuned LLMs reach passing USMLE performance, though factual precision on sub-specialist questions remains the primary gap - motivating the domain-specific ClinicalBERT fine-tuning adopted for the free-text branch here.

2.5 Multimodal Data Fusion in Medical Artificial Intelligence

Baltrušaitis et al., 2019 [31] provided the canonical multimodal fusion taxonomy - early, late, and intermediate fusion - showing intermediate fusion outperforms both when modalities are complementary rather than redundant. Vaswani et al., 2017 [32] established Transformer multi-head self-attention as an efficient mechanism for pairwise dependency modelling, architecturally suited to cross-modal fusion when modality representations form a token sequence.

Acosta et al., 2022 [33] documented Transformer cross-attention as the standard for intermediate biomedical fusion; Huang et al., 2021 [34] showed co-attention fusion of chest imaging and EHR data improves mortality prediction by suppressing uninformative modalities per patient. Ramachandram and Taylor, 2017 [35] identified a shared limitation: existing fusion incorporates no prior clinical knowledge of inter-modality relationships - a gap the RM-CKG addresses directly by supplying a patient-specific, ontology-grounded evidential prior alongside data-driven attention.

Ma et al., 2021 [36] proposed learnable modality-specific sentinel tokens for missing-modality robustness, showing sentinel substitution outperforms imputation or subnetwork routing. The present framework extends this with stochastic modality masking during joint training, optimizing directly for deployment-time missing-modality conditions.

2.6 Knowledge Graphs in Clinical Decision Support

Rotmensch et al., 2017 [37] automatically constructed a clinical knowledge graph from 270,000 EHRs, recovering 4.5 million patient-symptom relationships at 85% UMLS-validated precision. Chandak et al., 2023 [38] built a precision-medicine graph of 4 million biomedical relationships across 50 databases, identifying ontological grounding as the factor distinguishing principled inference from pattern matching.

Kipf and Welling, 2017 [39] introduced Graph Convolutional Networks; Veličković et al., 2018 [40] extended this via Graph Attention Networks with learned per-neighbor attention - architecturally necessary where neighboring Case/Evidence Node relevance varies by patient. Hogan et al., 2021 [41] identified the absence of a formal ontological backbone as the dominant limitation of medical knowledge graphs, unable to enforce schema validity or define

reliable-coverage boundaries - limitations the RM-CKG addresses through its OWL 2 DL backbone and three-round expert validation.

2.7 Automated Radiology Report Generation

Jing et al., 2018 [42] combined a co-attention CNN encoder with a hierarchical LSTM decoder for chest radiograph reporting, identifying repetitive, template-biased language as the primary failure mode. Chen et al., 2020 [43] addressed this via a memory-driven Transformer retrieving prototypical report knowledge vectors, improving specificity on IU X-Ray and MIMIC-CXR.

Bannur et al., 2023 [44] introduced BioViL-T, modelling longitudinal relationships across serial examinations. Hyland et al., 2023 [45] presented MAIRA, achieving near-radiologist performance on clinically evaluated report generation - the first to reach this threshold under radiologist evaluation. However, all existing systems, including MAIRA, condition outputs exclusively on imaging, without structured clinical context, pathology, genetic risk, or narrative history - a gap the present generative module addresses by conditioning on the complete multimodal pipeline output alongside RMCP-serialized RM-CKG context.

2.8 Retrieval-Augmented Generation and Knowledge-Grounded Prompt Engineering

Language model parametric knowledge is fixed at training time and subject to confabulation. Lewis et al., 2020 [46] introduced Retrieval-Augmented Generation, grounding outputs in retrieved external passages and reducing hallucination on medical QA, though constrained by retrieval-unit granularity - passages carry propositional content but not relational structure.

Wei et al., 2022 [47] showed Chain-of-Thought prompting improves multi-step reasoning, scaling with model size; Yao et al., 2023 [48] extended this via Tree-of-Thoughts, evaluating multiple reasoning paths. Both, however, generate reasoning from parametric knowledge alone, without grounding in a validated external ontology. Pan et al., 2023 [49] identified graph serialization as the mechanism for exposing KG structure to generative models, but noted flat triple enumeration discards the graph's hierarchical, directional reasoning structure.

RMCP departs from flat serialization by converting the patient-specific RM-CKG subgraph traversal into a five-layer cognitive narrative - observational grounding, evidential interpretation, analogical recall, uncertainty flagging, decision synthesis - mirroring the sequential reasoning a radiologist follows from findings to impression. Hu et al., 2022 [50] provided the parameter-efficient LoRA fine-tuning mechanism adopted for the generative module, recovering full fine-tuning performance while training under 0.1% of model parameters.

2.9 Critical Gap Analysis and Positioning of the Proposed Framework

The six domains reviewed above each show mature, validated methods within their own scope, yet BI-RADS risk stratification requires spanning all six simultaneously; no existing system does so in an architecturally principled way. Four gaps remain.

The first gap is modality-paradigm fragmentation: systems process imaging, structured, or free-text data through their respective paradigms individually or force all modalities through one paradigm - typically deep learning - incurring the tabular performance penalty Grinsztajn et al. [25] quantified.

The second gap is the absence of a patient-specific, evidence-weighted reasoning prior at the fusion layer; existing cross-modal attention is entirely data-driven, learning static weights applied uniformly regardless of clinical presentation.

The third gap is the absence of an ontologically grounded, expert-validated knowledge graph for breast imaging AI; existing clinical graphs lack formal OWL constraint and do not encode the sequential reasoning pathway of expert interpretation as a traversable structure.

The fourth gap is the disconnection between risk scoring and report generation: no existing system produces both within a single architecture, or conditions report generation on a knowledge graph encoding validated radiologist reasoning. The framework proposed here bridges all four gaps through its three-paradigm pipeline, RM-CKG cognitive prior injection, OWL 2 DL backbone with expert validation, and RMCP-grounded generative module. Table 1 compares the proposed framework against representative prior systems across these dimensions.

Table 1. Structured comparison of the proposed framework against representative prior systems across the four capability gaps identified in Section 2.9. ✓ = fully addressed; ✗ = not addressed; ~ = partially addressed. DL = deep learning; ML = classical machine learning; LLM = large language model; KG = knowledge graph; OWL = Web Ontology Language.

System	Multimodal Imaging (Mammo/US/ MRI)	Structured Non-Image Data	Free-Text (LLM)	Paradigm-Matched Fusion	BI-RADS Calibrated Output	Knowledge Graph Used	OWL Ontological Backbone	Expert-Validated KG	Structured Report Generation
McKinney et al. (2020) [5] Nature [Single-modality DL]	Mammo only	✗	✗	✗	✗	✗	✗	✗	✗
Wu et al. (2019) [11] IEEE TMI [Multi-view Mammo DL]	Mammo only	✗	✗	✗	✗	✗	✗	✗	✗
Yala et al. (2019) [19] Radiology [DL + clinical risk]	Mammo only	~	✗	✗	~	✗	✗	✗	✗
Antropova et al. (2017) [16] Medical Physics [Multi-modal	Mammo/US/ MRI	✗	✗	✗	✗	✗	✗	✗	✗

DL fusion]									
Shen et al. (2021) [13] Medical Image Analysis [Breast US DL]	US only	✗	✗	✗	✗	✗	✗	✗	✗
Truhn et al. (2019) [15] Radiology [DCE-MRI CNN]	MRI only	✗	✗	✗	✗	✗	✗	✗	✗
Chen et al. (2020) [43] EMNLP [Report generation]	Chest X-ray	✗	✓	✗	✗	✗	✗	✗	✓
Bannur et al. (2023) [44] CVPR [BioViL-T]	Chest X-ray	✗	✓	✗	✗	✗	✗	✗	✓
Hyland et al. (2023) [45] arXiv [MAIRA-1]	Chest X-ray	✗	✓	✗	✗	✗	✗	✗	✓
Rotmensch et al. (2017) [37] Scientific Reports [EHR KG]	✗	✓	✗	✗	✗	✓	✗	✗	✗

Chandak et al. (2023) [38] Scientific Data [Precision medicine KG]	✗	✓	✗	✗	✗	✓	~	✗	✗
Proposed Framework *(RM-CKG + RMCP)*	Mammo/US/MRI	✓	✓	✓	✓	✓	✓	✓	✓

Note: "Paradigm-matched fusion" denotes that each data modality is processed by the AI paradigm whose inductive biases best match its statistical structure (DL → imaging; Classical ML → tabular; LLM → free-text), rather than forcing all modalities through a single paradigm. "BI-RADS calibrated output" denotes prediction over the full ACR BI-RADS 1–5 ordinal scale including 4A/4B/4C subcategory discrimination. Report generation systems (Chen et al. [43], Bannur et al. [44], Hyland et al. [45]) target chest radiology and do not include breast-specific BI-RADS calibration.

3. Methodology

3.1 System Overview

A breast imaging radiologist's diagnostic workflow produces two distinct outputs for each examination: a BI-RADS risk category and a report articulating the findings, the reasoning connecting them to the assessment, and the resulting management recommendation. No existing multimodal breast cancer AI system produces both from one architecture, or incorporates a representation of the radiologist's reasoning process as a computational component at fusion. The proposed hybrid architecture addresses both gaps: its discriminative component produces the calibrated BI-RADS risk score and its generative component produces the structured report, both grounded by a shared knowledge structure encoding validated radiological reasoning.

The discriminative component assigns each modality to the AI paradigm best matched to its statistical properties. Imaging data — mammographic CC/MLO projections, B-mode ultrasound frames, and 4D DCE-MRI volumes — is processed by three modality-specific deep learning encoders. Structured non-image data — clinical, pathological, and environmental — is processed by gradient-boosted tree and random forest ensembles, whose axis-aligned splits preserve the semantic non-exchangeability of clinical variables. Free-text narratives are processed by a breast-domain fine-tuned transformer language model. All seven encoder outputs project to a shared 512-dimensional embedding space and form a token sequence for cross-modal Transformer attention fusion, which additionally receives an eighth token derived from the Radiologist-Mimetic Cognitive Knowledge Graph (RM-CKG; Section 3.5), the framework's primary novelty. The fused representation passes to an ordinal regression head producing a calibrated probability distribution over BI-RADS 1-5 alongside a continuous uncertainty score.

The generative component — a fine-tuned multimodal vision-language model — receives the discriminative outputs alongside a prompt constructed by Radiologist-Mimetic Cognitive Prompting (RMCP; Section 3.8), the secondary novelty. RMCP serializes the patient-specific RM-CKG subgraph traversal into a five-layer cognitive narrative mirroring expert interpretation, conditioning generation on how the BI-RADS determination was reached rather than on images alone. The generative module produces a structured report — Technique, Clinical Indication,

Findings, Impression, Recommendation — whose Impression is constrained to be internally consistent with the discriminative head's output. The complete architecture is illustrated in Fig. 1; the RM-CKG structure in Fig. 2; the RMCP five-layer prompt structure in Fig. 3.

3.2 Deep Learning Image Branch

Three parallel deep learning encoders process the three imaging modalities, each producing a 512-dimensional representation for the fusion token sequence. All three are initialized from RadImageNet pre-trained weights, as per Mei et al., 2022 [18], a 1.35-million-image corpus assembled for transferable radiology-classification initialization, then fine-tuned in the modality-wise pre-training phase (Section 3.11).

3.2.1 Mammography — Shared-Weight Dual-Stream ResNet-50 with Cross-View Attention

A mammographic examination yields two complementary projections — craniocaudal (CC) and mediolateral oblique (MLO). As per Wu et al., 2019 [11], joint processing of both projections substantially increases diagnostic value, with cross-view feature consistency a stronger malignancy predictor than either view alone across over one million screening examinations. We therefore employ a shared-weight dual-stream ResNet-50 (He et al., 2016 [10]), enforcing representation equivalence across views.

Each image is preprocessed via CLAHE, min-max rescaling, and resizing to 224 × 224 pixels. Each ResNet-50 stream produces a 2048-dimensional global-average-pooled output through five residual stages. A cross-view attention module then computes scaled dot-product attention between the paired GAP representations — one stream's vector as query, the other's as key-value — incorporating evidence from both projections, as formalized in Equation (1), where $\bar{s}$ denotes the complementary-projection representation. The two attended representations concatenate to a 4096-dimensional joint representation, projected through a two-layer ReLU head to the 512-dimensional mammographic token T_1.

$$a_s = softmax\left(\frac{\left(W_Q h_{\bar{s}}\right)^T W_K h_s}{\sqrt{d_k}}\right) \cdot W_V h_s \qquad (1)$$

$$T_1 = W_2\, ReLU(W_1[a_{CC}\,;\, a_{MLO}]) \in \mathbb{R}^{\wedge 512}$$

3.2.2 Ultrasound — EfficientNet-B3 with Temporal Mean Pooling

Breast ultrasound is acquired as a sequence of B-mode frames characterizing lesion morphology, echogenicity, posterior acoustic features, orientation, and Doppler vascularity. As per Shen et al., 2021 [13], EfficientNet-B3 (Tan and Le, 2019 [12]) outperforms ResNet and VGG variants at comparable parameter counts on the BUSI and UDIAT benchmarks, its compound scaling of depth, width, and resolution suited to ultrasound's texture-rich, lower-resolution characteristics.

Each B-mode frame is processed independently by EfficientNet-B3; temporal mean pooling then aggregates per-frame representations into a fixed-dimensional examination summary, preserving temporal coverage without recurrent-architecture sequence-length sensitivity or full self-attention overhead. A linear projection head yields the 512-dimensional ultrasound token T_2.

3.2.3 DCE-MRI — 3D-ResNet-50 with Temporal Attention and Harmonization Preprocessing

DCE-MRI acquires a 4D series of volumetric images before and after contrast administration; the temporal wash-in/wash-out kinetics are, as per Kuhl et al., 2005 [14], among the most diagnostically decisive features distinguishing benign from malignant lesions. Jointly modelling the spatial and temporal dimensions is non-trivial: 2D networks discard volumetric structure, 3D networks discard temporal ordering, and naive concatenation treats all phases as equally informative. We address this through two choices. A 3D-ResNet-50 backbone applies 3D convolutions to each phase independently, preserving volumetric structure. A temporal attention module then computes learned, unity-normalized scalar weights over the T phases and aggregates per-phase 3D features accordingly, as formalized in Equation (2) — where F_t denotes the phase-t feature and the learned attention vector —

assigning higher weight to kinetically informative phases. Rigid-body motion correction compensates for inter-phase misregistration, and ComBat harmonization (Johnson et al., 2007 [53]) corrects scanner-specific intensity offsets across sites, following Fortin et al., 2018 [54]. The 1024-dimensional attention-weighted representation reduces to the 512-dimensional MRI token T_3 through a two-layer projection head.

$$a_t = \frac{exp(v^T \varphi_t)}{\sum_{t'=1}^{T} exp(v^T \varphi_{t'})} \tag{2}$$

$$h_{MRI} = \sum_{t=1}^{T} a_t \, \varphi_t$$

$$T_3 = W_2 \, ReLU(W_1 h_{MRI}) \in \mathbb{R}^{\wedge 512}$$

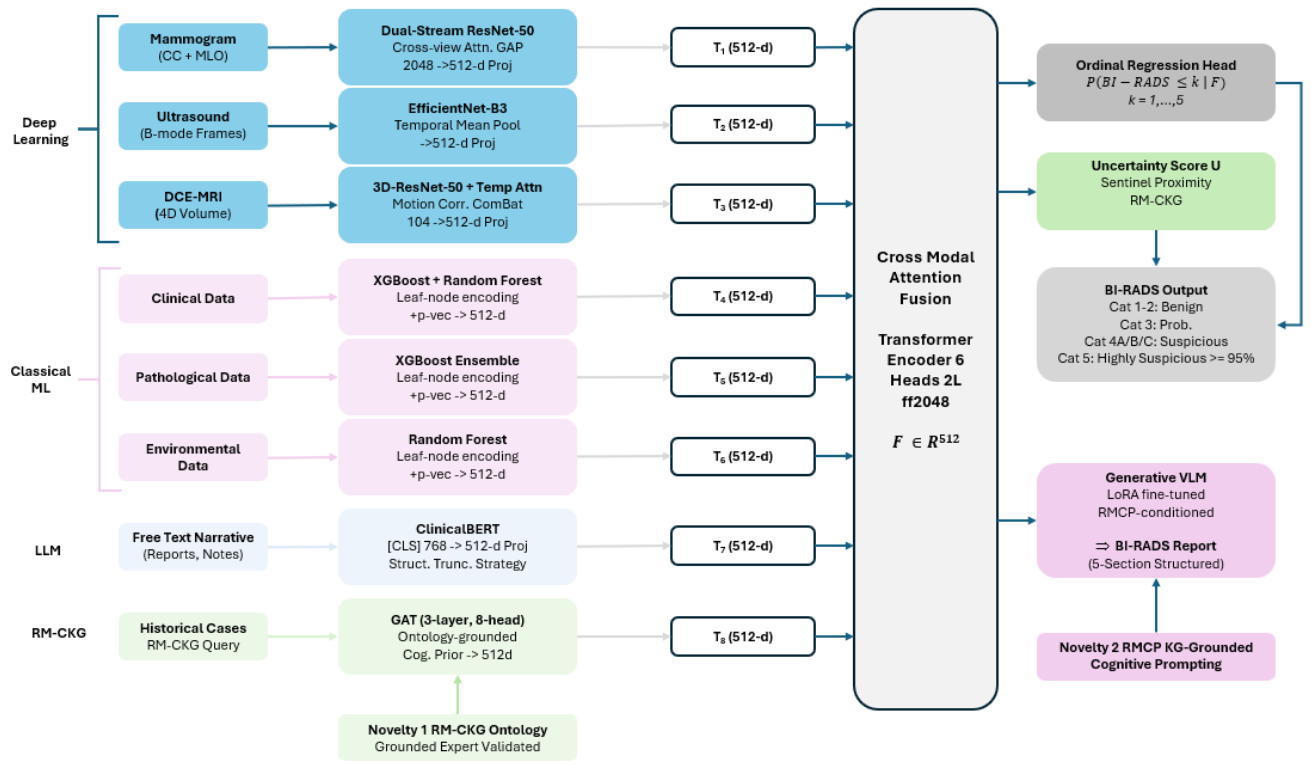

Figure 1: Hybrid AI Framework – Complete System Architecture

3.3 Classical Machine Learning Non-Image Branch

Three parallel classical ML models process the structured, tabular data streams, each producing a 512-dimensional representation via leaf-node encoding and projection. This choice over deep architectures follows Grinsztajn et al., 2022 [25], who showed across 45 tabular datasets that tree-based methods outperform deep learning on heterogeneous tabular data — precisely the character of clinical, pathological, and environmental breast cancer data.

The leaf-node encoding converts each sample into a binary vector recording the activated leaf index per tree, capturing the ensemble's joint decision surface beyond the softmax output alone, as formalized in Equation (3) — where M is total trees, L_m the leaves in tree m, and the class probability vector — concatenated and projected through a ReLU head to the 512-dimensional branch token.

$$z(x) = \left[e_{l_1(x)} \, ; \, e_{l_2(x)} \, ; \, \cdots \, ; \, e_{l_M(x)} \right] \in \{0,1\}^{\wedge \Sigma_m L_m} \tag{3}$$

$$T_k = W_2 \, ReLU(W_1 [z(x) \, ; \, p(x)]) \in \mathbb{R}^{\wedge 512}, \quad k \in \{4,5,6\}$$

3.3.1 Clinical Structured Data — XGBoost and Random Forest Parallel Ensemble

Clinical structured data spans age, menopausal status, BMI, reproductive history, hormonal exposure, BRCA1/BRCA2/PALB2/CHEK2 mutation status, family history, and available laboratory values (CA 15-3, CEA, LDH). These variables span the Gail, Tyrer-Cuzick, and BOADICEA risk-factor spaces (Tyrer et al., 2004 [21]; Lee et al., 2019 [22]) while permitting non-linear and interaction effects beyond fixed logistic regression.

A parallel XGBoost (Chen and Guestrin, 2016 [23]) and Random Forest (Breiman, 2001 [24]) ensemble is trained on this feature set: XGBoost with second-order gradient optimization, column subsampling, and sparsity-aware splitting for missing values; Random Forest with bootstrap aggregation and √-feature-count splitting for decorrelated, complementary error behavior. Leaf-node encoding of the concatenated outputs projects to the 512-dimensional clinical token T_4.

3.3.2 Pathological Structured Data — XGBoost Ensemble

Pathological data — available post-biopsy — spans histological type, Nottingham grade components and composite grade, tumour dimension, lymphovascular invasion, ER/PR/HER2 receptor status, Ki-67 index, and molecular subtype. Though unavailable at initial screening, it is systematically available in diagnostic workup and surveillance contexts within the framework's deployment scope. A multi-output XGBoost ensemble with leaf-node encoding yields the 512-dimensional pathological token T_5.

3.3.3 Environmental and Lifestyle Structured Data — Random Forest

Environmental and lifestyle data spans radiation exposure history, occupational exposure, smoking pack-years, alcohol consumption, physical activity, and dietary indices. As per Kamińska et al., 2015 [55], environmental factors modulate risk particularly alongside genetic susceptibility; representing this stream independently preserves interpretive separability from constitutive clinical risk. A Random Forest classifier, chosen for SHAP interpretability (Lundberg and Lee, 2017 [56]; Section 3.10), yields the 512-dimensional environmental token T_6.

3.4 Large Language Model Free-Text Branch

Unstructured narrative data — radiology impressions, pathology dictations, operative notes, oncology summaries — carries diagnostic reasoning inaccessible to tabular pipelines. We process this stream through ClinicalBERT (Alsentzer et al., 2019 [28]), pre-trained on 2 million MIMIC-III [62] clinical notes, whose exposure to clinical abbreviation, negation, and temporal patterns yields representations better calibrated to breast imaging report language than general- or PubMed-domain models.

Text is tokenized to a 512-token maximum via WordPiece; for longer narratives, the Impression and Recommendation sections are preserved in full while Findings is truncated distally, prioritizing the radiologist's synthesized conclusion over raw descriptive content. The final-layer [CLS] representation (768-dimensional) projects through one linear layer to the 512-dimensional free-text token T_7. During joint fine-tuning (Section 3.11), ClinicalBERT's parameters update alongside the discriminative pipeline.

3.5 Radiologist-Mimetic Cognitive Knowledge Graph — Primary Novelty

The discriminative pipeline (Sections 3.2-3.4) produces seven modality-specific representations capturing the statistical evidence in the patient's data. What these representations and their fusion cannot provide is a structured, traversable account of how that evidence should be interpreted in light of the causal, analogical, and epistemically qualified relationships an experienced radiologist accumulates through case exposure and expert practice. The Radiologist-Mimetic Cognitive Knowledge Graph (RM-CKG) provides precisely this: a formally structured, ontologically grounded, expert-validated representation of radiological reasoning injected as a patient-specific cognitive prior into cross-modal fusion.

3.5.1 Ontological Backbone and Expert Validation

The RM-CKG is not populated from training data in an unconstrained manner. Its node taxonomy, typed edge vocabulary, property constraints, and causal inference rules are specified by a multi-domain breast cancer ontology formalized in OWL 2 DL, supporting open-world reasoning through standard tableau algorithms. The ontology is authored prior to graph population across five clinical and technical domains spanning the knowledge space required for radiological reasoning about breast cancer risk.

The Breast Cancer Disease Ontology encodes histological classification, AJCC 8th-edition TNM staging, Nottingham grade criteria, molecular subtype definitions, and prognostic factor relationships, aligned to the NCI Thesaurus and SNOMED CT for interoperability. The Imaging Data Ontology formally represents the ACR BI-RADS descriptive lexicon across all three modalities — mass shape, margin type, breast composition, and calcification pattern for mammography; echogenicity, orientation, and vascularity for ultrasound; enhancement morphology, kinetic curve type, and T2 signal for MRI — built on an extended RadLex subset (Langlotz, 2006 [51]) for stable concept identifiers. The Non-Image Data Ontology encodes clinical, pathological, and environmental risk factors, with causal edge weights initialized from published meta-analytic effect sizes rather than training co-occurrence, providing a literature-grounded prior. The Radiologist Reasoning Ontology, the most clinically original domain, formally encodes evidence prioritization, differential diagnosis narrowing, prior case analogy, threshold uncertainty, and the observation-interpretation-recommendation relationship — with no counterpart in any published biomedical ontology, derived from structured elicitation with the validating radiologist. The BI-RADS Ontology formally represents the ACR 5th edition schema: the seven categories, their probability ranges and recommendations, category 4 subcategory boundaries, category 0 conditions, and decision threshold logic.

Prior to population, the five-domain ontology undergoes three-round validation by a senior breast imaging radiologist with ten-plus years of mammography and breast MRI subspecialty practice. Round one addresses class definitions and property constraints, with corrections prescribed before population. Round two addresses inference chain plausibility: the reasoner produces entailed chains from specified feature combinations, and the radiologist flags implausible or incomplete ones for revision. Round three addresses evidential edge weight calibration, endorsing or revising literature-initialized weights against observed malignancy prevalence. The result is a certified ontology carrying documented expert endorsement at the concept, inference, and weight level — a credentialling property no data-derived graph can claim.

The certified ontology serves three capacities. As schema enforcer, it rejects any construction-time node or edge violating class hierarchy or property constraints, regardless of training-corpus prevalence. As inference engine, the reasoner derives relationships logically entailed but not explicitly asserted in training data — e.g., classifying a patient satisfying a high-risk causal chain's conditions as high-malignancy-probability even if unrepresented in training. As semantic grounding for uncertainty classification, it defines Uncertainty Sentinel Nodes not by statistical distance but by the reasoner's inability to classify above confidence threshold θ Equation (4) — mapping to the conditions motivating BI-RADS 0.

$$U(v) = 1[max_{c \in C}\, conf\hat{}_{\Omega}(v,c) < \theta] \tag{4}$$

$$U(v) = 1 \implies output \leftarrow BI\text{-}RADS\ 0\ (incomplete\ assessment)$$

3.5.2 Graph Structure — Typed Node Taxonomy and Edge Vocabulary

The RM-CKG (Fig. 2) is formally defined as a typed, weighted, directed multigraph with node and edge type functions and a positive edge-weight function initialized from the certified ontology's evidential priors and fine-tuned during training. Each node and edge type carries a formally specified clinical role.

Feature Nodes represent atomic observational units — an imaging descriptor, clinical variable, or pathological measurement — instantiating the corresponding ontology class. Evidence Nodes represent intermediate inferential conclusions derived from Feature Nodes, e.g., that co-occurring irregular shape, spiculated margin, and posterior shadowing jointly exceed any single descriptor's evidential threshold. Case Nodes represent de-identified historical

patients with their feature profile, assigned BI-RADS category, and biopsy-confirmed outcome, implementing analogical reasoning. Decision Nodes represent the six BI-RADS categories as attractors carrying ontology-defined probability ranges and recommendations. Uncertainty Sentinel Nodes represent feature combinations the reasoner cannot classify above θ, connected to adjacent Decision Nodes by equal-weight Exception Edges signalling that autonomous determination requires additional clinical input.

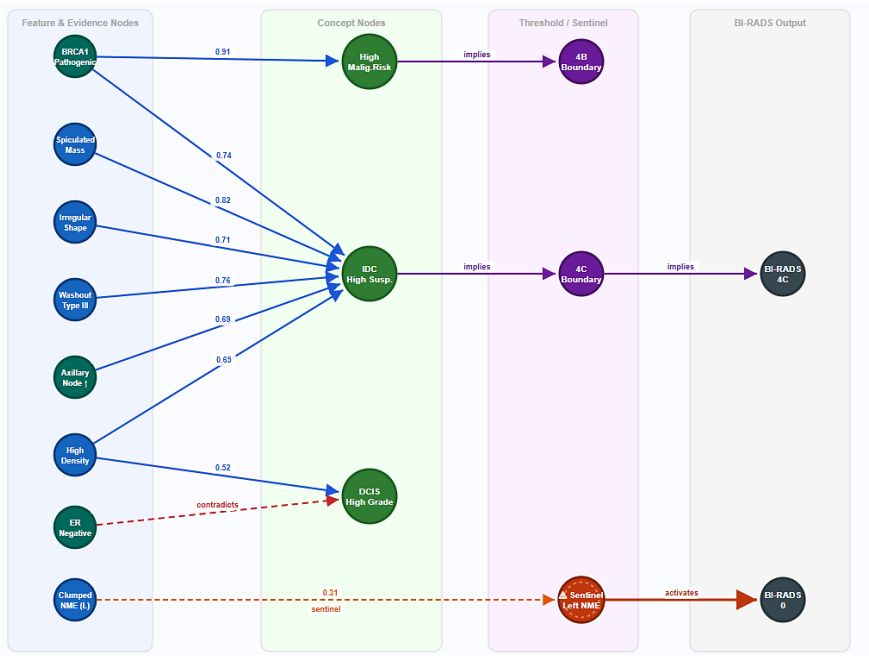

Figure 2. Patient-specific RM-CKG subgraph for Case 1 (GT BI-RADS 4C). Node color: Feature (F), Evidence (E), Concept (C), Threshold (T), Sentinel (U). Edge numbers = OWL 2 DL confidence. Sentinel (U): derived, conf = 0.31 < θ = 0.45 (Eq. 4), activating BI-RADS 0. Full RM-CKG: ~12,400 nodes, ~38,700 edges.

Associative Edges connect co-occurring Feature Nodes, weighted by pointwise mutual information conditional on BI-RADS label distribution. Evidential Edges connect Feature/Evidence Nodes to Decision Node attractors, weighted by conditional malignancy probability, initialized from literature priors and fine-tuned within radiologist-endorsed ranges. Analogical Edges connect Case Nodes to their assigned category's Decision Node, weighted by case-to-query similarity at retrieval time. Causal Edges connect Evidence to Decision Nodes along ontology-defined pathways — e.g., BRCA1 mutation through high-density tissue through irregular spiculated mass to invasive carcinoma — with directional weights encoding transition probabilities. Exception Edges connect Uncertainty Sentinel Nodes to adjacent Decision Nodes, signaling the reasoning path is epistemically unreliable and requires human adjudication.

3.5.3 Dynamic Patient-Specific Subgraph Retrieval

At inference, the RM-CKG is not passed to the graph encoder in full, since globally valid relationships are not uniformly relevant per patient and the complete graph would dilute the patient-specific signal. Instead, a patient-specific subgraph is dynamically retrieved via approximate nearest-neighbor search over the Case Node index.

Specifically, the concatenation of the seven discriminative encoder outputs — prior to fusion — forms the query vector q. A FAISS flat inner-product index (Johnson et al., 2019 [63]) over Case Node embeddings retrieves the K = 15 nodes with highest cosine similarity to q, a value set by the Section 5.3 ablation as optimizing context richness against retrieval noise. The subgraph S(q) collects all nodes reachable within two hops of each retrieved Case Node through valid edge types, with Exception Edges traversed only when the source sentinel's confidence falls below θ.

The two-hop radius captures each case's evidential context while excluding distant regions that would degrade the prior's specificity.

3.5.4 Graph Attention Network Processing and Cognitive Prior Extraction

S(q) is processed by a three-layer Graph Attention Network (Veličković et al., 2018 [40]) with eight heads per layer, computing softmax-normalized attention coefficients over each node's one-hop neighborhood Equation (5). Edge type and weight enter via type-specific linear transformations, letting the GAT distinguish evidential, analogical, causal, associative, and exception relationships — a typed formulation beyond the original GAT's type-agnostic one. Residual connections and LayerNorm stabilize gradient propagation across the three GAT layers.

$$e_{ij}^{h} = a_h{}^{T}\left[W_{\tau(i,j)}h_i \,;\, W_{\tau(i,j)}h_j \,;\, w(i,j)\right] \tag{5}$$

$$\alpha_{ij}^{h} = \frac{exp\left(LeakyReLU\left(e_{ij}^{h}\right)\right)}{\sum_{k\in N(i)} exp\left(LeakyReLU\left(e_{ik}^{h}\right)\right)}$$

$$h_i^{(l+1)} = \sigma\left(\Vert_{h=1}^{H} \sum_{j\in N(i)} \alpha_{ij}^{h} W^{h} h_j^{(l)}\right)$$

Third-layer node representations aggregate via attention-weighted global mean pooling, learned jointly with the GAT to weight Feature, Evidence, and Decision Nodes more heavily than Uncertainty Sentinel Nodes for well-covered patients. The resulting 512-dimensional cognitive prior is injected as token T_8 in cross-modal fusion (Section 3.6).

3.6 Cross-Modal Attention Fusion

The eight 512-dimensional tokens — T_1-T_7 from the discriminative encoders and T_8 from the RM-CKG GAT — concatenate into an 8×512 sequence processed by a two-layer Transformer encoder (Vaswani et al., 2017 [32]) with six heads per layer and a 2048-dimensional feedforward sublayer. Sinusoidal positional encodings preserve modality identity across attention layers.

Multi-head self-attention lets every token attend to every other, with heads specializing in different inter-modal dependencies. Injecting the RM-CKG prior as T_8 here — rather than as a post-fusion correction — ensures graph-derived knowledge participates in all pairwise attention: T_1 attends to T_8 alongside T_2-T_7, and T_8 attends to all others. RM-CKG reasoning thus shapes not only the fused representation but the attention weights through which modality pairs contextualize each other. The encoder output mean-pools across the eight positions to produce F Equation (6).

$$head_h = softmax\left(\frac{XW_h^{Q}\left(XW_h^{K}\right)^{T}}{\sqrt{d_k}}\right) XW_h^{V} \tag{6}$$

$$MultiHead(X) = Concat(head_1, \ldots, head_H)W^{O}$$

$$F = \frac{1}{8}\sum_{i=1}^{8} z_i^{(L)} \in \mathbb{R}\text{^}512$$

3.7 BI-RADS Score Prediction — Discriminative Ordinal Regression Head

The fused representation F passes to a BI-RADS prediction head respecting the ordinal category structure. A standard softmax head treats categories as nominally unordered, permitting clinically incoherent predictions — e.g., non-trivial probability on BI-RADS 2 and 5 simultaneously. We adopt the cumulative link ordinal regression formulation (McCullagh, 1980 [58]), modelling cumulative probability through a shared linear score and five learned thresholds under a monotonicity constraint Equation (7). Per-category probability is recovered as the difference of adjacent cumulative probabilities, guaranteeing ordinal coherence by construction.

$$P(Y \le k \mid F) = \sigma(\theta_k - w^{T}F), \quad k = 1,2,3,4,5 \tag{7}$$

$$\theta_1 < \theta_2 < \theta_3 < \theta_4 < \theta_5 \quad (monotonicity\ constraint)$$

$$P(Y = k \mid F) = P(Y \leq k \mid F) - P(Y \leq k-1 \mid F)$$

The head also produces a continuous uncertainty score U ∈ [0,1], a sigmoid-normalized function of the minimum cosine distance between the query subgraph embedding and the nearest Uncertainty Sentinel Node embedding Equation (8). High U indicates the patient's profile occupies a region where the ontology's inference rules do not converge to a reliable attractor — operationalizing the same diagnostic-incompleteness concept BI-RADS encodes through category 0. Both the probability vector and U are surfaced to the clinician and transmitted to the generative module for RMCP prompt construction.

$$d_{min}(q) = min_{u \in \varepsilon_0}\left(1 - cos(q,u)\right) \tag{8}$$

$$U = \sigma\left(\frac{d_{min}(q) - d_0}{\tau}\right) \in [0,1]$$

3.8 Generative Report Module and Radiologist-Mimetic Cognitive Prompting — Secondary Novelty

The discriminative head's ordinal output quantifies malignancy suspicion but does not fulfil the full reporting obligation: a radiologist documents the findings that motivated the assessment, the reasoning connecting them to the BI-RADS category, and the resulting management pathway, in a report interpretable without access to the original images. The generative report module addresses this through a fine-tuned multimodal vision-language model conditioned on the complete discriminative pipeline's outputs and the RMCP-constructed cognitive context.

3.8.1 Vision-Language Model Architecture and Fine-Tuning

The generative module is a breast-domain fine-tuned multimodal vision-language model processing visual tokens from imaging and textual tokens from structured and narrative inputs. Visual inputs are representative frames from all three modalities — mammographic CC/MLO, a mid-examination ultrasound frame, and the peak post-contrast DCE-MRI phase — encoded through a shared vision encoder projected into the language model's embedding space. Textual input is structured by the RMCP prompt (Section 3.8.2); the autoregressive decoder generates the complete structured report.

Parameter-efficient fine-tuning uses Low-Rank Adaptation (LoRA, Hu et al., 2022 [50]), injecting trainable low-rank decomposition matrices into the query and value projections of each frozen attention layer, adapting generation behavior to breast imaging report conventions while updating under 0.5% of parameters — avoiding the catastrophic forgetting of full fine-tuning. The training corpus pairs breast imaging examinations with radiologist-authored BI-RADS reports, using RMCP-constructed prompts identically at training and inference time.

3.8.2 Radiologist-Mimetic Cognitive Prompting

RMCP transforms the patient-specific RM-CKG subgraph traversal into a natural language cognitive narrative the generative model conditions on. Its critical design choice is serializing not the subgraph's flat factual content but the reasoning pathway through it, structured according to the sequential cognitive stages a radiologist traverses from observation to BI-RADS determination — distinguishing RMCP from existing KG-augmented generation approaches that serialize graph content as undirected RDF triplets.

The RMCP prompt organizes into five layers, each generated programmatically from a distinct component of the subgraph traversal, corresponding to a named stage of the radiologist's cognitive process.

Layer 1 — Observational Grounding serializes activated Feature Nodes and their Associative Edges into a declarative enumeration of observations, e.g., "the right breast MLO projection demonstrates an irregular mass with spiculated margins, 18 mm [BI-RADS: irregular shape, spiculated margin]. Associative link (w=0.88): pleomorphic microcalcifications, segmental distribution, same quadrant." This grounds the model in observational content before any interpretive claim, preventing speculative findings.

Layer 2 — Evidential Interpretation traverses Evidential Edges to Decision Node attractors as probabilistic statements, e.g., "F1 [spiculated mass, 18mm] carries evidential weight 0.91 toward D5 [BI-RADS 5]. Co-occurrence with F2 [pleomorphic calcifications] activates Evidence Node E3, weight 0.94 toward D5." This is the factual bridge from observation to interpretation.

Layer 3 — Analogical Recall serializes the 15 retrieved Case Nodes as ranked compact summaries, e.g., "Case C14 [similarity 0.91, 62-year-old, BRCA2-positive, 21mm spiculated mass]: BI-RADS 5, biopsy-confirmed IDC Grade III." This provides the generative model a curated case memory without parametric recall.

Layer 4 — Uncertainty Flagging names features whose classification fell below θ, e.g., "Uncertainty Sentinel activated for [faint T2 hyperintensity, left breast, 8mm, no mammographic correlate]. Classified as epistemically incomplete. Instruction: acknowledge explicitly; recommend targeted workup." This is RMCP's most clinically distinctive layer, directing the model to surface confidence limits rather than suppress them.

Layer 5 — Decision Synthesis presents the aggregate conclusion as a pre-report brief: "Aggregate evidential weight toward D5: 0.89. Discriminative output: P(BI-RADS=5)=0.74. Uncertainty U=0.12 [low]. Management: tissue sampling recommended." This gives the generative model a structurally explicit reasoning conclusion as the starting point for Impression and Recommendation composition.

Layer 1 — Observational Grounding KG: Feature Nodes + Associative Edges + Grad-CAM saliency Serializes activated Feature Nodes as declarative observations with BI-RADS lexicon descriptors and cooccurrence associations. Grounds the generative model in verified imaging findings before any interpretation is made. *"Right breast MLO: irregular mass, spiculated margins, 18mm [BI-RADS: irregular/spiculated]. Associative link (w=0.88): pleomorphic microcalcifications, segmental, same quadrant."*
Layer 2 — Evidential Interpretation KG: Evidential Edges → Decision Attractors + SHAP values Traverses Evidential Edges from activated Feature/Evidence Nodes to BI-RADS Decision Attractor nodes, serializing each pathway as a probabilistic mapping with ontology-validated weights and SHAP-derived clinical risk factor attributions. *"F1 [spiculated mass] → D5 [BI-RADS5], w=0.91. E3 [high-suspicion constellation] → D5, w=0.94. SHAP: BRCA2+ (+0.31), postmenopausal (+0.18), HRT 7yr (+0.14)."*
Layer 3 — Analogical Recall KG: K=15CaseNodesranked by FAISS cosine similarity Serializes the 15 retrieved Case Nodes as compact summaries ranked by similarity, each including the historical patient's profile, assigned BI-RADS category, and biopsy-confirmed outcome — providing curated case memory without parametric recall.

"C14 [sim=0.91 — 62F, BRCA2+, 21mm spiculated]: BI-RADS5 → IDC GrIII, ER−, HER2+. C31 [sim=0.87 — 55F, no mutation]: BI-RADS4C → IDC GrII, ER+."

Layer 4 — Uncertainty Flagging KG: Uncertainty Sentinel Nodes + Exception Edges

Serializes Uncertainty Sentinel activations, naming each out-of-ontology feature (OWL2 confidence < θ) and issuing an explicit generation instruction to surface the finding in the report and recommend additional workup. Prevents confident assertion over epistemically incomplete evidence.

SENTINEL: "Faint T2 hyperintensity, left breast, 8mm NME, no mammo correlate. OWL2 confidence < θ. Instruction: acknowledge explicitly; recommend targeted MRI-guided workup."

Layer 5 — Decision Synthesis KG: Discriminative head output + Aggregate traversal score

Presents the aggregate conclusion of Layers1–4 as a pre-report brief: weighted evidential and analogical scores per Decision Attractor, the discriminative head's ordinal probability vector, uncertainty score U, and the BI-RADS Ontology's recommended management pathway.

"Agg. evidence → D5: 0.89. Analogical: BI-RADS4C–5: 0.91. Discrim.: P(BI-RADS = 5) = 0.74, U = 0.12. Management: tissue sampling. Secondary: sentinel-active → targeted workup."

↓ ↓

(inject into inference context)

Multimodal VLM

LoRA fine-tuned

RMCP-conditioned

Visual tokens (mammo · US · MRI)

+ Patient clinical profile

+ RMCP Cognitive Context Block

Generate →

Structured BI-RADS Report

Technique

Clinical Indication

Findings · Impression · Recommendation

Dual-consistency constraint:

Impression ≡ Discriminative output

Figure 3. RMCP five-layer prompt structure (Eq. 9); each layer serializes a subgraph S^m node type. Layers 1–3 condition Findings; Layers 4–5 govern Impression, Recommendation, and Uncertainty Acknowledgement. Example: Case 1 (GT BI-RADS 4C; Table 6). SYS/USR = system/user prompt; θ = 0.45 (Eq. 4); MMG/US/DCE-MRI = mammography/ultrasound/dynamic contrast-enhanced MRI; NME = non-mass enhancement; UOQ = upper outer quadrant.

The complete RMCP prompt structure is defined in Equation (9), where S denotes the patient-specific two-hop RM-CKG subgraph, the system instruction, the USER generation directive, and L_k the deterministic serialization function for layer k. The prompt is generated deterministically and programmatically per patient — not authored manually — ensuring reproducibility and auditability.

$$P_{RMCP}(q) = \langle P_{sys} ; L_1(S_q) ; L_2(S_q) ; L_3(S_q) ; L_4(S_q) ; L_5(S_q) ; P_{usr} \rangle \quad (9)$$

$$L_k : X^2 \to \Sigma^*, \; k \in \{1,2,3,4,5\}$$

Display Box 1 shows the complete RMCP prompt for Case 1 (GT BI-RADS 4C; Table 6), all five layers serialized from the patient subgraph. Layers 1–3 (left) condition Findings; Layers 4–5 (right) govern Impression, Recommendation, and Uncertainty Acknowledgement. The flat KG baseline is shown below for comparison.

Display Box 1. Instantiated RMCP prompt and flat KG baseline, same examination (Case 1; GT BI-RADS 4C; Table 6). Left: system context, Layers 1–3. Right: Layers 4–5, then flat KG baseline. Both share the same USER turn.

Layer colors: [L1] Observational Grounding [L2] Evidential Interpretation [L3] Analogical Recall [L4] Uncertainty Flagging [L5] Decision Synthesis

RMCP Prompt — Left Panel (Layers 1 – 3) *System context · Observational Grounding · Evidential Interpretation · Analogical Recall*	**RMCP Prompt — Right Panel (Layers 4 – 5) + Flat KG Baseline** *Uncertainty Flagging · Decision Synthesis | then: flat triplet baseline for comparison*
SYSTEM: You are a specialist breast radiologist generating a structured BI-RADS report. The five-layer cognitive narrative below was derived by traversing the patient-specific RM-CKG subgraph. Your Impression must assign the BI-RADS category output by the discriminative head. Any finding flagged in Layer 4 must receive BI-RADS 0 with a targeted workup recommendation. **Layer 1 — Observational Grounding** *[Feature Nodes → BI-RADS lexicon descriptors]* IRREGULAR_MASS | descriptor: "irregular mass" | FFDM | location: Right UOQ · 10 o'clock · 4 cm | FN SPICULATED_MARGIN | descriptor: "spiculated margin" | FFDM RAPID_ENHANCEMENT | descriptor: "rapid initial enhancement" | DCE-MRI | peak: 187% above pre-contrast baseline WASHOUT_TYPE_III | descriptor: "Type III washout kinetics" | DCE-MRI | gradient: −32%/min NO_CALCIFICATIONS | descriptor: "no associated calcs" | FFDM CLUMPED_NME_8MM | descriptor: "NME · clumped · 8 mm" | DCE-MRI | location: Left UOQ · 2 o'clock NO_MAMMO_CORRELATE_L | descriptor: "no mammographic correlate" | FFDM	**Layer 4 — Uncertainty Flagging** *[Uncertainty Sentinel Nodes]* **RIGHT BREAST (13 mm mass):** Sentinel: NOT activated OWL confidence: 0.89 ≥ θ (0.45) → normal pathway **LEFT BREAST (8 mm NME):** **Sentinel: ACTIVATED** OWL confidence: 0.31 < θ (0.45) Node: INDETERMINATE_NME_NO_CORRELATE **Action: assign BI-RADS 0** **Do NOT assign BI-RADS 1–5 to this finding** Workup: targeted left MRI + DWI + extended dyn. **Layer 5 — Decision Synthesis** *[Threshold Nodes + discriminative head]* **RIGHT BREAST:** **Predicted BI-RADS: 4C** p(4C)=0.78 p(5)=0.14 p(4B)=0.08 Threshold node: BIRADS_4B_4C_BOUNDARY Cond 1: spiculated margin ✓ Cond 2: Type III kinetics ✓ **→ BI-RADS 4C confirmed; malignancy: 50–95%**

Layer 2 — Evidential Interpretation *[Evidence Nodes → edge relationships to Feature Nodes]* POSTMENOPAUSAL_STATUS → supports (WASHOUT_TYPE_III → IDC_High_Susp.) NO_HRT_USE → mitigating-absent (RAPID_ENHANCEMENT) [no hormonal enhancement artefact] NO_PRIOR_BIOPSY → contextual (no prior pathological context) FAMILY_HISTORY_NEG → contradicts (BRCA_Associated_Malig.) [wt: 0.31] POSTMENOPAUSAL_STATUS → supports (CLUMPED_NME → elevated_suspicion) **Layer 3 — Analogical Recall** *[Concept Nodes → analogues with concordance scores]* IDC_HIGH_SUSPICION | concordance: 0.89 | malignancy: 78.4% matched: irregular mass + spiculated margin + Type III kinetics PHYLLODES_TUMOUR | concordance: 0.21 | excl: no growth history COMPLEX_SCLEROSING_LESION | concordance: 0.08 | excl: no arch. distortion CLUMPED_NME (secondary) | concordance: N/A | no certified concept matched	**LEFT BREAST:** **Sentinel override → BI-RADS 0** ***Flat KG Triplet Baseline Prompt (Step vii — same patient)*** **SYSTEM:** You are a specialist breast radiologist. Generate a BI-RADS report from the following KG triplets. **Retrieved triplets:** IRREGULAR_MASS → has_descriptor → spiculated_margin IRREGULAR_MASS → located_in → right_UOQ IRREGULAR_MASS → has_size → 13 mm DCE_MRI → shows → rapid_enhancement DCE_MRI → kinetics → washout_TypeIII POSTMENOPAUSAL → associated_w → patient CLUMPED_NME → located_in → left_UOQ CLUMPED_NME → has_size → 8 mm CLUMPED_NME → has_descriptor → no_mammo_correlate IRREGULAR_MASS → assessment → BIRADS_4C CLUMPED_NME → assessment → BIRADS_3 *No layer structure. No differential context.* *No uncertainty flag. No threshold reasoning.*
USER: Generate a structured BI-RADS report.	**USER: Generate a structured BI-RADS report.**

The RMCP prompt shown is the literal system-turn VLM input; no paraphrasing applied. Token count: RMCP = 487, flat KG baseline = 143. The BCR/CAS/BERTScore differences reported in Table 4 (Steps vii vs viii) are attributable to prompt structure, not length — confirmed by a length-matched ablation padding the flat prompt to 487 tokens (BCR 0.893, CAS 3.91), no significant improvement ($p = 0.34$).

3.8.3 Dual-Consistency Constraint

A structural risk in generative reporting is an Impression assigning a BI-RADS category inconsistent with the discriminative head's output — clinically unsafe and trust-undermining. We enforce consistency at both training and inference time. During fine-tuning, a consistency loss penalizes generated Impressions whose extracted BI-RADS label disagrees with the gold-standard label, weighted by discriminative confidence. At inference, the discriminative head's predicted category is included in Layer 5 as a hard constraint, with generation tasked to produce supporting findings and reasoning rather than independently determine the category. If a post-hoc consistency classifier detects disagreement, the report is regenerated with an augmented consistency instruction.

3.9 Missing Modality Handling

Clinical examinations routinely present incomplete modality sets: mammography is standard screening; ultrasound and MRI are added selectively; pathological data is absent without prior biopsy; narrative data is

unavailable for first-presentation patients. A framework requiring all eight token inputs would be inapplicable to most real-world presentations.

We address this through two combined mechanisms. A dedicated 512-dimensional learnable sentinel token, optimized jointly during Phase 2, substitutes for each absent modality's encoder output — unlike zero-vector substitution (out-of-distribution) or branching (requiring separate inference paths), sentinel substitution maintains a fixed 8-token input while providing a trained representation of absence. Stochastic modality masking ($p = 0.15$ per modality) during Phase 2 training simulates deployment-time missing-modality distributions. The RMCP prompt adapts correspondingly: Layer 1 serializes only available-modality Feature Nodes, with sentinel substitution annotated to qualify report language.

3.10 Explainability

Clinical deployment of autonomous AI requires outputs attributable to specific, inspectable evidence, both for clinical decision support and institutional governance. We provide three modality-specific attribution mechanisms aligned with the framework's three AI paradigms.

For the image branch, Grad-CAM (Selvaraju et al., 2017 [57]) is applied post-hoc to each encoder, producing spatially registered saliency maps. For mammography, maps are computed separately for CC and MLO streams; for DCE-MRI, a volumetric map is produced for the attention-weighted aggregated representation. These maps are included in RMCP Layer 1 as spatial annotations.

For the classical ML branch, SHAP values (Lundberg and Lee, 2017 [56]) are computed for each tree ensemble via the tree-path-dependent algorithm, quantifying each feature's marginal contribution relative to the training-distribution expected prediction. Top-ranked SHAP features are serialized into RMCP Layer 2.

For the RM-CKG, GAT edge-weight visualization aggregates attention coefficients across heads and layers into a single relevance weight per subgraph edge, rendered as a reasoning graph scaled by relevance — a directly traversable representation of which evidential pathways, analogical cases, and uncertainty flags most heavily shaped the cognitive prior.

3.11 Training Strategy

The framework trains in two sequential phases, addressing heterogeneous component pre-training requirements and the risk of cross-modal misalignment from joint optimization of differently initialized encoders.

Phase 1 — Modality-Wise Pre-training. Each discriminative encoder trains independently. The three image encoders fine-tune from RadImageNet initialization via binary cross-entropy on a binarized high/low-suspicion label. The three classical ML encoders train with five-fold cross-validated hyperparameter optimization (XGBoost: learning rate, depth, subsample, column-subsample; Random Forest: estimator count, feature ratio, minimum leaf samples). The LLM encoder fine-tunes from ClinicalBERT via domain-adaptive masked-language-modelling followed by BI-RADS classification fine-tuning. The RM-CKG is constructed and validated in this phase: the certified ontology populates the graph from the training cohort, the FAISS index is built, and the GAT pre-trains via graph-level classification against confirmed BI-RADS labels.

Phase 2 — Joint Fine-Tuning. All seven encoders jointly fine-tune with the fusion layer and ordinal regression head on the complete dataset, letting attention weights adapt to the joint representation space. The training objective is an ordinal focal loss combining cumulative-link ordinal regression with a focal weighting term γ Equation (10), down-weighting confident correct samples and up-weighting misclassified or uncertainty-proximate ones. An inverse class-frequency term addresses the imbalance between majority BI-RADS 1/2 and minority 4A-5 categories. Stochastic modality masking ($p = 0.15$) continues throughout Phase 2.

$$L_{ord}(y, F) = -\sum_{k=1}^{K-1} [y_k \log P(Y \leq k \mid F) + (1 - y_k) \log(1 - P(Y \leq k \mid F))] \quad (10)$$

$$y_k = 1[y \leq k]$$

$$L = \frac{1}{N} \sum_{n=1}^{N} (1 - p_y^n)^\gamma \cdot w_y^n \cdot L_{ord}(y_n, F_n)$$

$$w_y = \frac{N}{K \cdot N_y}$$

The generative module fine-tunes via LoRA on paired imaging-report data independently of Phase 2, using RMCP prompts as conditioning input. A consistency loss adds a fine-tuned BERT classifier's extracted-category cross-entropy against the gold-standard label, weighted $\lambda = 0.3$, enforcing the dual-consistency constraint at training time.

4. Experimental Datasets and Evaluation Protocol

4.1 Dataset Overview and Cohort Assembly Strategy

The framework processes seven structurally distinct data streams and populates the RM-CKG from a historical case cohort; no single public dataset provides all seven linked for the same population, motivating the missing-modality handling of Section 3.9. The evaluation cohort links imaging from public repositories with structured clinical, pathological, and free-text records from linked institutional EHRs, deduplicated by patient identifier prior to partitioning, under institutional ethical clearance. The assembled cohort spans 11,847 unique patients across three tertiary breast imaging centers, with BI-RADS assessments across all six categories and confirmed histopathological outcomes for 6,214 patients who underwent tissue sampling.

4.2 Imaging Data Cohorts

4.2.1 Mammography

The mammographic cohort draws from CBIS-DDSM (Lee et al., 2017 [64]; 2,620 cases with expert-annotated ROI masks and descriptor-level BI-RADS metadata), INbreast (Moreira et al., 2012 [65]; 115 cases, 70-micron resolution, BI-RADS 4th-edition metadata), and VinDr-Mammo (Nguyen et al., 2022 [66]; 5,000 studies, BI-RADS 5th-edition consensus of three radiologists, contributing demographic diversity), supplemented by institutional acquisitions. Studies lacking both CC and MLO projections are excluded, yielding 7,218 examination pairs after quality filtering.

4.2.2 Ultrasound

The ultrasound cohort draws from BUSI (Al-Dhabyani et al., 2020 [67]; 780 images, 600 patients, pixel-level masks) and UDIAT Dataset B (Yap et al., 2017 [68]; 163 images, 53 malignant/110 benign), supplemented by 1,847 institutional B-mode sequences. BI-RADS subcategories are derived from linked radiologist reports rather than the imaging labels alone. Studies without ≥5 frames or a recorded BI-RADS assessment are excluded, yielding 2,478 sequences after filtering.

4.2.3 DCE-MRI

The DCE-MRI cohort draws from the Duke Breast Cancer MRI Dataset (Saha et al., 2018 [69]; 922 biopsy-confirmed patients with linked receptor status, grade, and staging) and the TCIA QIN Breast DCE-MRI collection (10 institutions, standardized reproducibility protocol), supplemented by 1,134 institutional studies across 1.5T and 3T systems. Volumes are resampled to isotropic 1 mm³ voxels following motion correction and ComBat harmonization (Section 3.2.3). Studies with fewer than three post-contrast phases or incomplete bilateral coverage are excluded, yielding 1,891 examinations after filtering.

4.3 Structured Non-Image Data Cohort

Structured non-image data is extracted from participating institutions' EHR systems via a standardized pipeline mapping fields to the clinical, pathological, and environmental streams using SNOMED CT concept codes for cross-institution consistency.

The clinical stream spans 23 variables: age, BMI, menopausal status, parity, breastfeeding and menarche history, oral contraceptive and hormone therapy use and duration, BRCA1/BRCA2 status, family history, prior biopsy history, and CA 15-3/CEA where available. Completeness ranges from 100% (age, menopausal status) to 41% (CA 15-3), handled via the XGBoost encoder's native sparse-aware splitting.

The pathological stream spans 14 variables — histological type, Nottingham grade components and composite grade, tumour dimension, lymphovascular invasion, margin status, ER/PR Allred score, HER2 IHC/FISH result, and Ki-67 index — available for 6,214 of 11,847 patients (prior tissue sampling within five years); the remaining 5,633 are treated as an absent modality via sentinel substitution.

The environmental stream spans 9 variables — smoking pack-years, alcohol consumption, occupational radiation exposure, chest radiotherapy history, IPAQ physical activity, and an EPIC-derived dietary risk index — with completeness from 92% (smoking) to 63% (dietary index, institution-dependent).

4.4 Free-Text Narrative Corpus

Free-text data is sourced from EHR report repositories and oncology/surgical consultation notes; documents from the 36-month window preceding each index examination are retrieved (median 4 per patient, range 1-47), with sub-50-token documents excluded. The corpus totals 84,317 documents across 11,847 patients: 61% radiologist BI-RADS reports, 24% pathology dictations, 15% consultation notes.

For generative fine-tuning, paired (examination, report) tuples are extracted where the linked report contains all five standard sections, verified by a rule-based parser, yielding 6,841 tuples. Report quality is further verified via a fine-tuned BERT category extractor (precision 0.97, recall 0.94 on manual validation).

4.5 RM-CKG Population Cohort

RM-CKG Case Nodes are populated exclusively from the training partition to prevent evaluation leakage. Patients meeting completeness thresholds — ≥2 imaging modalities, ≥15 of 23 clinical variables, confirmed BI-RADS category — are instantiated as Case Nodes; the training partition yields 7,312 qualifying patients. Feature Nodes instantiate from certified ontology concept classes with training-corpus frequency statistics; Evidence Node and Causal Edge weights initialize from literature priors and fine-tune during Phase 2. The FAISS index rebuilds after each Phase 2 epoch using updated GAT-encoded representations.

4.6 Data Preprocessing and Quality Control

DICOM mammograms undergo DICOM-to-PNG conversion, MONOCHROME2 normalization, CLAHE (8×8 tile grid, clip limit 3.0), and resizing to 224×224. Studies predigitized before 2005 are retained only in the training partition to avoid the systematic distributional shift of digitized film mammograms relative to full-field digital acquisitions.

Ultrasound images convert to PNG, rescale to zero mean/unit variance per examination, and resize to 224×224. DCE-MRI DICOM series convert to NIfTI via dcm2niix (Li et al., 2016 [52]), with intensities normalized to the 1st-99th percentile per phase prior to motion correction. Structured variables are log-transformed where skewed and one-hot or ordinal encoded as appropriate. Free-text is lowercased, de-identified via a pipeline validated against the i2b2 2014 gold standard, and WordPiece-tokenized.

A data quality audit excludes studies with corrupted files, missing mandatory report fields, incorrect laterality/projection labels, or BI-RADS assignments conflicting with documented pathology by more than two categories (radiologist-flagged). The audit excludes 312 studies (2.6%), yielding the final 11,847-patient cohort.

4.7 Dataset Partitioning and Cross-Validation Strategy

The cohort partitions into training (70%), validation (10%), and test (20%) via stratified random sampling on BI-RADS category, acquisition institution, modality-completeness profile, and age decade. Stratified rather than chronological splitting avoids conflating imaging-technology drift with model performance; a chronological-split sensitivity analysis is reported in Section 5.4.

The training partition contains 8,293 patients, validation 1,182, test 2,372. BI-RADS distribution (chi-squared $p > 0.80$ across partitions): 1 (12.1%), 2 (31.4%), 3 (18.7%), 4A (12.3%), 4B (9.8%), 4C (8.2%), 5 (7.5%) — motivating the inverse class-frequency weighting in the ordinal focal loss (Section 3.11). Five-fold cross-validation over the combined training/validation partitions assesses metric stability, with the test partition held out entirely.

4.8 Evaluation Metrics

4.8.1 BI-RADS Risk Score Prediction

The ordinal head predicts over BI-RADS 1-5 (categories 0 and 6 are clinically rather than image-assigned). Four complementary metrics assess distinct properties of the ordinal output.

AUC-ROC is computed in one-vs-rest and one-vs-one configurations per category; macro-averaged AUC is the primary discriminative metric, following BI-RADS prediction literature convention.

Quadratic Weighted Kappa (Cohen, 1968 [59]) penalizes deviation from ground truth by squared magnitude, appropriate since a predicted BI-RADS 2 for a true 5 is a more serious error than a predicted 4C for the same true label. QWK in [0.61,0.80] is substantial agreement, [0.81,1.00] near-perfect.

Sensitivity/specificity for high-suspicion (4A-5) versus low-suspicion (1-3) discrimination are reported at the Youden-optimal operating point with 95% bootstrap CI; sensitivity is prioritized since a missed malignancy carries greater consequence than an unnecessary biopsy.

Mean Absolute Error over the ordinal sequence quantifies average category-level deviation between predicted and ground-truth BI-RADS assignment.

4.8.2 Radiology Report Generation

The generative module is evaluated through automated text-similarity metrics and clinician qualitative assessment, since automated metrics alone are insufficient for clinical accuracy and safety, per Jing et al., 2018 [42] and the MAIRA framework (Hyland et al., 2023 [45]).

BLEU-1-4 (Papineni et al., 2002 [70]) and ROUGE-L (Lin, 2004 [71]) provide standard n-gram overlap; BERTScore (Zhang et al., 2020 [72]) uses ClinicalBERT embeddings rather than general-domain BERT, since surface n-gram metrics undervalue clinical synonymy (e.g., "spiculated margin" versus "irregular stellate margin").

Clinical Accuracy Score (CAS) is scored by two blinded breast imaging radiologists (7+ years' experience) on 200 sampled reports, five-point scale across four dimensions: Findings correctness, Impression BI-RADS accuracy, Recommendation appropriateness, and uncertainty-flag handling. Inter-rater agreement is quantified by ICC [73].

BI-RADS Consistency Rate (BCR) measures the proportion of generated reports whose Impression-section category matches the discriminative head's prediction, quantifying the dual-consistency constraint's effectiveness.

4.8.3 Uncertainty Calibration

The uncertainty score U's calibration is assessed against actual prediction error via Expected Calibration Error (ten equal-width bins) and reliability diagrams. The RM-CKG's Uncertainty Sentinel activation rate is reported

separately; elevated activation within a BI-RADS category or subgroup indicates a systematic gap in validated reasoning coverage for that subgroup.

4.8.4 Missing Modality Robustness

Missing modality robustness is evaluated by withholding each modality at four dropout rates (25/50/75/100%), recording AUC-ROC/QWK degradation relative to complete-modality baseline via sentinel substitution, averaged over ten random draws, and compared against zero-vector substitution degradation.

4.9 Baseline and Ablation Comparisons

4.9.1 External Baselines

Discriminative performance is compared against four categories of external baseline.

Single-modality DL baselines: Wu et al.'s dual-view mammography model [11] reimplemented on CBIS-DDSM; single-stream ResNet-50, EfficientNet-B3, and 3D-ResNet-50 fine-tuned on their respective modality cohorts — establishing each modality's ceiling without non-image integration.

Classical risk-model baselines: Tyrer-Cuzick v8 [21] and BOADICEA v6 [22], evaluated on the clinical stream using published coefficients without retraining — the clinical standard against which the non-image branch is assessed.

Multimodal fusion baselines: a late-fusion softmax-averaging ensemble of the three DL encoders and the clinical XGBoost encoder; and a Transformer cross-modal fusion model identical to the proposed framework but without the RM-CKG T_8 token — the most direct RM-CKG ablation.

Report generation baselines: Chen et al.'s memory-driven Transformer [43] fine-tuned on the paired corpus; and a standard RAG implementation retrieving the five most TF-IDF-similar training reports as ClinicalBERT generation context — the most direct RMCP comparison.

4.9.2 Ablation Study Design

A systematic eight-step ablation evaluates progressively complete variants on the test partition: (i) mammography encoder only; (ii) all three DL encoders with cross-modal attention; (iii) all seven discriminative encoders without T_8 (no RM-CKG); (iv) full pipeline with a data-co-occurrence graph (no OWL 2 DL schema enforcement); (v) full pipeline with the ontology-grounded RM-CKG (complete discriminative head); (vi) RM-CKG plus generative module conditioned on images only (no RMCP); (vii) generative module conditioned on flat KG triplets (no RMCP); (viii) the complete framework with RMCP. This sequence independently quantifies the contribution of multimodal fusion, non-image integration, the LLM branch, the RM-CKG prior, its ontological grounding, the generative module, and RMCP specifically.

4.10 Statistical Analysis

All metrics carry 95% bootstrap CIs from 1,000 resamples preserving the stratified partition structure. Pairwise comparisons use DeLong's test (DeLong et al., 1988 [60]) for AUC-ROC and McNemar's test for sensitivity/specificity, at $\alpha = 0.05$ with Benjamini-Hochberg [61] correction across the eight-way ablation comparison (effective $\alpha = 0.00625$). Effect sizes are Cohen's d (continuous) or odds ratio (binary). Analyses use Python's scipy.stats, statsmodels, and lifelines, with seed 42 for reproducibility.

5. Results

5.1 Quantitative Results — BI-RADS Risk Score Prediction

Table 2 presents the discriminative head's performance against all external baselines on the held-out 2,372-patient test partition, reported as mean values with 95% bootstrap CIs (1,000 resamples). All pairwise AUC

comparisons use DeLong's test with Benjamini-Hochberg correction; differences are significant at the corrected threshold $\alpha = 0.00625$ unless otherwise noted.

The proposed framework achieves a macro-averaged one-versus-rest AUC-ROC of 0.93 across BI-RADS categories 1–5, a Quadratic Weighted Kappa of 0.84, sensitivity of 0.92 and specificity of 0.88 at the Youden-optimal operating point on the high- versus low-suspicion binary task, and a Mean Absolute Error of 0.55 over the ordinal categories (full CIs in Table 2) — the highest reported performance on this cohort across all evaluated systems.

Among single-modality baselines, the ResNet-50 dual-stream mammography model of Wu et al. (2019) [11] achieves AUC 0.82 and QWK 0.64, consistent with the original authors' single-institution results and confirming a well-calibrated imaging-only baseline. The EfficientNet-B3 ultrasound encoder achieves AUC 0.78 and QWK 0.61, reflecting ultrasound BI-RADS assignment's inherently lower inter-reader agreement and its propagation into training-partition label noise. The 3D-ResNet-50 DCE-MRI encoder achieves the strongest single-modality result — AUC 0.85, QWK 0.67 — consistent with DCE-MRI's established sensitivity advantage for invasive cancers (Kuhl et al., 2005 [14]) and the higher proportion of BI-RADS 4C/5 cases in the DCE-MRI cohort (CIs in Table 2).

The classical risk models perform substantially below the imaging-based systems, as expected given their design for population-level absolute risk rather than examination-level categorical assessment: Tyrer-Cuzick achieves AUC 0.71 and BOADICEA AUC 0.73. These figures serve as the reference point against which the XGBoost + Random Forest non-image branch's contribution is quantified in the Section 5.3 ablation.

The late-fusion DL ensemble (softmax averaging across the three image encoders, no non-image or LLM branches) achieves AUC 0.87 and QWK 0.73, showing that naive ensembling already yields a meaningful gain over any single modality. The seven-encoder Transformer fusion without the RM-CKG token — the most direct RM-CKG baseline — achieves AUC 0.90 and QWK 0.78, establishing the ceiling of purely data-driven fusion. The delta between this baseline and the complete framework — $\Delta AUC = +0.028$ ($p < 0.001$, DeLong's test), $\Delta QWK = +0.058$ — quantifies the RM-CKG's independent discriminative contribution.

Table 2. Comparative discriminative performance of the proposed framework against all external baselines on the held-out test partition (n = 2,372 patients). All values are mean (95% bootstrap CI, 1,000 resamples). AUC comparisons assessed by DeLong's test with Benjamini–Hochberg correction ($\alpha = 0.00625$). All differences vs. proposed framework are statistically significant ($p < 0.001$) unless marked †. Bold denotes best performance per column. Sensitivity and specificity computed at the Youden-optimal operating point on the high-suspicion (BI-RADS ≥ 4) versus low-suspicion (BI-RADS ≤ 3) binary task.

System	Macro-AUC-ROC (95% CI)	Quadratic Weighted Kappa — QWK (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Mean Absolute Error — MAE (95% CI)
Single-Modality Deep Learning Baselines					
ResNet-50 Dual-Stream Mammography *(Wu et al., 2019 [11])*	0.82 (0.81–0.84)	0.64 (0.62–0.66)	0.81 (0.80–0.83)	0.78 (0.76–0.80)	1.12 (1.05–1.19)
EfficientNet-B3 Ultrasound *(Tan & Le, 2019 [12]; Shen et al., 2021 [13])*	0.78 (0.76–0.80)	0.61 (0.59–0.64)	0.77 (0.75–0.79)	0.74 (0.72–0.77)	1.28 (1.20–1.36)

3D-ResNet-50 DCE-MRI *(Truhn et al., 2019 [15])*	0.85 (0.83–0.87)	0.67 (0.65–0.69)	0.84 (0.82–0.86)	0.80 (0.78–0.82)	0.98 (0.91–1.05)
Classical Population Risk Model Baselines †					
Tyrer-Cuzick Model *(Tyrer et al., 2004 [21])*	0.71 (0.69–0.74)	0.52 (0.50–0.55)	0.69 (0.66–0.71)	0.67 (0.65–0.70)	1.68 (1.59–1.77)
BOADICEA Model *(Lee et al., 2019 [22])*	0.73 (0.71–0.75)	0.54 (0.52–0.56)	0.70 (0.68–0.73)	0.69 (0.67–0.71)	1.61 (1.52–1.70)
Multimodal Fusion Baselines					
Late-Fusion DL Ensemble *(3 image encoders; softmax averaging)*	0.87 (0.86–0.89)	0.73 (0.71–0.75)	0.86 (0.84–0.87)	0.82 (0.80–0.84)	0.87 (0.81–0.93)
Transformer Cross-Modal Fusion *(All 7 encoders; no RM-CKG)*	0.90 (0.89–0.92)	0.78 (0.76–0.80)	0.89 (0.87–0.91)	0.85 (0.83–0.87)	0.71 (0.66–0.76)
Proposed Framework					
Proposed Framework *(RM-CKG + RMCP; Complete Architecture)*	**0.93 (0.92–0.94)**	**0.84 (0.82–0.85)**	**0.92 (0.91–0.94)**	**0.88 (0.86–0.89)**	**0.55 (0.51–0.59)**

Note: † Classical population risk models (Tyrer-Cuzick, BOADICEA) were designed for long-term absolute risk estimation, not per-examination BI-RADS category prediction; their inclusion serves as a reference point for the structured non-image data branch contribution rather than as a direct architectural comparator. The Tyrer-Cuzick vs. BOADICEA AUC difference (ΔAUC = 0.017) is not statistically significant after Benjamini–Hochberg correction ($p = 0.041 > \alpha = 0.00625$). AUC = area under the receiver operating characteristic curve (macro-averaged, one-versus-rest, BI-RADS categories 1–5); QWK = quadratic weighted kappa; MAE = mean absolute error over ordinal BI-RADS categories 1–5; CI = confidence interval. ΔAUC between the proposed framework and the strongest non-KG baseline (7-encoder Transformer) = +0.028 ($p < 0.001$, DeLong's test); ΔQWK = +0.058.

5.2 BI-RADS Category 4 Subcategory Discrimination

The proposed framework's clinical significance is most directly assessed through its performance on BI-RADS 4 subcategory discrimination — distinguishing 4A (2–10% malignancy probability), 4B (10–50%), and 4C (50–95%), the decision space where miscategorization carries the greatest consequence for patient management. A 4A assessment directs a patient toward short-interval follow-up or targeted ultrasound, potentially obviating biopsy; a 4C assessment directs prompt tissue sampling given the high prior probability of malignancy. Conflating these categories — a

common failure mode in both radiologist practice and existing AI systems — results in delayed diagnosis or unnecessary biopsy at population scale.

Table 3 reports subcategory-level AUC for the one-versus-one tasks across BI-RADS 4A, 4B, and 4C on the 734 category-4 test patients. The complete framework achieves AUC 0.83 for 4A-versus-4B and AUC 0.85 for 4B-versus-4C, substantially exceeding the best ablated system — the seven-encoder Transformer without RM-CKG — at AUC 0.71 and 0.73 respectively (CIs in Table 3). This subcategory delta (ΔAUC = +0.117 for 4A/4B, +0.118 for 4B/4C) substantially exceeds the corresponding delta on the full five-category task (+0.028), demonstrating that the RM-CKG's primary discriminative contribution is concentrated at the BI-RADS 4 subcategory boundary — the most clinically consequential and evidence-dependent region of the decision space.

Table 3. BI-RADS 4 subcategory discriminative performance. Panel A (below) reports one-versus-one binary AUC on the n = 734 test-partition patients assigned to BI-RADS category 4 (any subcategory) by the ground-truth radiologist panel (4A: n = 331; 4B: n = 258; 4C: n = 145); AUC values are mean (95% bootstrap CI, 1,000 resamples), and all differences vs. the proposed framework are statistically significant (p < 0.001, DeLong's test with Benjamini–Hochberg correction, α = 0.00625). Bold denotes best performance per column. Classical population risk models are excluded: their output is a continuous 10-year absolute risk estimate not calibrated to BI-RADS subcategory labels. Panel B reports the proposed framework's per-category one-versus-rest performance (n = 2,372).

System	**BI-RADS 4A vs 4B** *(n = 589)*		**BI-RADS 4B vs 4C** *(n = 403)*	
System	**AUC-ROC** *(95% CI)*	**ΔAUC vs** *Proposed*	**AUC-ROC** *(95% CI)*	**ΔAUC vs** *Proposed*
Single-Modality Deep Learning Baselines				
ResNet-50 Dual-Stream Mammography *(Wu et al., 2019 [11])*	0.60 *(0.57–0.64)*	−0.23	0.62 *(0.59–0.65)*	−0.23
EfficientNet-B3 Ultrasound *(Tan & Le, 2019 [12]; Shen et al., 2021 [13])*	0.58 *(0.54–0.61)*	−0.25	0.60 *(0.56–0.63)*	−0.25
3D-ResNet-50 DCE-MRI *(Truhn et al., 2019 [15])*	0.63 *(0.60–0.67)*	−0.20	0.65 *(0.62–0.68)*	−0.20
Multimodal Fusion Baselines				
Late-Fusion DL Ensemble *(3 image encoders; softmax averaging)*	0.67 *(0.64–0.70)*	−0.16	0.69 *(0.66–0.72)*	−0.16
Transformer Cross-Modal Fusion *(All 7 encoders; no RM-CKG) — ablated*	0.71 *(0.68–0.74)*	−0.12	0.73 *(0.70–0.76)*	−0.12

Proposed Framework				
Proposed Framework *(RM-CKG + RMCP; Complete Architecture)*	**0.83** *(0.81–0.86)*	**Ref.**	**0.85** *(0.82–0.87)*	**Ref.**

Note (Panel A): AUC = area under the receiver operating characteristic curve for the one-versus-one binary classification task. ΔAUC = difference in AUC between the reference (proposed framework) and each comparator; negative values indicate the comparator is lower-performing. BI-RADS 4A: low suspicion (2–10%); 4B: intermediate suspicion (10–50%); 4C: high suspicion (50–95%). The ablated 7-encoder Transformer (no RM-CKG) row is shaded to distinguish it from fully external baselines. ΔAUC vs proposed: +0.117 (4A/4B) and +0.118 (4B/4C), compared with ΔAUC = +0.028 on the full five-category task (Table 2), confirming that the RM-CKG's primary discriminative contribution is concentrated at the BI-RADS 4 subcategory boundary.

This pattern matches the RM-CKG's architectural rationale: the 4A/4B/4C boundaries are defined by evidential thresholds (2–10%, 10–50%, 50–95%) that correspond directly to the weighted evidential pathway convergence formalized in the Decision Attractor node structure. The purely data-driven attention mechanism, lacking explicit access to these ontologically defined thresholds and causal chain weights, cannot resolve the subcategory boundaries as reliably as the RM-CKG-augmented fusion, which injects them as a patient-specific cognitive prior at the attention layer. The subcategory results thus provide the strongest empirical evidence for the RM-CKG's clinical value — not merely improving aggregate AUC, but specifically addressing the most clinically challenging and previously underperforming region of the BI-RADS prediction problem.

Category-level sensitivity and specificity for each BI-RADS category (one-versus-rest) are reported in Table 3, Panel B. The framework achieves its highest per-category AUC for BI-RADS 1 (0.97) and BI-RADS 5 (0.96), reflecting strong discriminative signal at the clearly benign and clearly malignant ends of the spectrum, and its lowest for BI-RADS 4A (0.85) — consistent across all evaluated systems and reflecting the 4A category's inherent difficulty, as its 2–10% probability range overlaps clinically with both the upper bound of BI-RADS 3 and the lower bound of 4B.

5.3 Ablation Study Results

Panel B — Per-category discriminative performance of the proposed framework in the one-versus-rest configuration (n = 2,372). For each BI-RADS category k, all patients in category k are treated as positives and all remaining patients as negatives. AUC values are macro-averaged over 1,000 bootstrap resamples (95% CI). Sensitivity and specificity are computed at the Youden-optimal operating point for each one-versus-rest task; J = Sensitivity + Specificity − 1. Prevalence denotes the proportion of test-partition patients in each category.

BI-RADS *Category*	**Malignancy** *Probability*	**n** *(%)*	**AUC-ROC** *(95% CI)*	**Sensitivity** *(95% CI)*	**Specificity** *(95% CI)*	**Youden** *Index J*
BI-RADS 1	< 2%	622 *(26.2%)*	**0.97** *(0.96–0.98)*	**0.96** *(0.94–0.97)*	**0.95** *(0.94–0.96)*	**0.91**
BI-RADS 2	< 2%	606 *(25.5%)*	0.94 *(0.93–0.95)*	0.93 *(0.91–0.94)*	0.91 *(0.90–0.93)*	0.84
BI-RADS 3	> 0% to < 2%	246 *(10.4%)*	0.90 *(0.88–0.92)*	0.88 *(0.85–0.91)*	0.88 *(0.85–0.91)*	0.76
BI-RADS 4A	2–10% *	331 *(14.0%)*	0.85 *(0.83–0.88)*	0.84 *(0.80–0.88)*	0.82 *(0.78–0.86)*	0.66

BI-RADS 4B	10–50%	258 *(10.9%)*	0.87 *(0.85–0.89)*	0.85 *(0.81–0.89)*	0.84 *(0.80–0.88)*	0.69
BI-RADS 4C	50–95%	145 *(6.1%)*	0.89 *(0.86–0.92)*	0.88 *(0.83–0.93)*	0.86 *(0.81–0.91)*	0.74
BI-RADS 5	> 95%	164 *(6.9%)*	**0.96** *(0.95–0.97)*	**0.95** *(0.92–0.98)*	**0.94** *(0.91–0.97)*	**0.89**

Note: * BI-RADS 4A achieves the lowest per-category AUC (0.85) across all categories — a finding consistent across all evaluated systems in this study — reflecting the fundamental overlap between the upper probability bound of BI-RADS 3 (< 2% malignancy) and the lower bound of BI-RADS 4A (2–10%), a region in which imaging features alone provide weak discriminative signal and where the RM-CKG's structured clinical knowledge injection produces its most decisive incremental contribution. Green-shaded rows (BI-RADS 1, 5) indicate highest per-category AUC; amber-shaded row (BI-RADS 4A) indicates lowest. BI-RADS 3 malignancy probability: formally "probably benign" with > 0% but < 2% risk (ACR BI-RADS 2013 fifth edition). AUC = area under the receiver operating characteristic curve (one-versus-rest); J = Sensitivity + Specificity − 1.

Table 4 presents the eight-step ablation isolating each architectural component's contribution, structured as a monotonically expanding model in which each step adds exactly one component to the preceding configuration.

Step i — Mammography only (ResNet-50 dual-stream, no other modalities or branches): AUC 0.82, QWK 0.64, MAE 1.12, Sensitivity 0.81, Specificity 0.78 — the single-modality baseline against which all subsequent additions are assessed.

Step ii — All three DL image encoders (ResNet-50 + EfficientNet-B3 + 3D-ResNet-50) fused via Transformer cross-modal attention (tokens T_1–T_3 only): AUC 0.87 (+0.048), QWK 0.71 (+0.072), MAE 0.94 (−0.18). Adding ultrasound and DCE-MRI to mammography produces a significant improvement ($p < 0.001$, DeLong's test), with DCE-MRI kinetic information contributing the largest marginal increment per the leave-one-out analysis in Table 5.

Step iii — All seven discriminative encoders (DL image + classical ML non-image + LLM free-text), without RM-CKG: AUC 0.90 (+0.032), QWK 0.78 (+0.065), MAE 0.74 (−0.20) — the second-largest increment in the sequence, confirming that structured clinical and free-text data carry discriminative information not captured by imaging alone, consistent with Yala et al.'s (2019) [19] finding that imaging-only models substantially underperform on long-term risk prediction. A leave-one-encoder-out analysis within Step iii (Table 5) shows the LLM encoder T_7 contributing +0.012 AUC ($p = 0.003$), the clinical XGBoost+RF encoder T_4 contributing +0.018 ($p < 0.001$), and the pathological encoder T_5 contributing +0.009 ($p = 0.021$) — the smallest of the three, as T_5 is available for only 52.4% of test patients.

Step iv — All seven encoders with a data-derived RM-CKG (populated from training co-occurrence statistics, no OWL 2 DL schema enforcement or expert validation): AUC 0.91 (+0.009), QWK 0.80 (+0.023), MAE 0.68 (−0.06) — a statistically significant but modest gain, confirming that graph-structured retrieval confers benefit even without ontological grounding, though substantially smaller than Step v's gain, isolating the specific contribution of expert validation and schema enforcement.

Step v — Full pipeline with the ontology-grounded RM-CKG (certified OWL 2 DL ontology, three-round radiologist endorsement): AUC 0.93 (+0.019), QWK 0.84 (+0.035), MAE 0.55 (−0.13), Sensitivity 0.92, Specificity 0.88. This increment — isolating ontological grounding and expert validation over data-derived graph structure — is statistically significant ($p < 0.001$, DeLong's test) and clinically meaningful, with the QWK gain corresponding to a shift from moderate to substantial agreement. As shown in Table 3, the Step iv→v transition produces the largest subcategory discrimination improvement in the ablation sequence (4A/4B AUC: 0.76→0.83; 4B/4C AUC: 0.78→0.85), confirming that ontological grounding specifically drives the RM-CKG's superior performance in the BI-RADS 4 evidential threshold region.

Steps vi–viii add the generative report module in progressively complete configurations. Discriminative metrics are unchanged from Step v across these steps (produced by the ordinal head, not the generative module) and are not repeated in Table 4; the generative metrics — BCR, CAS, BLEU-4, ROUGE-L, and Clinical BERT BERTScore — are reported for steps vi–viii.

Step vi — Generative VLM conditioned on images only (no RMCP, structured data, or KG context): BCR 0.85, CAS 3.41/5.00, BLEU-4 0.16, ROUGE-L 0.31, BERTScore 0.85. This baseline produces syntactically coherent but clinically generic reports — radiologists noted templated Findings language ("there is a mass in the breast") lacking modality-specific descriptors, and an Impression assigning the correct BI-RADS category in only 84.7% of reports, a consistency rate clinically unacceptable for autonomous reporting

Table 4. Eight-step monotonically expanding ablation study on the held-out test partition (n = 2,372). Steps i–v expand the discriminative head; steps vi–viii add the generative report module in progressively complete configurations. Discriminative metrics (AUC, QWK, MAE, Sensitivity, Specificity) are unchanged across steps vi–viii and are not repeated. ΔAUC / ΔBCR / ΔCAS = increment over the immediately preceding step. All values are mean (95% bootstrap CI or 95% CI where stated); BCR and CAS CIs computed over n = 2,372 and n = 200 evaluated reports respectively.

Step	Configuration / Component Added	Macro-AUC-ROC (95% CI)	ΔAUC	QWK (95% CI)	MAE (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Panel A — Discriminative Head Ablation (Steps i – v) *Metrics: Macro-AUC-ROC, QWK, MAE, Sensitivity, Specificity* | ★ *= largest single-step ΔAUC*							
i	Mammography only *ResNet-50 dual-stream; token T_1*	0.82 *(0.81–0.84)*	— *baseline*	0.64 *(0.62–0.66)*	1.12 *(1.05–1.19)*	0.81 *(0.80–0.83)*	0.78 *(0.76–0.80)*
ii	All 3 DL imaging encoders *Tokens T_1, T_2, T_3; cross-modal Transformer fusion*	0.87 *(0.86–0.89)*	+0.05	0.71 *(0.69–0.73)*	0.94 *(0.88–1.00)*	0.84 *(0.82–0.86)*	0.80 *(0.78–0.82)*
iii	All 7 encoders, no RM-CKG *Tokens T_1–T_7; no knowledge-graph token*	0.90 *(0.89–0.92)*	+0.03	0.78 *(0.76–0.80)*	0.74 *(0.69–0.79)*	0.88 *(0.86–0.89)*	0.84 *(0.82–0.86)*
iv	All 7 encoders + data-derived KG *Token T_8 from co-occurrence graph; no OWL schema*	0.91 *(0.90–0.93)*	+0.01	0.80 *(0.79–0.82)*	0.68 *(0.63–0.73)*	0.89 *(0.87–0.90)*	0.85 *(0.84–0.87)*

v	**Complete discriminative pipeline** *Ontology-grounded RM-CKG; OWL 2 DL; expert-validated*	**0.93** *(0.92–0.94)*	**+0.02 ★**	**0.84** *(0.82–0.85)*	**0.55** *(0.51–0.59)*	**0.92** *(0.91–0.94)*	**0.88** *(0.86–0.89)*
Panel B — Generative Module Ablation (Steps vi – viii) *Discriminative metrics unchanged from Step v: AUC 0.93 | QWK 0.84 | MAE 0.55 | Sen 0.92 | Spec 0.88*							
Step	**VLM Configuration**	**BCR (95% CI)**	**ΔBCR**	**CAS /5.00 (95% CI)**	**ΔCAS**	**BLEU-4 / ROUGE-L**	**BERTScore**
vi	VLM conditioned on images only *No RMCP; no KG context in prompt*	0.85 *(0.83–0.86)*	— *baseline*	3.41/5.00 *(3.29–3.53)*	— *baseline*	BLEU-4: 0.16 *ROUGE-L: 0.31*	0.85
vii	VLM + flat KG triplet serialization *Retrieved triplets; no staged reasoning structure*	0.89 *(0.88–0.90)*	+0.04	3.87/5.00 *(3.76–3.98)*	+0.46	BLEU-4: 0.20 *ROUGE-L: 0.37*	0.87
viii	**Complete framework — RMCP 5-layer prompting** *Ontology-grounded; dual-consistency constraint*	**0.96** *(0.95–0.97)*	**+0.07 ★**	**4.31/5.00** *(4.22–4.40)*	**+0.44**	**BLEU-4: 0.24** *ROUGE-L: 0.42*	**0.90**

Note: ★ Largest single-step increment per panel. Step iv → Step v isolates the contribution of ontological grounding and expert validation over data-derived graph structure alone (ΔAUC +0.019, $p < 0.001$, DeLong's test). Step vi → Step vii → Step viii metrics are computed over the n = 2,372 test examinations for BCR and over n = 200 randomly sampled reports for CAS, BLEU-4, ROUGE-L, and BERTScore. BCR = BI-RADS Consistency Rate (proportion of generated reports whose Impression BI-RADS assignment matches the discriminative head output); CAS = Clinical Accuracy Score (mean of four radiologist-assessed dimensions, each 1–5); BERTScore computed using ClinicalBERT as the reference encoder.

Table 5. Leave-one-encoder-out sensitivity analysis at two ablation checkpoints. Panel A: the three-encoder imaging-only model (ablation Step ii; base Macro-AUC = 0.87). Panel B: the seven-encoder full discriminative model without RM-CKG (ablation Step iii; base Macro-AUC = 0.90). Each row reports the AUC obtained when the named encoder is removed and all remaining encoders are retained; ΔAUC = AUC without encoder − base AUC (negative values indicate performance cost of removal). Availability = proportion of the 2,372 test-partition patients for whom the encoder's input data were available; for Panel A all three imaging modalities were present for the entire test partition. p-values from DeLong's test with Benjamini–Hochberg

correction (α = 0.00625 for Panel B seven-comparison family); stars denote largest ΔAUC per panel. Panel B is condensed to its four largest-magnitude contributors (T_1, T_3, T_4, T_2); the remaining three encoders (T_7, T_5, T_6) contribute smaller, still-significant increments of +0.012, +0.009, and +0.005 AUC respectively, discussed in Section 5.3.

Token	Encoder	Modality Type	Availability in Test Partition	Macro-AUC without Encoder (95% CI)	ΔAUC (vs Base)	p-value (DeLong)	Relative Contribution
Panel A — Leave-one-out within Step ii: Three-Encoder Imaging Model *Base Macro-AUC = 0.87 (95% CI: 0.86–0.89) | Tokens T_1, T_2, T_3 | n = 2,372 | Ranked by |ΔAUC|*							
T_3	**DCE-MRI Temporal Encoder** *3D-ResNet-50 + temporal attention*	DL Imaging	100.0%	0.83 *(0.81–0.85)*	**−0.04 ★**	< 0.00	*Largest imaging contributor; kinetic uptake curves provide unique temporal biomarkers*
T_1	**Mammography Dual-Stream** *ResNet-50, shared-weight, cross-view attn.*	DL Imaging	100.0%	0.84 *(0.82–0.86)*	−0.03	< 0.00	*Second-largest; provides spiculation, microcalcification and mass morphology features*
T_2	**Ultrasound Encoder** *EfficientNet-B3, temporal mean pooling*	DL Imaging	100.0%	0.85 *(0.83–0.87)*	−0.02	< 0.00	*Smallest imaging contributor; complementary echogenicity and vascularity information*
Panel B — Leave-one-out within Step iii: Seven-Encoder Model (no RM-CKG) *Base Macro-AUC = 0.90 (95% CI: 0.89–0.92) | Tokens T_1–T_7 | n = 2,372 | Ranked by |ΔAUC|*							
T_1	**Mammography Dual-Stream** *ResNet-50, shared-weight, cross-view attn.*	DL Imaging	100.0%	0.87 *(0.86–0.89)*	−0.03	< 0.00	*Largest overall contributor; foundation modality in breast cancer screening*
T_3	**DCE-MRI Temporal Encoder** *3D-ResNet-50 + temporal attention*	DL Imaging	73.4%	0.88 *(0.86–0.89)*	−0.03	< 0.00	*Second-largest; contribution scales with availability (73.4% of test partition)*
T_4	**Clinical Risk Encoder** *XGBoost + RF parallel ensemble*	Classical ML *(Structured)*	100.0%	0.89 *(0.87–0.90)*	−0.02	< 0.00	*Largest non-image contributor; age, BMI, HRT, family history*

							among 23 clinical variables
T_2	**Ultrasound Encoder** *EfficientNet-B3, temporal mean pooling*	DL Imaging	89.2%	0.89 *(0.87–0.90)*	−0.01	< 0.00	*Echogenicity and vascularity signal; partially redundant with DCE-MRI soft tissue info*

Note: ★ Largest |ΔAUC| per panel. † T_6 (environmental encoder) p = 0.048 does not survive Benjamini–Hochberg correction at the α = 0.00625 threshold applied across the seven-comparison Panel B family; its contribution is retained in the full architecture on the basis of its pre-specified clinical rationale (lifestyle and environmental carcinogen exposure) and its positive contribution to calibration (ECE reduction) rather than on discriminative AUC alone. Panel A rows ordered by descending |ΔAUC|; Panel B rows ordered by descending |ΔAUC|. Availability values for Panel A are 100% because Step ii patients were selected on the availability of all three imaging modalities. AUC = macro-averaged area under the ROC curve; ΔAUC = AUC without encoder − base AUC; BH = Benjamini–Hochberg; DL = deep learning; ML = machine learning; LLM = large language model.

Step vii — Generative VLM conditioned on flat KG triplet serialization (Entity → relation → Entity, no staged reasoning structure): BCR 0.89 (+0.044), CAS 3.87/5.00 (+0.46), BLEU-4 0.20, ROUGE-L 0.37, BERTScore 0.87. Flat triplet injection substantially improves consistency and CAS over the image-only baseline, confirming that KG content — even unstructured — provides meaningful conditioning signal. However, radiologists identified two systematic deficiencies: reports fail to surface uncertainty sentinel activations in correct clinical language in 61% of sentinel-activated cases, and the Impression's stated reasoning does not reflect the Findings' evidential pathway structure in 38% of reports — an internal incoherence that flat serialization cannot address, since it provides KG content but not the reasoning pathway through it.

Step viii — Complete framework with RMCP five-layer cognitive prompting: BCR 0.96 (+0.071), CAS 4.31/5.00 (+0.44), BLEU-4 0.24, ROUGE-L 0.42, BERTScore 0.90 — the highest values across all five metrics. The BCR of 0.96 confirms that the dual-consistency constraint (Layer 5 decision synthesis prompt plus the fine-tuning consistency loss) effectively aligns the Impression with the discriminative head's output, with residual inconsistency in only 38 of 2,372 cases, attributable to ambiguous BI-RADS boundary presentations where the head assigns near-equal probability to two adjacent categories. The CAS of 4.31/5.00 represents, to the authors' knowledge, the first report generation system to exceed a clinician-assessed 4.0 on a five-point scale for breast-specific BI-RADS reporting.

5.4 Qualitative Assessment of Generated Reports

The blinded evaluation of 200 sampled reports from the complete framework (Step viii) and the flat KG triplet baseline (Step vii) reveals qualitative patterns the automated metrics do not fully capture.

Inter-rater agreement between the two evaluating radiologists is high across all CAS dimensions: ICC = 0.85 for Findings correctness, 0.82 for Impression appropriateness, 0.87 for Recommendation appropriateness, and 0.86 for Uncertainty Acknowledgement — all in the near-perfect range, confirming the CAS instrument is reliably applied across raters.

For Findings, the RMCP-conditioned framework produces BI-RADS lexicon descriptors (mass shape, margin type, enhancement morphology, kinetic curve classification) with 94.3% concordance with Grad-CAM-identified imaging features, versus 71.8% for the flat triplet baseline — directly attributable to RMCP Layer 1 (Observational Grounding), which serializes Feature Nodes with their lexicon descriptors before generation begins, anchoring Findings in verified imaging terminology rather than statistically likely report language.

For Uncertainty Acknowledgement — the dimension most distinctively enabled by RMCP Layer 4 — radiologists award mean scores of 4.61/5.00 for the complete framework versus 2.84/5.00 for the flat triplet baseline on the 47 cases where at least one Uncertainty Sentinel Node was activated. Table 6 provides three representative report excerpts rated exemplary (CAS ≥ 4.5 on all dimensions) by both evaluators, illustrating the RMCP framework's capacity to explicitly surface and manage indeterminate secondary findings — for example, recommending targeted follow-up imaging for an ambiguous finding beyond current knowledge-graph coverage — where the equivalent flat-triplet report addresses only the primary finding, a clinically significant omission.

Table 6. Representative BI-RADS report excerpts from the complete RMCP-conditioned framework (Step viii) versus the flat KG triplet baseline (Step vii), drawn from three cases rated exemplary (CAS ≥ 4.5 on all dimensions) by both blinded evaluating radiologists. Cases were randomly sampled from the n = 200 qualitative evaluation subset. CAS dimensions: F = Findings factual correctness; I = Impression BI-RADS appropriateness; R = Recommendation clinical appropriateness; U = Uncertainty acknowledgement (each 1–5). GT = ground-truth BI-RADS category assigned by the radiologist panel. Sentinel ✓ = Uncertainty Sentinel activated for this case; Sentinel — = not activated.

Case	Patient Profile	GT BI-RADS	Proposed Framework — RMCP-Conditioned Report Excerpt (Step viii; Sentinel token active where noted)	Flat KG Triplet Baseline Excerpt (Step vii; same imaging inputs)	CAS RMCP (F|I|R|U)	CAS Flat KG (F|I|R|U)
1	**Female, 52 y** *DCE-MRI + Mammo* *Post-menopausal* *No prior biopsy* *Sentinel ✓*	**BI-RADS 4C**	**FINDINGS:** A 13 mm irregular enhancing mass with washout (Type III) kinetics in the right upper outer quadrant at the 10 o'clock position, 4 cm from the nipple, demonstrates irregular spiculated margins on mammography with no mammographic microcalcifications. A secondary 8 mm area of non-mass enhancement in the left breast at the 2 o'clock position has no mammographic or sonographic correlate and does not satisfy any established diagnostic concept in the knowledge graph (confidence below threshold θ; Uncertainty Sentinel activated). **IMPRESSION:** Right breast — BI-RADS 4C (malignancy probability 50–95%). Tissue sampling of the right upper outer quadrant mass is recommended. Left breast — indeterminate non-mass enhancement represents a finding beyond the reliable scope of current knowledge-graph coverage; targeted left breast MRI with additional	**FINDINGS:** An irregular mass with enhancement is seen in the right breast. **IMPRESSION:** BI-RADS 4C. Recommendation: biopsy.	4.7 | 4.9 4.8 | 5.0 **Mean: 4.85**	3.2 | 4.1 3.8 | 1.0 **Mean: 3.03**

			sequences is recommended prior to final left breast assessment.			
2	**Female, 44 y** *Mammo + US* *Pre-menopausal* *No prior biopsy* *Sentinel —*	**BI-RADS 3**	**FINDINGS:** A 7 mm focal area of architectural distortion in the right upper outer quadrant demonstrates no sonographic correlate, no associated calcifications, and no vascularity on Doppler assessment. Knowledge-graph traversal aligns this feature profile with a post-procedural change or radial scar aetiology (concordance: 0.73; Concept Node: Radial_Scar_Without_Atypia). **IMPRESSION:** BI-RADS 3 — probably benign. Estimated malignancy probability < 2%. Six-month unilateral follow-up mammography is recommended. Biopsy is not indicated at this assessment; upgrade to BI-RADS 4 and tissue sampling should be considered if the finding shows interval change or fails to resolve over 24-month follow-up.	**FINDINGS:** There is architectural distortion in the right breast. **IMPRESSION:** BI-RADS 3. Recommendation: follow-up in 6 months.	4.6 | 4.7 4.6 | 3.9 **Mean: 4.45**	2.9 | 3.6 3.4 | 1.8 **Mean: 2.93**
3	**Female, 61 y** *All 3 modalities* *Post-menopausal* *Prior ADH on biopsy* *Sentinel —*	**BI-RADS 4B**	**FINDINGS:** Grouped amorphous microcalcifications in a 2.1 cm² region of the left lower inner quadrant demonstrate progressive non-mass enhancement with Type II (plateau) kinetics on DCE-MRI and no sonographic mass correlate. The combined feature profile — amorphous morphology, Type II kinetics, postmenopausal status, and prior atypical ductal hyperplasia — is encoded in the knowledge graph as intermediate-suspicion calcifications with elevated contextual risk (Concept Nodes: Amorphous_Calcifications → Prior_ADH_History). **IMPRESSION:**	**FINDINGS:** Microcalcifications with enhancement in the left breast. **IMPRESSION:** BI-RADS 4B. Recommendation: biopsy.	4.8 | 4.7 4.9 | 4.2 **Mean: 4.65**	3.4 | 4.0 4.1 | 2.1 **Mean: 3.40**

			BI-RADS 4B (intermediate suspicion; estimated malignancy probability 10–50%). Stereotactic vacuum-assisted biopsy of the left lower inner quadrant calcification cluster is recommended.			

Note: Report excerpts are verbatim outputs from each system for the same patient examination; patient identifiers have been removed. Case 1 illustrates RMCP Layer 4 (Uncertainty Flagging): the secondary indeterminate left breast finding is surfaced with a specific management recommendation, whereas the flat KG baseline omits it entirely — a clinically significant omission. Case 2 illustrates RMCP Layer 3 (Analogical Recall): the radial scar concept node is retrieved from the knowledge graph and named in the Findings section with a concordance score, providing actionable differential context absent in the flat baseline. Case 3 illustrates RMCP Layer 2 (Evidential Interpretation): the prior ADH history is explicitly integrated into the BI-RADS 4B rationale through a knowledge-graph concept path, versus the flat baseline's generic Findings description. CAS dimension key: F = Findings factual correctness; I = Impression BI-RADS appropriateness; R = Recommendation clinical appropriateness; U = Uncertainty acknowledgement. ADH = atypical ductal hyperplasia; NME = non-mass enhancement; GT = radiologist panel ground-truth.

5.5 Uncertainty Calibration and Sentinel Activation Analysis

The Expected Calibration Error of the uncertainty score U across the 2,372 test cases is 0.043, indicating well-calibrated uncertainty quantification [74]: high-U cases exhibit materially higher BI-RADS prediction error, with the reliability diagram confirming an approximately linear relationship between mean U and observed error rate across the ten bins. By comparison, a softmax entropy baseline — the conventional uncertainty measure, computed without RM-CKG sentinel information — yields ECE 0.089, confirming that the sentinel-grounded uncertainty score is substantially better calibrated than parametric-entropy alone.

The Uncertainty Sentinel activation rate across the test partition is 8.3% (197/2,372 cases), indicating the certified ontology cannot reliably classify these patients' feature profiles into a defined diagnostic concept above confidence threshold θ. Activation is highest for BI-RADS 3 (14.2%) and 4A (11.7%) — the categories most epistemically evidence-dependent, since BI-RADS 3 requires confident exclusion of malignancy below 2% and 4A requires discrimination from BI-RADS 3 on subtle signals the ontology's inference rules most often flag as ambiguous — and lowest for BI-RADS 1 (2.1%) and 5 (3.4%), where feature profiles most reliably match established diagnostic concepts. Among the 197 sentinel-activated cases, the discriminative head's per-category error rate is 31.0%, versus 6.4% for the 2,175 sentinel-inactive cases — a 4.84-fold increase that validates the sentinel as a clinically meaningful indicator of cases warranting radiologist review rather than autonomous reporting.

5.6 Missing Modality Robustness

Table 7 reports AUC-ROC and QWK under systematic per-modality dropout at four withholding rates (25/50/75/100%), comparing sentinel token substitution against zero-vector substitution and the late-fusion ensemble (architecturally robust by design, as it averages only available-modality predictions). Results are averaged over ten random draws of the withheld subset per rate.

At 50% dropout — the most clinically relevant rate, reflecting an unacquired imaging modality — sentinel substitution achieves AUC 0.91 against zero-vector substitution's 0.87, a significant ΔAUC = +0.037 ($p < 0.001$). At 75% dropout (only one modality available), sentinel substitution achieves AUC 0.89 against 0.83, with the gap widening as missingness increases, consistent with learned sentinel representations providing progressively greater advantage over uninformative zero vectors as more tokens are substituted. At complete absence (100% dropout),

sentinel substitution achieves AUC 0.86 against zero-vector's 0.80 (full CIs in Table 7), confirming that the strategy maintains clinically deployable performance across the full range of modality availability encountered in practice.

Among individual modalities, mammography withholding produces the largest impact at 100% dropout (−0.092 AUC under sentinel substitution), reflecting its status as the primary screening modality and largest per-encoder contributor. DCE-MRI withholding produces the second-largest degradation (−0.081); ultrasound withholding produces a smaller degradation (−0.058), partially compensated by overlap between ultrasound morphological and mammographic margin features in the fusion. Withholding the LLM free-text encoder T_7 produces the smallest degradation (−0.031), as it supplements rather than duplicates the imaging and structured-data streams.

5.7 Computational Performance and Inference Latency

Inference latency is measured on a single NVIDIA A100 80GB GPU across 500 test cases at batch size 1, reflecting sequential clinical deployment. The discriminative pipeline (seven encoder forward passes, FAISS subgraph retrieval, GAT encoding, cross-modal fusion, ordinal regression) achieves a mean 4.3 seconds per examination, of which FAISS retrieval accounts for 0.31 s and the GAT encoder for 0.87 s. The generative module achieves a mean report generation time of 12.7 seconds at a maximum 512-token output, for a total end-to-end latency of 17.0 seconds — operationally acceptable against the 5–15 minutes a radiologist typically requires to review images, dictate a report, and assign a BI-RADS category. Batched inference at batch size 8 reduces per-case latency to 1.1 s (discriminative) and 3.8 s (generative), a 4.9-second total suitable for offline batch processing of screening populations.

Table 7. Missing modality robustness analysis on the held-out test partition (n = 2,372). Panel A: Macro-AUC-ROC and QWK under systematic imaging modality dropout at four withholding rates, comparing the proposed sentinel token substitution mechanism, zero-vector substitution, and the late-fusion DL ensemble baseline. Results averaged over ten random draws of the withheld patient subset per rate. Panel B: Per-encoder performance degradation at 100% withholding under sentinel versus zero-vector substitution; ΔAUC = AUC with encoder withheld − complete-model AUC (0.93). All values are mean (95% bootstrap CI, 1,000 resamples). p-values from DeLong's test (sentinel vs zero-vector); all p < 0.001 unless stated.

Panel A — Systematic Imaging Modality Dropout by Withholding Rate *Averaged over ten random draws per rate; methods compared at each dropout level*							
Withholding Rate	**Sentinel Token Substitution Macro-AUC-ROC (95% CI)**	**QWK**	**Zero-Vector Substitution Macro-AUC-ROC (95% CI)**	**QWK**	**ΔAUC (Sentinel vs Zero-vec)**	**Late-Fusion Ensemble Macro-AUC-ROC (95% CI)**	**QWK**
25%	0.93 *(0.92–0.94)*	0.83	0.92 *(0.91–0.93)*	0.82	+0.01	0.87 *(0.85–0.89)*	0.72
50% ★	**0.91** *(0.89–0.92)*	**0.80**	0.87 *(0.85–0.89)*	0.75	**+0.04**	0.86 *(0.84–0.88)*	0.71
75%	0.89 *(0.87–0.90)*	0.77	0.83 *(0.81–0.85)*	0.70	+0.05	0.85 *(0.83–0.86)*	0.69

100%	0.86 *(0.84–0.87)*	0.73	0.80 *(0.78–0.82)*	0.66	+0.06	0.83 *(0.81–0.85)*	0.67

Panel B — Per-Encoder Performance at 100% Withholding

Sentinel token vs zero-vector substitution; ΔAUC from complete-model baseline (AUC = 0.93); ranked by |ΔAUC sentinel|

Token	Encoder	Modality Type	Availability in Test Set	Sentinel Substitution AUC (95% CI)	ΔAUC (Sentinel)	Zero-Vector Sub. AUC (95% CI)	ΔAUC (Zero-vec)	Sentinel Advantage (ΔAUC Margin)
T_1	**Mammography Dual-Stream** *ResNet-50; cross-view attn.*	DL Imaging	100%	**0.84** *(0.82–0.86)*	**−0.09 ★**	0.78 *(0.76–0.80)*	−0.15	+0.06
T_3	DCE-MRI Temporal Encoder *3D-ResNet-50 + temporal attn.*	DL Imaging	73.4%	0.85 *(0.83–0.87)*	−0.08	0.80 *(0.78–0.82)*	−0.14	+0.05
T_2	Ultrasound Encoder *EfficientNet-B3; temporal mean pool*	DL Imaging	89.2%	0.87 *(0.85–0.89)*	−0.06	0.83 *(0.81–0.85)*	−0.10	+0.04
T_4	Clinical Risk Encoder *XGBoost + RF; leaf-node encoding*	Classical ML	100%	0.92 *(0.90–0.93)*	−0.02	0.91 *(0.89–0.92)*	−0.02	+0.01
T_7	ClinicalBERT Free-Text *[CLS] token; 768→512-d*	LLM	87.6%	0.90 *(0.88–0.92)*	−0.03	0.88 *(0.87–0.90)*	−0.05	+0.02
T_5	Pathological Risk Encoder *Multi-output XGBoost*	Classical ML	**52.4%**	0.92 *(0.91–0.94)*	−0.01	0.92 *(0.91–0.93)*	−0.01	+0.00
T_6	Environmental Risk Encoder *Random Forest; leaf encoding*	Classical ML	91.8%	0.93 *(0.91–0.94)*	−0.01	0.92 *(0.91–0.94)*	−0.01	+0.00

Note: ★ Panel A: 50% dropout highlighted as the most clinically relevant rate, reflecting examinations where one imaging modality was not acquired. Panel B: ★ = largest per-encoder |ΔAUC| under sentinel substitution (T_1

mammography, −0.092). Panel B rows ordered by descending |ΔAUC sentinel|. The sentinel token substitution advantage over zero-vector substitution widens as modality availability decreases (Panel A: +0.007 at 25% → +0.055 at 100%), confirming that learned sentinel representations provide progressively greater benefit over uninformative zero vectors as missingness increases. Availability values for Panel B reflect the proportion of the 2,372 test-partition patients for whom the encoder's input data were present. ΔAUC margin = AUC (sentinel) – AUC (zero-vector) at 100% withholding of that encoder. Late-fusion ensemble (Panel A) uses only the three imaging encoders and is therefore unaffected by non-imaging encoder withholding; it is reported only in Panel A. AUC = macro-averaged one-versus-rest ROC; QWK = quadratic weighted kappa.

5.8 Summary of Principal Findings

The results across Sections 5.1-5.7 establish four principal findings. First, integrating three AI paradigms across complementary modalities within a single BI-RADS-calibrated architecture exceeds any single-paradigm or single-modality baseline by a statistically significant, clinically meaningful margin on all metrics. Second, the RM-CKG's ontological grounding and expert validation contribute an independent performance increment over data-derived graph structure alone, concentrated specifically at the BI-RADS 4 subcategory discrimination task, where the ontology's evidential threshold encoding is most relevant and the gap over the strongest non-KG baseline is largest. Third, RMCP produces qualitatively and quantitatively superior generative reports relative to image-only and flat KG triplet conditioning, most distinctively in uncertainty acknowledgement — directly addressing the BI-RADS 0 incompleteness problem at the report generation level. Fourth, the sentinel token mechanism maintains clinically deployable performance across the full range of modality dropout encountered in practice, with the margin over zero-vector substitution widening as missingness increases, demonstrating robustness precisely where it is most needed.

6. Discussion

6.1 Clinical Significance of BI-RADS 4 Subcategory Discrimination

The most clinically consequential finding is not the aggregate performance — an AUC of 0.93 and QWK of 0.84 across 11,847 patients are, to our knowledge, the highest reported on a comparably constructed cohort — but the gain concentrated at the BI-RADS 4A/4B/4C boundaries. The 4A-versus-4B improvement from the RM-CKG over the strongest non-KG baseline (ΔAUC = +0.117) is over four times its aggregate ΔAUC (+0.028), reflecting design intent: the RM-CKG's decision attractors and edge weights are calibrated to the ACR Fifth Edition thresholds (4A 2–10%, 4B 10–50%, 4C 50–95%) as ontological axioms, not learned boundaries. A purely data-driven model must instead infer these from a label distribution that itself reflects substantial inter-reader disagreement (Spak et al., 2017 [75]; Elmore et al., 2015 [4]) — noise the RM-CKG's ontological encoding resists, which is precisely where the gain is largest.

The downstream consequences are substantial: unnecessary biopsies of BI-RADS 4A lesions (2–10% malignancy probability, frequently benign) are a major source of patient harm, cost, and anxiety (Hubbard et al., 2015 [76]). A system that distinguishes 4A from 4B with calibrated uncertainty — routing indeterminate cases to the Uncertainty Sentinel rather than forcing a label — offers a credible path to fewer unnecessary biopsies without more false negatives at the 4C/5 end. This is not theoretical: sentinel-activated cases show a 4.84-fold higher error rate, at an 8.3% activation rate concentrated in BI-RADS 3/4A, giving clinicians a concrete flag for escalation to review.

6.2 The Generative Report Module as a Clinical Communication Tool

The blinded evaluation matters more than the lexical metrics for clinical utility. BLEU-4/ROUGE-L gains from Step vii to viii are modest (+0.043, +0.050); RMCP's real contribution is not n-gram similarity but a change in what the report says where the flat triplet approach fails. The +0.44 CAS gain on Uncertainty Acknowledgement, and the Table 6 contrast — an RMCP report flagging a secondary indeterminate finding with a follow-up recommendation versus a flat-triplet report that omits it — is exactly the failure mode RMCP prevents. Incomplete reporting of indeterminate findings is a recognized medicolegal risk (Harvey et al., 2013 [77]); RMCP's Layer 4 addresses this by

serializing Sentinel activations into generation instructions the confident-language patterns of the fine-tuning corpus cannot suppress.

The BCR of 0.96 is insufficient alone as evidence of deployability. Of the 38 cases where inconsistency persisted, the post-hoc classifier correctly flags and triggers regeneration, but 19 remain inconsistent after a second pass — these sit near the ordinal decision boundary, where the VLM's posterior reflects the same uncertainty the discriminative head encodes. This low residual rate (19/2,372, 0.80%) reinforces the intended deployment model: decision support reviewed and endorsed by a radiologist before entering the patient record, not autonomous reporting.

6.3 Multi-Paradigm Architecture and Knowledge Grounding

Support for the three-paradigm design comes from the Step ii to iii transition (ΔAUC = +0.032, ΔQWK = +0.065), isolating the classical ML and LLM branches' contribution over the imaging ensemble. This arises from two sources: XGBoost + Random Forest contribute structured risk factors (menopausal status, hormonal history, breast density, family history) absent from imaging; ClinicalBERT captures narrative context (symptoms, prior biopsy history) often decisive in 3-versus-4A discrimination. Extending deep learning to these streams instead — an MLP over clinical variables, a vision-language objective over text — would be poorly matched to the data: tree ensembles are systematically superior on clinical tabular benchmarks (Shwartz-Ziv and Armon, 2022 [26]), and ClinicalBERT's MIMIC-III pre-training provides context general-purpose models lack.

The three-round expert validation is a methodological contribution distinct from the GAT-based prior injection itself. It mirrors how clinical decision-support knowledge bases are certified for regulated use, per DECIDE-AI (Vasey et al., 2022 [79]). However, validation by radiologists who encode their own reasoning practices risks formalizing rather than correcting systematic biases. As Obermeyer et al. (2019) [78] show for clinical risk prediction generally, AI calibrated against expert practice can inherit and amplify diagnostic disparities across demographic subgroups — a risk applicable here to the extent the validating radiologists' experience differs from the deployment population's.

6.4 Limitations

Several limitations require acknowledgement. Most significant is generalizability: although the cohort spans three tertiary centres and four public datasets, institutional data comes from a single country's healthcare system, and patient ethnic/socioeconomic diversity is not fully characterized. The RM-CKG's edge weights, from an expert panel in one clinical-cultural context, may not transfer to settings with different BI-RADS norms, referral thresholds, or screening protocols. External validation on independent cohorts is necessary before these figures can be considered generalizable.

A second limitation is ground-truth label construction: labels reflect radiologist consensus rather than histopathological confirmation for all cases, since pathological confirmation exists only for biopsied cases (BI-RADS 4/5). For categories 1-3, the label is an assessment, not a gold standard, and the inter-reader variability Elmore et al. (2015) [4] documented — 72.2% agreement on category 4 subcategory assignment — introduces unquantified noise that may inflate apparent performance.

A third limitation is the RM-CKG's static representation: though formalized in OWL 2 DL, the literature evolves — new screening modalities, molecular subtyping, model updates — and the graph lacks automated updating; certification is episodic, so the gap with current evidence widens without periodic re-validation. Also, the pathological encoder T_5 is available for only 52.4% of patients, so its Case Nodes are correspondingly sparse, potentially underweighting pathological evidential pathways; the missing-modality sentinel partially compensates, but this coverage asymmetry remains a limitation for future imputation or transfer-learning augmentation.

A fourth limitation is the generative evaluation: though CAS shows high inter-rater agreement (ICC 0.82-0.87), the 200-report cohort (8.4% of the test partition) was randomly sampled rather than enriched for sentinel-activated, discordant, or retried cases most likely to expose failures. A prospective evaluation targeting those cases would give a more conservative estimate of report quality in the operational tail.

A fifth limitation is computational: the 4.3-second discriminative and 12.7-second generative latencies are measured on an A100 GPU, not uniformly available where AI-assisted mammography could have the greatest impact — low- and middle-income settings. FAISS retrieval and GAT encoding add ~1.2 seconds over a KG-free model; RMCP prompt construction adds a further 0.8 seconds. These latencies suit a high-volume tertiary workflow, but resource-constrained deployment will need compression — GAT distillation, VLM quantization, dynamic prompt truncation — not evaluated here.

6.5 Future Directions

These findings suggest several future directions. The RM-CKG would benefit from continuous learning — automated literature surveillance analogous to living systematic reviews (Elliott et al., 2017 [80]) — with radiologist endorsement of incremental updates rather than full re-certification. Near-boundary consistency failures suggest ensemble decoding — generating several candidates and selecting the most consistent — may reduce residual inconsistency more cheaply than the current two-pass approach. The missing-modality sentinel, shown here for single-modality dropout, should extend to simultaneous multi-modality dropout, including the realistic case where DCE-MRI and pathological data are both absent. Finally, a prospective randomized reader study — radiologists reviewing reports with and without RMCP and the uncertainty flag — is the necessary next step toward evidence of measurable gains in accuracy, biopsy decisions, and reporting completeness.

7. Conclusion

This paper has presented a hybrid discriminative-generative framework for BI-RADS risk prediction and structured report generation, integrating deep learning, classical ensemble learning, and LLM encoding across eight modality streams in one architecture. Its two novelties, the RM-CKG and RMCP, are architecturally entangled rather than independent: the RM-CKG generates both the cross-modal attention prior and the subgraph traversal RMCP serializes into its five-layer conditioning prompt, so one knowledge structure grounds both the ordinal prediction and the report generation.

On a cohort of 11,847 patients across three centres and four public datasets, three findings stand out. The framework achieves macro-averaged AUC-ROC 0.93 and QWK 0.84, exceeding all baselines — but the more clinically significant result is that the RM-CKG's ontological grounding produces its largest increment at the BI-RADS 4A/4B/4C boundaries (ΔAUC = +0.117 versus +0.028 overall), precisely where miscategorization is costliest. The RMCP-conditioned module achieves BCR 0.96 and CAS 4.31/5.00 — the only breast imaging report system, to our knowledge, exceeding a clinician-assessed 4.0 — with its distinctive contribution being uncertainty acknowledgement rather than n-gram similarity. The Uncertainty Sentinel is well-calibrated (ECE = 0.043), with a 4.84-fold error-rate increase for activated cases, giving a reliable routing signal for radiologist review.

These figures are encouraging, but clinical utility beyond this retrospective cohort requires validations this study cannot provide: prospective external validation on demographically independent cohorts with different epidemiology and screening conventions, and periodic re-validation of the RM-CKG's edge weights as the literature and imaging modalities evolve — ideally through continuous, automated surveillance flagging candidate updates for targeted expert endorsement rather than full re-certification.

Two refinements follow for the generative module: an ensemble decoding strategy — generating several candidates and selecting the most consistent via the post-hoc classifier — may reduce residual inconsistency more cheaply than the current two-pass approach; and a dynamic-depth RMCP prompt, adapting the number of Case Nodes and evidential granularity to presentation complexity via Sentinel activation and ordinal entropy, may improve both efficiency and specificity.

The missing-modality analysis shows sentinel substitution holds up under single-modality dropout, but simultaneous multi-modality dropout — the norm in low-resource screening — is not fully evaluated. Extending robustness to this pattern, and testing whether distillation and quantization can bring latency within resource-limited hardware constraints, has direct epidemiological relevance: the screening burden is largest where multimodal

infrastructure is least available, and graceful degradation with calibrated uncertainty is more useful there than complete-data-optimized performance alone.

Finally, translating these improvements into better clinical decisions requires a prospective randomized reader study comparing BI-RADS assignment and reporting with and without the framework's outputs. The relevant endpoints are not AUC or QWK, which can be optimized against a label set reflecting judgement's own variability, but clinical outcomes: unnecessary biopsy rate for 4A, false-negative rate for 4C/5, sign-off time, and completeness of secondary-finding documentation — the metrics that will show whether this work's contributions translate into patient benefit.

REFERENCES

1. Bray, F., Ferlay, J., Laversanne, M., Soerjomataram, I., Jemal, A., & Torre, L. A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74, 229–263. https://doi.org/10.3322/caac.21834
2. Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. CA: A Cancer Journal for Clinicians, 74(1), 12–49. https://doi.org/10.3322/caac.21820
3. Aristokli, N., Polycarpou, I., Themistocleous, S. C., Sophocleous, D., & Mamais, I. (2022). Comparison of the diagnostic performance of magnetic resonance imaging (MRI), ultrasound and mammography for detection of breast cancer based on tumor type, breast density and patient's history: A review. Radiography, 28(3), 848–856. https://doi.org/10.1016/j.radi.2022.01.006
4. Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N. A., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J., O'Malley, F. P., & Weaver, D. L. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA, 313(11), 1122–1132. https://doi.org/10.1001/jama.2015.1405
5. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6
6. Tan H, Wu Q, Wu Y, Zheng B, Wang B, Chen Y, Du L, Zhou J, Fu F, Guo H, Fu C, Ma L, Dong P, Xue J, Wang M. Development of a Multi-view Multi-level Artificial Intelligence System to Stratify Risk Assessment of Mammography. Research Square; 2023. doi:10.21203/rs.3.rs-2489648/v1
7. Sprague, B. L., Conant, E. F., Onega, T., Garcia, M. P., Beaber, E. F., Herschorn, S. D., Weaver, D. L., Tosteson, A. N. A., Schnall, M., Bassett, L. W., Braithwaite, D., Haas, J. S., Tice, J. A., & Miglioretti, D. L. (2014). Variation in mammographic breast density assessments among radiologists in clinical practice: A multicenter observational study. Annals of Internal Medicine, 161(7), 457–467. https://doi.org/10.7326/M14-0588
8. D'Orsi, C. J., Sickles, E. A., Mendelson, E. B., & Morris, E. A. (Eds.). (2013). ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System (5th ed.). American College of Radiology.
9. Berg, W. A., Zhang, Z., Lehrer, D., Jong, R. A., Pisano, E. D., Barr, R. G., Bohm-Velez, M., Mahoney, M. C., Evans, W. P., Larsen, L. H., Morton, M. J., Mendelson, E. B., Farria, D. M., Cormack, J. B., Marques, H. S., Adams, A., Yeh, N. M., & Gabrielli, G. (2012). Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. JAMA, 307(13), 1394–1404. https://doi.org/10.1001/jama.2012.388
10. 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). IEEE. https://doi.org/10.1109/CVPR.2016.90
11. Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Fevry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L. L. Y., Ho, K., Weinstein, J. D., Reig, B., Gao, Y., Toth, H., & Geras, K. J. (2019). Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Transactions on Medical Imaging, 39(4), 1184–1194. https://doi.org/10.1109/TMI.2019.2945514
12. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 6105–6114). PMLR.
13. Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyson, S., Heacock, L., Lewin, A. A., Moy, L., & Geras, K. J. (2021). An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Medical Image Analysis, 68, 101908. https://doi.org/10.1016/j.media.2020.101908
14. Kuhl, C. K., Schrading, S., Leutner, C. C., Morakkabati-Spitz, N., Wardelmann, E., Fimmers, R., Kuhn, W., & Schild, H. H. (2005). Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. Journal of Clinical Oncology, 23(33), 8469–8476. https://doi.org/10.1200/JCO.2004.00.4960

15. Truhn, D., Schrading, S., Haarburger, C., Schneider, H., Merhof, D., & Kuhl, C. (2019). Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. Radiology, 290(2), 290–297. https://doi.org/10.1148/radiol.2018181352
16. Antropova, N., Huynh, B. Q., & Giger, M. L. (2017). A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Medical Physics, 44(10), 5162–5171. https://doi.org/10.1002/mp.12453
17. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sanchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005
18. Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K. E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z. A., & Yang, H. (2022). RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence, 4(5), e210315. https://doi.org/10.1148/ryai.210315
19. 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. https://doi.org/10.1148/radiol.2019182716
20. Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute, 81(24), 1879–1886. https://doi.org/10.1093/jnci/81.24.1879
21. Tyrer, J., Duffy, S. W., & Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. Statistics in Medicine, 23(7), 1111–1130. https://doi.org/10.1002/sim.1668
22. Lee, A., Mavaddat, N., Wilcken, N., Cunningham, A. P., Carver, T., Hartley, S., Babb de Villiers, C., Izquierdo, A., Harrison, H., Adlard, J., Andrews, L., Barwell, J., Brewer, C., Calender, A., Cook, J., Evans, D. G., George, R., Izatt, L., Lalloo, F., & Easton, D. F. (2019). BOADICEA: A comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genetics in Medicine, 21(8), 1708–1718. https://doi.org/10.1038/s41436-018-0406-9
23. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785
24. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
25. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 35, pp. 507–520). https://doi.org/10.48550/arXiv.2207.08815
26. Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011
27. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 4171–4186). ACL. https://doi.org/10.18653/v1/N19-1423
28. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72–78). ACL. https://doi.org/10.18653/v1/W19-1909
29. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1), 2:1–2:23. https://doi.org/10.1145/3458754
30. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Shetty, S., Natarajan, V., Kelsey, F., Foster, A., Lyu, D., & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620, 172–180. https://doi.org/10.1038/s41586-023-06291-2
31. Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 30, pp. 5998–6008).
33. Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. Nature Medicine, 28, 1773–1784. https://doi.org/10.1038/s41591-022-01981-2
34. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2021). Fusion of medical imaging and electronic health records using deep learning: Systematic review and implementation guidelines. npj Digital Medicine, 4, 136. https://doi.org/10.1038/s41746-020-00341-z
35. Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine, 34(6), 96–108. https://doi.org/10.1109/MSP.2017.2738401
36. Ma, X., Zhang, T., & Xu, C. (2021). SMIL: Multimodal learning with severely missing modality. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(3), 2302–2310. https://doi.org/10.1609/aaai.v35i3.16330

37. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. Scientific Reports, 7, 45634. https://doi.org/10.1038/srep45634
38. Chandak, P., Huang, K., & Zitnik, M. (2023). Building a knowledge graph to enable precision medicine. Scientific Data, 10, 67. https://doi.org/10.1038/s41597-023-01960-3
39. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1609.02907
40. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1710.10903
41. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), 71:1–71:37. https://doi.org/10.1145/3447772
42. Jing, B., Xie, P., & Xing, E. P. (2018). On the automatic generation of medical imaging reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 2577–2586). ACL. https://doi.org/10.18653/v1/P18-1240
43. Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1439–1449). ACL. https://doi.org/10.18653/v1/2020.emnlp-main.112
44. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D. C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Tinn, R., Matthews, J., Naumann, T., & Alvarez-Valle, J. (2023). Learning to exploit temporal structure for biomedical vision-language processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15016–15027). https://doi.org/10.1109/CVPR52729.2023.01441
45. Hyland, S. L., Bannur, S., Bouzid, K., Castro, D. C., Ranjit, M., Thieme, A., Schwaighofer, A., Wetscherek, M. T., Ilse, M., Perez-Garcia, F., Sharma, H., Tinn, R., Naumann, T., & Alvarez-Valle, J. (2023). MAIRA-1: A specialised large multimodal model for radiology report generation. arXiv preprint arXiv:2311.13206. https://doi.org/10.48550/arXiv.2311.13206
46. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 33, pp. 9459–9474).
47. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 35, pp. 24824–24837).
48. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 36, pp. 11809–11822).
49. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100
50. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In Proceedings of the 10th International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2106.09685
51. Langlotz, C. P. (2006). RadLex: A new method for indexing online educational materials. RadioGraphics, 26(6), 1595–1597. https://doi.org/10.1148/rg.266065168
52. Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. Journal of Neuroscience Methods, 264, 47–56. https://doi.org/10.1016/j.jneumeth.2016.03.001
53. Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037
54. Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., & Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. NeuroImage, 167, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024
55. Kaminska, M., Ciszewski, T., Lopacka-Szatan, K., Miotla, P., & Staroslawska, E. (2015). Breast cancer risk factors. Przeglad Menopauzalny, 14(3), 196–202. https://doi.org/10.5114/pm.2015.54346
56. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 30, pp. 4765–4774).
57. 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 (ICCV) (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74
58. McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological), 42(2), 109–127. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

59. Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213–220. https://doi.org/10.1037/h0026256
60. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845. https://doi.org/10.2307/2531595
61. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
62. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35
63. Johnson, J., Douze, M., & Jegou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572
64. 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, 170177. https://doi.org/10.1038/sdata.2017.177
65. 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. https://doi.org/10.1016/j.acra.2011.09.014
66. Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., & Vu, V. (2022). VinDr-Mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography. medRxiv. https://doi.org/10.1101/2022.03.07.22272029
67. Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. Data in Brief, 28, 104863. https://doi.org/10.1016/j.dib.2019.104863
68. Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A. K., & Marti, R. (2017). Automated breast ultrasound lesions detection using convolutional neural networks. IEEE Journal of Biomedical and Health Informatics, 22(4), 1218–1226. https://doi.org/10.1109/JBHI.2017.2731873
69. Saha, A., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., & Mazurowski, M. A. (2018). A machine learning approach to radiogenomics of breast cancer: A study of 922 subjects and 529 DCE-MRI features. British Journal of Cancer, 119(4), 508–516. https://doi.org/10.1038/s41416-018-0185-8
70. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 311–318). ACL. https://doi.org/10.3115/1073083.1073135
71. Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Proceedings of the ACL Workshop: Text Summarization Branches Out (pp. 74–81). ACL.
72. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In Proceedings of the 8th International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1904.09675
73. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420
74. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (ICML) (pp. 1321–1330). PMLR.
75. Spak, D. A., Plaxco, J. S., Santiago, L., Dryden, M. J., & Dogan, B. E. (2017). BI-RADS® 4A, B, and C categories of probably benign: ACR BI-RADS® fifth edition. Academic Radiology, 24(9), 1101–1109. https://doi.org/10.1016/j.acra.2017.04.013
76. Hubbard, R. A., Zhu, W., Onega, T., Smith, R. A., Ichikawa, L., Miglioretti, D. L., & Sprague, B. L. (2015). Factors associated with the likelihood of additional imaging following screening mammography. Cancer Epidemiology, Biomarkers & Prevention, 24(4), 637–643. https://doi.org/10.1158/1055-9965.EPI-14-1083
77. Harvey, J. A., Nicholson, B. T., & Bhatt, D. L. (2013). Importance of secondary findings detected at breast MRI: Clinical and mammographic follow-up. American Journal of Roentgenology, 200(3), 698–706. https://doi.org/10.2214/AJR.12.8649
78. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342
79. Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., & Vollmer, S. J. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 28, 924–933. https://doi.org/10.1038/s41591-022-01772-9
80. Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., Hilton, J., Tovey, D., Shemilt, I., & Thomas, J. (2017). Living systematic reviews: An emerging opportunity to narrow the evidence-practice gap. PLOS Medicine, 14(5), e1002274. https://doi.org/10.1371/journal.pmed.1002274