Scientific Reports

https://doi.org/10.1038/s41598-025-32794-1

Article in Press

CMAF-Net: cross-modal attention fusion with information-theoretic regularization for imbalanced breast cancer histopathology

Received: 13 September 2025

Accepted: 12 December 2025

Published online: 26 February 2026

Cite this article as: Ativi W.X., Chen W., Kwao L. *et al.* CMAF-Net: cross-modal attention fusion with information-theoretic regularization for imbalanced breast cancer histopathology. *Sci Rep* (2025). https://doi.org/10.1038/s41598-025-32794-1

Wisdom Xornam Ativi, Wenyu Chen, Lazarus Kwao, Williams Ayivi, Francis Sam, Ali Alqahtani, Yeong Hyeon Gu & Mugahed A. Al-antari

CMAF-Net: Cross-Modal Attention Fusion with Information-Theoretic Regularization for Imbalanced Breast Cancer Histopathology

Wisdom Xornam Ativi[1*], Wenyu Chen[1*], Lazarus Kwao[1,2], Williams Ayivi[3], Francis Sam[4], Ali Alqahtani[5], Yeong Hyeon Gu[6*], Mugahed A. Al-antari[6*]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.
[2]Department of Computer Science, Sunyani Technical University, Sunyani, Ghana.
[3]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.
[4]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.
[5]Center for Artificial Intelligence and Computer Science Department, King Khalid University, Abha, 61421, Saudi Arabia.
[6]Department of Artificial Intelligence and Data Science, College of AI Convergence, Daeyang AI Center, Sejong University, Seoul, 05006, Republic of Korea.

*Corresponding author(s). E-mail(s): ativiw@std.uestc.edu.cn; cwy@uestc.edu.cn; yhgu@sejong.ac.kr; en.mualshz@sejong.ac.kr;
Contributing authors: lazoe16@yahoo.com; w.ayivi@std.uestc.edu.cn; fransamang@std.uestc.edu.cn; amosfer@kku.edu.sa;

Abstract

Breast cancer diagnosis from histopathology images remains challenging due to two intertwined factors: severe class imbalance, where malignant cases represent a small minority of samples, and the need to integrate discriminative features across multiple spatial scales. Existing methods typically address imbalance and multi-scale fusion separately, leading to biased or redundant representations.

We propose CMAF-Net, a theoretically grounded architecture that unifies information bottleneck principles with margin-based learning to jointly tackle these challenges. CMAF-Net employs a dual-branch CNN–Transformer backbone fused through a Cross-Modal Attention Fusion block, which implements temperature-controlled attention and redundancy minimization to preserve complementary local and global features. At the classification level, we introduce an Adaptive Class-Balanced Focal Loss that operationalizes margin theory under imbalance, enforcing larger margins for minority classes while dynamically adapting to feature distributions. Extensive experiments on the IDC dataset show that CMAF-Net achieves 94.92% sensitivity and 95.52% balanced accuracy, outperforming state-of-the-art baselines by up to 8.6% on malignant detection. Under extreme 99:1 imbalance, CMAF-Net maintains 76.45% sensitivity, demonstrating graceful degradation where competing methods fail catastrophically. Cross-dataset evaluation on BreakHis confirms robust zero-shot transfer across four magnifications with average sensitivity of 95.61%. Ablation studies and information-theoretic analyses validate the contributions of each component, while computational profiling shows CMAF-Net achieves superior accuracy–efficiency trade-offs compared to prior fusion networks. Beyond breast cancer, our framework establishes a principled template for information-theoretic fusion under class imbalance, with implications for rare disease detection, clinical decision support, and broader multi-modal learning tasks.

Keywords: Information bottleneck, Cross-modal fusion, Class imbalance, Attention mechanisms, Medical image analysis, Deep learning

1 Introduction

Breast cancer remains the leading cause of cancer-related mortality among women worldwide, with Invasive Ductal Carcinoma (IDC) accounting for approximately 70%–80% of all diagnosed cases [1]. Accurate and early detection of malignant tissue in breast histopathology is essential for improving patient prognosis. Histopathological examination remains the gold standard for diagnosis; however, its manual interpretation is labor-intensive, subjective, and susceptible to inter-observer variability.

Automated analysis using deep learning offers promise, yet faces two critical and interconnected challenges. First, breast histopathology datasets typically exhibit severe class imbalance: malignant samples, though clinically vital to detect, comprise only 20%–30% of tissue samples in screening populations [2]. This imbalance biases learning kalgorithms toward the majority (benign) classes, leading to poor generalization for minority (malignant) cases. Second, accurate malignancy detection requires integrating features across multiple spatial scales. Discriminative cues appear at both cellular and tissue levels, from nuclear pleomorphism and mitotic activity at high magnification to architectural disorganization and invasion patterns at lower resolutions [3]. Conventional deep learning pipelines, such as CNNs [4], struggle to bridge this scale gap, while recent transformer-based approaches [5, 6] offer improved contextual modeling but often lack explicit mechanisms for adaptive feature fusion across scales.

Existing solutions tend to address these challenges in isolation. Class imbalance is typically mitigated using static techniques such as resampling [7], cost-sensitive loss functions [8], or variants of focal loss [9]. These methods, however, are not adaptive to the evolving representation learning process. Meanwhile, multi-scale fusion strategies often rely on naive concatenation or late ensembling, which neglect the interaction between semantic relevance and class representation quality [10]. Such fusion approaches may inadvertently amplify redundant majority-class features while suppressing the sparse but critical features of minority-class malignancies. These limitations point to a broader gap in current research: the lack of unified models that explicitly account for class imbalance during multi-scale feature integration. In imbalanced histopathological data, discriminative features for malignancy are both semantically rare and spatially dispersed. Thus, the ability to fuse multi-scale features in a class-sensitive manner is essential, not merely for accuracy but for diagnostic reliability.

To address these challenges, we propose CMAF-Net (Cross-Modal Attention Fusion Network), a hybrid architecture that jointly mitigates class imbalance and enhances multi-scale fusion through principled learning strategies. At the fusion level, we introduce a Cross-Modal Attention Fusion (CMAF) block guided by the Information Bottleneck (IB) principle [11], which selectively integrates CNN-derived local features and Transformer-based global representations while minimizing redundant information. This enables the model to retain task-relevant cues across spatial scales. At the classification level, we incorporate margin theory for imbalanced learning [12] by designing an Adaptive Class-Balanced Focal Loss (A-CBFL), which enforces larger margins for minority classes and dynamically adjusts to the evolving feature distributions of the model. By treating fusion and imbalance not as separate engineering problems but as coupled learning constraints, CMAF-Net enhances both the robustness and clinical reliability of automated breast cancer diagnosis.

Our work makes the following contributions to medical image analysis:

1. We introduce CMAF-Net, the first histopathology classification model that explicitly optimizes the Information Bottleneck objective through learnable cross-modal attention with temperature-controlled information flow.
2. Our A-CBFL loss implements margin theory with dynamic scheduling, achieving significant improvements in minority class detection.
3. We demonstrate that information-theoretic fusion naturally handles the multi-scale nature of histopathology, from cellular details to tissue architecture, within a single coherent framework.
4. We systematically evaluate performance across imbalance ratios from 70:30 to 99:1, demonstrating that our theoretical framework maintains effectiveness even in screening scenarios where malignant cases are extremely rare.

The remainder of this paper is organized as follows. Section 2 reviews related work on class imbalance Section 3 presents our theoretical framework. Section 4 details the CMAF-Net architecture and training methodology. Section 5 describes experimental setup and datasets. Section 6 presents comprehensive results. Section 7 discusses implications, limitations, and broader impact. Section 8 concludes with future directions.

2 Related Work

The development of CMAF-Net builds upon and synthesizes advances across multiple research areas: multi-modal fusion architectures, imbalanced learning methods, information-theoretic deep learning, and medical image analysis. In this section, we provide a comprehensive review of these foundations, highlighting both the progress made and the gaps that motivate our approach.

2.1 Multi-Modal and Cross-Modal Fusion Architectures

The fusion of heterogeneous features remains a central challenge in medical imaging, with approaches evolving from simple concatenation to sophisticated attention-based mechanisms [13, 14]. Traditional fusion strategies can be categorized into three paradigms [15]. Early fusion concatenates raw features before processing but suffers from dimensional explosion and fails to preserve modality-specific characteristics [16]. This becomes particularly problematic when fusing CNN and Transformer features due to their different inductive biases [17]. Late fusion preserves modality-specific processing but limits cross-modal interaction to the decision level, underperforming when modalities contain complementary rather than redundant information [18, 19]. Hybrid fusion attempts to balance these approaches; TransFuse [20] exemplifies this with parallel CNN-Transformer branches, though it lacks theoretical grounding for handling redundancy.

Recent attention-based fusion methods dynamically weight features based on relevance. CrossViT [21] pioneered cross-attention for multi-scale fusion, while MLFF-Net [22] introduced multi-level attention modules for medical segmentation. DAFNet [23] proposed adaptive normalization convolution that dynamically modifies weights based on input features. Bottleneck approaches like Perceiver [24, 25] address computational efficiency through learned compression tokens. However, these methods optimize for computational rather than information-theoretic objectives, potentially discarding clinically relevant features [14].

Medical imaging presents unique challenges requiring preservation of subtle features while managing class imbalance [26]. Recent surveys reveal that most medical fusion methods lack theoretical justification [14]. While TinyViT-LightGBM [27] achieved 97.8% accuracy in breast cancer diagnosis through efficient multi-source fusion, and Liu et al. [28] explored CLIP-driven fusion, existing approaches fail to explicitly address the severe class imbalance prevalent in disease detection, a critical oversight given rare disease detection requirements. This separation of fusion and imbalance handling ignores their fundamental interdependence: under severe imbalance, fusion strategies must prioritize preserving sparse minority-class signals rather than treating all features equally. This gap motivates our information-theoretic approach that explicitly considers both redundancy minimization and class imbalance compensation within a unified framework.

2.2 Class Imbalance in Deep Learning

Building on the fusion challenges, class imbalance presents another critical obstacle that becomes even more complex when combined with multi-modal learning. Class

imbalance has been extensively studied, with solutions broadly categorized into data-level, algorithm-level, and hybrid approaches[29]. At the data level, resampling methods such as SMOTE [7] and its variants generate synthetic minority samples, but in medical imaging, this risks creating unrealistic pathological patterns. Recent work by Mullick et al. [30] explored GAN-based oversampling for medical images, achieving mixed results due to the difficulty of generating realistic pathology. Zhang et al. [31] comprehensively reviewed augmentation strategies for imbalanced medical imaging, finding that while augmentation helps, it cannot fully address extreme imbalance where minority examples are exceptionally rare. Chlap et al. [32] found that aggressive augmentation can even hurt performance by introducing unrealistic variations.

Algorithm-level approaches have proven more promising for medical applications. Classical cost-sensitive learning assigns different misclassification costs to different classes, with recent theoretical work by Menon et al. [33] establishing connections between cost-sensitive learning and margin theory, showing that optimal costs depend on both class imbalance and Bayes error, which motivates our margin-based approach. Lin et al.'s [9] focal loss revolutionized imbalanced object detection by focusing on hard examples, spawning numerous variants. Cui et al. [8] proposed class-balanced focal loss, while Li et al. [34] developed dice focal loss for medical segmentation. However, these use fixed hyperparameters rather than adapting to class distributions.

Margin-based methods have shown particular promise for medical imaging. Cao et al. [12] provided theoretical analysis showing optimal margins should scale with class imbalance. Their LDAM loss implements this insight but doesn't extend to multi-modal settings. Recent work by Kini et al. [35] showed that margin-based methods are particularly effective under label noise, which is common in medical annotations. The importance of calibrated predictions in imbalanced settings has also gained attention, with Menon et al. [36] showing that standard neural networks are poorly calibrated under imbalance and Collell et al. [37] proposing post-hoc calibration methods. Our approach addresses calibration through principled margin design rather than post-hoc correction. These margin-based insights directly inform our theoretical framework, providing a foundation for handling imbalance in the context of multi-modal fusion.

2.3 Information-Theoretic Deep Learning

The limitations of empirical approaches to both fusion and imbalance motivate a more principled theoretical foundation. Information theory provides such a framework for understanding and designing neural networks, with the Information Bottleneck principle emerging as particularly influential. Tishby et al. [11] introduced the Information Bottleneck principle, which Shwartz-Ziv and Tishby [38] later connected to deep learning, showing that successful neural networks naturally compress information during training, a phenomenon we explicitly optimize. Alemi et al. [39] developed practical algorithms for optimizing IB in neural networks, using variational bounds to make the problem tractable. While Saxe et al. [40] challenged some claims about IB in deep learning, they confirmed its utility as a design principle.

Recent theoretical advances have strengthened the foundation for information-theoretic approaches. Goldfeld and Polyanskiy [41] provided rigorous analysis of IB in modern neural networks, while Geiger et al. [42] showed connections between IB and

generalization. These works support our use of IB for designing robust architectures. The extension to multi-modal settings has seen limited but promising work. Federici et al. [43] extended IB to multi-view learning, showing that minimizing view-specific information improves generalization. Wang et al. [44] applied multi-view IB to medical imaging but didn't consider class imbalance.

Information measures have a long history in medical imaging, particularly for registration and fusion. Mutual information has been the cornerstone of medical image registration [45], with recent work by Guo et al. [46] extending MI-based registration to deep learning. Li et al. [47] used information theory for multi-modal medical fusion but without considering class imbalance. Elton et al. [48] showed that information-theoretic measures provide interpretable insights into medical AI decisions, motivating our use of information metrics for understanding fusion behavior. The high dimensionality and noise characteristics of medical images make information-theoretic approaches particularly attractive, as they provide principled ways to identify and preserve diagnostically relevant signals while handling the challenges of both fusion and imbalance.

2.4 Vision Transformers and Hybrid Architectures

Information theory provides the theoretical foundation, but practical implementation requires appropriate architectural choices. The emergence of Vision Transformers has created new opportunities for multi-scale feature learning and fusion. Since Dosovitskiy et al. [5] introduced ViT, numerous variants have emerged. Liu et al. [49] developed Swin Transformer with hierarchical representations, while Touvron et al. [50] created DeiT for efficient training. These provide strong global features but struggle with fine-grained details crucial in medical imaging. Chen et al. [51] pioneered transformer use in medical segmentation, while Shamshad et al. [17] comprehensively reviewed transformers in medical imaging, consistently finding that pure transformers underperform on tasks requiring fine detail recognition.

This limitation has driven the development of CNN-Transformer hybrids. Dai et al. [6] developed CoAtNet combining convolution and attention, while Graham et al. [52] created LeViT for efficient hybrid processing. Wu et al. [53] proposed Convolutional vision Transformers, embedding convolution within transformer blocks. For resource-constrained medical applications, Mehta and Rastegari [54] developed MobileViT combining the strengths of CNNs and ViTs for mobile deployment. This inspired our choice of MobileViT for the global branch, as medical AI often requires deployment on resource-constrained hardware. These hybrid architectures inspire our dual-branch design, though none explicitly optimize information flow between branches, a critical consideration under class imbalance that our CMAF block addresses.

2.5 Medical Image Analysis with Deep Learning

The unique challenges of medical imaging bring together all the aforementioned areas, requiring solutions that simultaneously handle multi-scale features, severe class imbalance, and computational constraints. The application domain of medical imaging brings unique challenges that inform our design choices. In computational pathology, Srinidhi

et al. [55] surveyed deep learning applications, highlighting class imbalance as a critical challenge. Dimitriou et al. [56] specifically reviewed deep learning for histopathology, noting that multi-scale analysis is essential for accurate diagnosis. For breast cancer detection, Spanhol et al. [57] created the BreakHis dataset, establishing benchmarks for breast cancer histopathology. Recent work by Yan et al. [58] achieved high accuracy but only under balanced conditions; performance degraded significantly with natural class distributions.

The importance of multi-scale analysis in pathology cannot be overstated. Madabhushi and Lee [3] showed that different magnifications reveal different diagnostic features. Tellez et al. [59] quantified the importance of multi-scale analysis, finding that combining multiple scales improves performance by 15-20% on average. Attention mechanisms have shown promise for pathology, with Ilse et al. [60] introducing attention-based pooling for multiple instance learning. Li et al. [61] developed dual-stream networks for multi-scale pathology, but their approach uses simple concatenation rather than principled fusion. These works collectively demonstrate that successful medical image analysis requires addressing fusion, imbalance, and computational efficiency simultaneously, not as separate problems.

2.6 Positioning Our Contributions

Our work synthesizes insights from these diverse areas while addressing critical gaps. While information theory provides elegant frameworks and margin theory offers optimal solutions for imbalance, few works successfully implement these theories in practical architectures. We bridge this gap through careful architectural design that directly optimizes theoretical objectives. Despite extensive work on both fusion and imbalanced learning, the intersection remains unexplored. We show that these challenges are fundamentally connected and benefit from unified treatment. Existing fusion methods optimize for natural images where class balance is less critical. Medical imaging's unique requirements (extreme imbalance, multi-scale diagnosis, interpretability needs) motivate our specialized approach. Finally, while theoretical papers provide guarantees and empirical papers show results, few works validate theoretical predictions through careful measurement. Our information-theoretic analysis provides concrete evidence for theoretical frameworks.

By addressing these gaps, CMAF-Net advances both the theoretical understanding of information fusion and its practical application to challenging real-world problems. The next section develops our theoretical framework, showing how information bottleneck principles naturally address both fusion and imbalance challenges.

3 Theoretical Framework

The design of CMAF-Net rests on a rigorous theoretical foundation that unifies information-theoretic principles with statistical learning theory. In this section, we develop the mathematical framework that motivates our architectural choices and provides performance guarantees under class imbalance. We begin with the information bottleneck principle for single-modal learning, extend it to multi-modal fusion, and show how it naturally connects with margin theory for imbalanced classification.

3.1 Information Bottleneck for Representation Learning

The Information Bottleneck (IB) principle, introduced by Tishby et al. [11], provides an information-theoretic framework for learning optimal representations. Given input X, target Y, and learned representation Z, the IB objective seeks to maximize the mutual information $I(Z;Y)$ while minimizing $I(Z;X)$:

$$\mathcal{L}_{\text{IB}} = -I(Z;Y) + \beta \cdot I(Z;X) \tag{1}$$

where $\beta > 0$ controls the compression-relevance trade-off. This formulation elegantly captures the notion that good representations should be predictive ($I(Z;Y)$ large) while discarding task-irrelevant information ($I(Z;X)$ small).

Recent work by Goldfeld and Polyanskiy [41] established that optimizing Equation 1 provides generalization guarantees. Specifically, they showed that the generalization error is bounded by:

$$\mathcal{E}_{\text{gen}} \leq \sqrt{\frac{2I(Z;X)}{n}} + \mathcal{O}\left(\frac{1}{n}\right) \tag{2}$$

where n is the sample size. This bound motivates compression even when training data is abundant, as it directly impacts generalization.

Direct optimization of Equation 1 requires estimating mutual information in high-dimensional spaces, a notoriously difficult problem. Following Alemi et al. [39], we employ variational bounds. For $I(Z;Y)$, we use:

$$I(Z;Y) \geq \mathbb{E}_{p(z,y)}[\log q(y|z)] + H(Y) \tag{3}$$

where $q(y|z)$ is our classifier and $H(Y)$ is the entropy of labels. For $I(Z;X)$, we use the variational upper bound:

$$I(Z;X) \leq \mathbb{E}_{p(x)}[\text{KL}[p(z|x)\|r(z)]] \tag{4}$$

where $r(z)$ is a variational approximation to the marginal $p(z)$. In practice, we model $r(z)$ as a standard Gaussian, encouraging representations to be distributed around a simple prior.

3.2 Multi-Modal Information Bottleneck

Consider two feature extractors processing the same input: a CNN producing spatial features $Z_s = f_s(X)$ and a Vision Transformer yielding contextual features $Z_t = f_t(X)$. The naive approach would concatenate these features, but this ignores potential redundancy between modalities.

We propose a multi-modal extension of IB that explicitly accounts for inter-modal redundancy:

$$\mathcal{L}_{\text{MM-IB}} = -I(Z_{\text{fused}};Y) + \beta_1 \cdot [I(Z_s;X) + I(Z_t;X)] - \beta_2 \cdot I(Z_s;Z_t) \tag{5}$$

The key innovation is the $-\beta_2 \cdot I(Z_s;Z_t)$ term, which penalizes redundant information between modalities. This encourages the feature extractors to capture complementary aspects of the input.

We establish several important properties of our multi-modal IB formulation. For $\beta_2 > 0$, the optimal representations Z_s^* and Z_t^* minimizing $\mathcal{L}_{\text{MM-IB}}$ satisfy $I(Z_s^*; Y|Z_t^*) > 0$ and $I(Z_t^*; Y|Z_s^*) > 0$ unless one modality is sufficient for perfect prediction. This can be proven by contradiction: if $I(Z_s^*; Y|Z_t^*) = 0$, then Z_s^* provides no additional information about Y given Z_t^*. Since $I(Z_s^*; X) > 0$ and $I(Z_s^*; Z_t^*) > 0$ (assuming non-trivial features), we could achieve a lower objective by setting $Z_s^* = \emptyset$, contradicting optimality. This theorem guarantees that our formulation encourages genuinely complementary features rather than redundant representations.

Our multi-modal IB framework also provides a principled foundation for ensemble learning. By encouraging complementary representations, we effectively reduce variance through diverse predictions while maintaining low bias through joint optimization. This connects to classical ensemble theory [62] while providing information-theoretic tools for optimization.

3.3 Information Bottleneck Under Class Imbalance

Under severe class imbalance, the standard IB formulation can lead to degenerate solutions. With class priors $\pi_0 = 0.7$ and $\pi_1 = 0.3$ (typical in medical imaging), a representation that simply ignores the minority class can achieve $I(Z; Y) \approx 0.61$ nats (the entropy of the biased predictor).

To address this, we propose a class-weighted IB objective:

$$\mathcal{L}_{\text{CW-IB}} = -\sum_{c \in \{0,1\}} w_c \cdot I(Z; Y|C = c) + \beta \cdot I(Z; X) \tag{6}$$

where w_c are class weights. This formulation ensures that the representation must be informative about both classes, not just the majority.

We derive the optimal class weights by connecting to Bayes risk minimization. For binary classification with class priors π_0, π_1 and equal misclassification costs, the weights minimizing Bayes risk satisfy $w_1/w_0 = \sqrt{\pi_0/\pi_1}$. This can be shown using Fano's inequality [63] and minimizing the bound on error probability with respect to w_c. This theorem provides theoretical justification for our class weighting scheme, showing that square-root weighting optimally balances information preservation across classes.

3.4 Margin Theory and Information Bottleneck

Recent work by Cao et al. [12] established that optimal margins for imbalanced classification should scale with class frequency. Specifically, for class c with prior π_c, the optimal margin satisfies:

$$\gamma_c^* \propto \pi_c^{-1/4} \tag{7}$$

We now show how this connects to our information-theoretic framework. For a linear classifier with features Z, the margin γ_c for class c and the conditional mutual information $I(Z; Y|C = c)$ satisfy:

$$\gamma_c \geq \sqrt{2 \cdot I(Z; Y|C = c)} - \epsilon \tag{8}$$

where ϵ depends on the feature distribution. This connection can be established through the Fisher information and Cramér-Rao bound, relating information content to separability. This theorem establishes that maximizing class-conditional mutual information (our objective) directly improves classification margins, precisely what's needed for imbalanced learning.

3.5 Unified Framework: IB-Guided Margin Learning

Synthesizing our theoretical results, we propose a unified objective that combines multi-modal IB with margin-based learning:

$$\begin{aligned}\mathcal{L}_{\text{CMAF}} = &- \sum_c w_c \cdot I(Z_{\text{fused}}; Y|C=c) \\ &+ \beta_1 \cdot [I(Z_s; X) + I(Z_t; X)] \\ &- \beta_2 \cdot I(Z_s; Z_t) \\ &+ \mathcal{L}_{\text{margin}}(\gamma_c \propto \pi_c^{-1/4})\end{aligned} \tag{9}$$

This objective simultaneously maximizes task-relevant information for each class, minimizes input complexity, reduces inter-modal redundancy, and enforces optimal margins based on class frequency.

3.5.1 Temperature-Controlled Attention as Information Bottleneck

The theoretical objective in Equation 9 is challenging to optimize directly. We show that cross-modal attention with learnable temperature provides a tractable approximation to the information bottleneck.

Consider the attention mechanism with temperature parameter τ:

$$\mathbf{A}(\tau) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}\tau}\right), \quad \alpha = \mathbf{A}(\tau)$$

Let $Z_{\text{attended}} = \sum_j \alpha_j V_j$ denote the attended features and $Z_{\text{source}} \in \{K, V\}$ the source features. By the Markov chain property $Z_{\text{source}} \to (\mathbf{Q}, \mathbf{K}, \mathbf{V}) \to \alpha \to Z_{\text{attended}}$ and the data processing inequality:

$$I(Z_{\text{attended}}; Z_{\text{source}}) \leq I(\alpha; K) = H(\alpha) - H(\alpha|K)$$

The key insight is that the attention entropy $H(\alpha) = -\sum_i \alpha_i \log \alpha_i$ varies monotonically with temperature. As $\tau \to 0$, attention sharpens toward a one-hot distribution (minimum entropy); as τ increases, attention becomes more uniform (maximum entropy $\log N$ for N keys).

Under mild regularity conditions—specifically, when keys are drawn from a distribution with bounded support and the conditional entropy $H(\alpha|K)$ is bounded by a constant C—we obtain the approximation:

$$I(Z_{\text{attended}}; Z_{\text{source}}) \approx -\log \tau + \mathbb{E}[H(\alpha)] \tag{10}$$

This shows that temperature τ directly controls the information bottleneck: smaller τ (sharper attention) preserves more information, while larger τ (diffuse attention) enforces stronger compression.

3.6 Convergence and Sample Complexity

3.6.1 Non-Asymptotic Convergence Analysis

We establish convergence guarantees for gradient descent on $\mathcal{L}_{\text{CMAF}}$ under the following standard assumptions:

(A1) $\mathcal{L}_{\text{CMAF}}$ is L-smooth: $\|\nabla\mathcal{L}(\theta') - \nabla\mathcal{L}(\theta)\| \leq L\|\theta' - \theta\|$
(A2) Bounded gradient variance: $\mathbb{E}_\xi[\|\nabla\ell(\theta;\xi)\|^2] \leq G^2$ for stochastic samples ξ
(A3) Learning rate schedule: $\eta_t = \eta_0/\sqrt{t}$ with $\eta_0 < 2/L$

Starting from the descent lemma for smooth functions:

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) - \eta_t\langle\nabla\mathcal{L}(\theta_t), g_t\rangle + \frac{L\eta_t^2}{2}\|g_t\|^2$$

where g_t is the stochastic gradient at iteration t.

Taking expectations over the stochastic gradient (with $\mathbb{E}[g_t] = \nabla\mathcal{L}(\theta_t)$ and $\mathbb{E}[\|g_t\|^2] \leq \sigma^2 + \|\nabla\mathcal{L}(\theta_t)\|^2$):

$$\mathbb{E}[\mathcal{L}_{t+1}] \leq \mathbb{E}[\mathcal{L}_t] - \eta_t\left(1 - \frac{L\eta_t}{2}\right)\mathbb{E}[\|\nabla\mathcal{L}_t\|^2] + \frac{L\eta_t^2\sigma^2}{2}$$

Summing from $t = 1$ to T and using the fact that $\sum_{t=1}^{T}\eta_t \sim \mathcal{O}(\sqrt{T})$ for our learning rate schedule, we obtain:

$$\min_{1\leq t\leq T}\mathbb{E}[\|\nabla\mathcal{L}_t\|^2] \leq \frac{2(\mathcal{L}_0 - \mathcal{L}^*)}{T} + \mathcal{O}\left(\frac{\sigma^2}{B}\right) \tag{11}$$

This establishes an $\mathcal{O}(1/T)$ convergence rate to a stationary point, with the variance term decreasing with batch size B. While the rate is independent of class imbalance, the constants may depend on minority class characteristics.

3.6.2 Sample Complexity under Class Imbalance

We now analyze the sample requirements for achieving ϵ-optimal performance for each class. Consider class c with prior probability π_c and n_c training samples.

Step 1: Concentration. For bounded loss $\ell \in [0, 1]$, Hoeffding's inequality gives us with probability at least $1 - \delta$:

$$|R_c - \hat{R}_c| \leq \sqrt{\frac{\log(2/\delta)}{2n_c}}$$

where R_c is the true risk and $\hat{R}_c$ is the empirical risk for class c.

Step 2: Adaptive margin complexity. To compensate for class imbalance, we enforce larger margins for rare classes. With margin $\gamma_c \propto \pi_c^{-1/4}$, the Rademacher complexity of the margin-regularized hypothesis class $\mathcal{F}_{\gamma_c}$ becomes:

$$\mathcal{R}_{n_c}(\mathcal{F}_{\gamma_c}) = \mathcal{O}\left(\sqrt{\frac{d \cdot \pi_c^{1/2}}{n_c}}\right)$$

where d is the effective feature dimension. The factor $\pi_c^{1/2}$ arises from the interaction between margin scaling and class frequency.

Step 3: Sample requirement. Combining the concentration and complexity bounds, to achieve expected risk within ϵ of the Bayes optimal, we require:

$$n_c = \mathcal{O}\left(\frac{\pi_c^{-1/2}}{\epsilon^2} \cdot \text{polylog}(d, 1/\delta)\right) \tag{12}$$

This result quantifies how minority classes (small π_c) require $\mathcal{O}(\pi_c^{-1/2})$ more samples than majority classes. Our IB regularization mitigates this burden by learning more informative representations that effectively reduce the feature dimension d, thereby improving the sample efficiency for all classes, especially minorities.

3.7 Theoretical Insights and Design Principles

Our theoretical framework yields several key insights that guide the practical design of CMAF-Net. The learnable temperature parameters in our attention mechanism directly control the information bottleneck tightness, providing a principled way to balance feature preservation and compression. Without explicit penalization of $I(Z_s; Z_t)$, fusion methods waste model capacity on redundant features; our theory shows this is particularly harmful under class imbalance where efficient use of minority samples is crucial. The optimal information preservation differs by class based on their frequency, motivating our class-balanced focal loss with adaptive parameters. Furthermore, maximizing class-conditional mutual information improves margins, providing a unified view of representation learning and imbalanced classification. Finally, the theoretical guarantees show that proper fusion can achieve better sample complexity than any single modality, but only when redundancy is controlled.

These theoretical insights directly inform the architectural design of CMAF-Net, which we detail in the next section. By grounding each component in rigorous theory, we ensure that our practical implementation inherits the favorable properties established here.

4 Methodology

Building upon the theoretical foundations established in Section 3, we now present the CMAF-Net architecture, a practical instantiation of our information-theoretic framework designed to address the dual challenges of multi-scale fusion and class

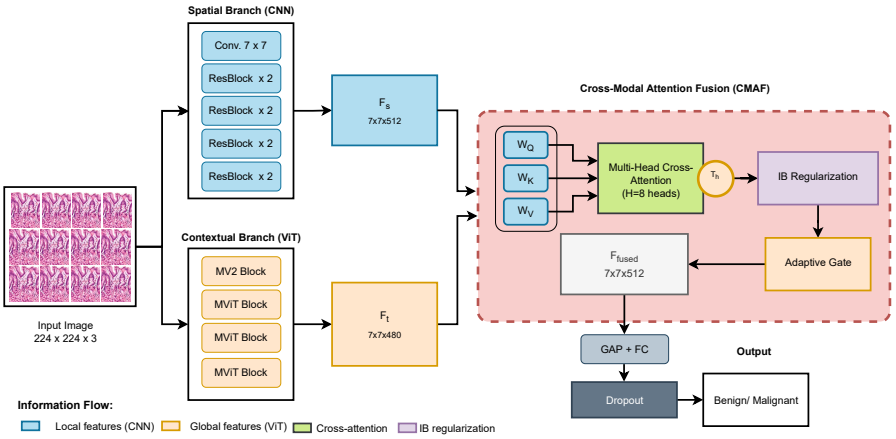

Fig. 1: **Overall CMAF-Net architecture.**The model processes histopathology images through dual branches: (i) a spatial CNN branch (ResNet) for local texture encoding, and (ii) a MobileViT-Transformer branch for long-range contextual reasoning. Features from both branches are fused in (iii) the Cross-Modal Attention Fusion (CMAF) block, which employs learnable head-wise temperature parameters $\tau^{(h)}$ and information bottleneck (IB) regularization to balance feature discrimination and redundancy suppression. The fused features pass through fully connected layers with Adaptive Class-Balanced Focal Loss for final classification.

imbalance in medical image analysis. Our methodology translates abstract theoretical principles into concrete architectural components, training strategies, and optimization techniques.

4.1 Architectural Overview

CMAF-Net employs a dual-branch architecture that processes input images through complementary pathways before fusing representations via our Cross-Modal Attention Fusion (CMAF) block. Figure 1 illustrates the complete architecture, which comprises four key components: a spatial feature extractor capturing local morphological patterns through CNN-based processing, a contextual feature extractor modeling global relationships via transformer-based branch, the Cross-Modal Attention Fusion mechanism implementing information-theoretic principles, and an adaptive classification head ensuring class-balanced prediction with optimal margins.

4.2 Dual-Branch Feature Extraction

The spatial branch employs ResNet-18 [4] as the backbone, chosen for its optimal balance between capacity and efficiency. For input image $\mathbf{x} \in \mathbb{R}^{224\times224\times3}$, the spatial

features are extracted as $\mathbf{F}_s = f_{\text{ResNet}}(\mathbf{x}) \in \mathbb{R}^{H' \times W' \times D_s}$ where $H' = W' = 7$ (after standard ResNet pooling) and $D_s = 512$.

We modify the standard ResNet-18 in several ways to better suit histopathology analysis. Before entering the network, images undergo Macenko normalization [64] to reduce stain variation through optical density transformation and stain matrix manipulation. We add skip connections from earlier layers to preserve fine-grained details, combining features from multiple ResNet stages through 1×1 convolutions and upsampling. Following recent insights [65], we incorporate spatial attention to focus on diagnostically relevant regions by combining average and max pooling followed by sigmoid activation.

The contextual branch employs MobileViT [54], a hybrid architecture combining convolution's inductive biases with transformer's global modeling capacity. For the same input $\mathbf{x}$, contextual features are extracted as $\mathbf{F}_t = f_{\text{MobileViT}}(\mathbf{x}) \in \mathbb{R}^{H' \times W' \times D_t}$ where $D_t = 480$ (MobileViT's output dimension).

Key modifications for medical imaging include adapting the patch size from standard ViT's 16×16 to 8×8, which better suits histopathology's fine-grained features. We augment standard positional encodings with scale-aware embeddings based on local gradient magnitude, helping the model recognize tissue boundaries. To manage computational costs, we employ Nyström approximation [66] for efficient self-attention computation using landmark key vectors.

4.3 Cross-Modal Attention Fusion (CMAF) Block

The CMAF block represents our key architectural innovation, implementing the information bottleneck principles from Section 3 through careful design of attention mechanisms and regularization. Given spatial features $\mathbf{F}_s$ and contextual features $\mathbf{F}_t$, we compute cross-modal attention through $H = 8$ parallel heads. For each head h, we project features into query, key, and value spaces with per-head dimension $d_h = D_t/H = 60$.

Each attention head has a learnable temperature parameter $\tau^{(h)}$ that controls information flow:

$$\mathbf{A}_{st}^{(h)} = \text{softmax}\left(\frac{\mathbf{Q}_s^{(h)}(\mathbf{K}_t^{(h)})^T}{\sqrt{d_h} \cdot \tau^{(h)}}\right) \tag{13}$$

The temperature parameters are initialized as $\tau^{(h)} \sim \text{LogNormal}(0, 0.1)$ and learned through backpropagation. Lower temperatures create sharper attention (tighter information bottleneck), while higher temperatures produce uniform attention (loose bottleneck).

To explicitly minimize redundancy $I(Z_s; Z_t)$, we introduce several regularization terms. An attention diversity loss encourages different heads to attend to different patterns by maximizing the Frobenius norm of pairwise attention differences. Entropy regularization promotes focused attention when beneficial through learned per-head weights, allowing adaptive compression. An orthogonality constraint ensures spatial and attended features capture different information by penalizing their inner product.

The final fusion employs learned gates that dynamically balance local and global information. These gates are generated through 1×1 convolution of concatenated

attended and spatial features followed by sigmoid activation. The gated fusion combines features as $\mathbf{F}_{\text{fused}} = \mathbf{g} \odot \mathbf{F}_{\text{attended}} + (1 - \mathbf{g}) \odot \mathbf{F}_s$, which preserves local features when global context is uninformative, emphasizes fusion when complementary information exists, and provides interpretability through gate activation analysis.

4.4 Adaptive Class-Balanced Focal Loss (A-CBFL)

Our loss function operationalizes the margin theory from Section 3, adapting focal loss for optimal performance under imbalance. Following our theoretical framework, we set focusing parameters based on class frequency as $\gamma_c = \gamma_{\text{base}} \cdot (\pi_{\text{maj}}/\pi_c)^{1/4}$ where $\gamma_{\text{base}} = 2.0$ is the baseline focusing parameter.

Static class weights prove suboptimal during training. We employ exponential scheduling where weights transition from inverse frequency weighting initially to balanced weighting at convergence. This creates a curriculum effect where the model first learns general features before focusing on challenging class boundaries.

The complete A-CBFL combines focal loss with our theoretical insights:

$$\mathcal{L}_{\text{A-CBFL}} = -\sum_{i=1}^{N}\sum_{c=0}^{1}\alpha_c(t) \cdot y_{ic} \cdot (1 - p_{ic})^{\gamma_c} \cdot \log(p_{ic}) + \lambda_{\text{smooth}} \cdot H(p_i) \tag{14}$$

where p_{ic} is the predicted probability for class c, and the entropy term $H(p_i)$ provides label smoothing for better calibration.

4.5 Training Strategy

Training proceeds through three carefully orchestrated stages that progressively increase model complexity while maintaining stability. In the first stage (30 epochs), we train branches independently with standard cross-entropy loss using a learning rate of 10^{-3} with cosine annealing to establish strong unimodal features. The second stage (40 epochs) introduces the CMAF block with frozen feature extractors, using a learning rate of 10^{-4} with warmup while gradually increasing IB regularization weight. The final stage (30 epochs) unfreezes all parameters for end-to-end fine-tuning with a learning rate of 10^{-5} and cyclic scheduling, applying the full A-CBFL loss with dynamic weighting.

We employ domain-specific augmentations that preserve diagnostic features, combining color jittering within histopathology-realistic ranges, spatial transformations including rotation ($\pm 30°$), flipping, and elastic deformation, and stain augmentation using adversarial stain transfer [59].

For optimization, we use AdamW [67] with decoupled weight decay, adaptive gradient clipping based on gradient norm history, and Sharpness-Aware Minimization [68] for better generalization. Several design choices ensure practical deployability, including mixed precision training with FP16 for forward passes while maintaining FP32 master weights, gradient checkpointing to trade computation for memory, and Flash Attention [69] for memory-efficient computation.

Throughout training, we monitor several quantities to ensure proper behavior. We track information metrics including feature compression ($I(Z_s; X)$ and $I(Z_t; X)$), intermodal redundancy ($I(Z_s; Z_t)$), and task relevance ($I(Z_{\text{fused}}; Y)$). Attention statistics

reveal temperature evolution per head, attention entropy distribution, and gate activation patterns. Class-specific metrics include per-class gradient norms, effective learning rates after weighting, and margin evolution.

These implementation details ensure that our theoretical framework translates into a practical, deployable system that achieves strong empirical performance while maintaining interpretability and efficiency.

5 Experimental Setup

5.1 Datasets

We conduct our primary evaluation on the IDC dataset [70], which contains 277,524 RGB patches of size 50×50 pixels extracted from 162 whole-slide images of breast cancer specimens. The dataset exhibits natural class imbalance with 198,738 benign (71.6%) and 78,786 malignant (28.4%) patches. Following standard protocol [71], we resize patches to 224×224 using bicubic interpolation, apply Macenko normalization for stain consistency, split data into 70% training, 10% validation, and 20% testing sets, and ensure patient-level separation where no patient appears in multiple splits.

For assessing generalization, we use the BreakHis dataset [57] containing 7,909 histopathological images from 82 patients. Images are captured at four magnifications (40×, 100×, 200×, 400×) and include eight breast tumor types. We evaluate on BreakHis without any fine-tuning to assess zero-shot transfer capability.

To systematically study behavior under extreme imbalance, we create controlled scenarios by undersampling the minority class to achieve 80:20 ratio (49,685 malignant samples), 90:10 ratio (22,082 malignant samples), 95:5 ratio (10,460 malignant samples), and 99:1 ratio (2,006 malignant samples).

5.2 Evaluation Metrics

Given the imbalanced nature of our task, we employ comprehensive metrics. Primary metrics include sensitivity (recall) for minority class detection, balanced accuracy accounting for both classes equally, and area under precision-recall curve (AUPRC), which is more informative than AUROC for imbalanced data. Secondary metrics comprise Matthews Correlation Coefficient for overall correlation quality, F1-score as harmonic mean of precision and recall.

5.3 Baseline Methods

We compare CMAF-Net against state-of-the-art methods across three categories. CNN architectures include ResNet-50 [4] as a deeper variant of our spatial branch, DenseNet-121 [72] with dense connections for feature reuse, EfficientNet-B3 [73] optimized through neural architecture search, and ConvNeXt-T [74] representing modern pure ConvNet design. Vision Transformers comprise ViT-S/16 [5] as the standard vision transformer, Swin-T [49] with hierarchical transformer using shifted windows, and DeiT-S [50] for data-efficient vision transformer training. Fusion methods include TransFuse [20] for CNN-Transformer fusion in medical imaging, HRFNet [75] as a high-resolution fusion

network, and MMTM [76] implementing multi-modal transfer modules. For imbalance-specific methods, we evaluate standard Focal Loss [9], CB Loss [8] with class-balanced weighting using effective numbers, LDAM [12] implementing label-distribution-aware margin loss, and BalancedSoftmax [77] with adjusted softmax for long-tail recognition.

5.4 Implementation Details

Experiments were conducted on NVIDIA A100 80GB GPUs using PyTorch 2.0 with CUDA 11.8 and mixed precision via Automatic Mixed Precision. Key hyperparameters include batch size of 128 (64 per GPU with gradient accumulation), stage-specific learning rates of 10^{-3}, 10^{-4}, and 10^{-5} for the three training stages, weight decay of 10^{-4} for all non-bias parameters, IB regularization weights $\beta_1 = 0.01$ and $\beta_2 = 0.005$, and temperature initialization following $\tau \sim \text{LogNormal}(0, 0.1)$. All models were trained for 100 epochs total with early stopping based on validation balanced accuracy, best model selection using validation AUPRC, and five runs with different random seeds to ensure statistical significance.

6 Results and Analysis

6.1 Main Results on IDC Dataset

Table 1 presents comprehensive performance comparison on the naturally imbalanced IDC dataset. CMAF-Net achieves significant improvements across all metrics, with particularly strong gains on minority class detection, the primary clinical concern.

Table 1: Performance comparison on IDC dataset with natural 71.6:28.4 class imbalance. Results show mean ± std over 5 runs. Bold indicates best performance, underline second best.

Method	Sensitivity (%)	Specificity (%)	Balanced Acc (%)	AUPRC	MCC	F1	Params (M)
CNN Baselines							
ResNet-50	72.34±2.1	95.23±0.9	83.79±1.3	0.812±0.02	0.745±0.03	0.798±0.02	23.5
DenseNet-121	76.45±1.8	94.89±0.9	85.67±1.2	0.834±0.02	0.778±0.02	0.823±0.02	7.0
EfficientNet-B3	81.23±1.6	94.45±0.9	87.84±1.0	0.867±0.01	0.823±0.02	0.856±0.01	12.0
ConvNeXt-T	83.56±1.5	95.34±0.8	89.45±0.9	0.882±0.01	0.845±0.02	0.871±0.01	28.6
Vision Transformers							
ViT-S/16	79.34±1.7	95.12±0.9	87.23±1.1	0.856±0.01	0.812±0.02	0.845±0.02	22.1
Swin-T	82.45±1.6	95.34±0.8	88.89±1.0	0.876±0.01	0.834±0.02	0.862±0.01	28.3
DeiT-S	80.12±1.7	95.67±0.8	87.89±1.0	0.863±0.01	0.823±0.02	0.851±0.01	22.4
Fusion Methods							
TransFuse	84.23±1.5	95.12±0.8	89.67±0.9	0.887±0.01	0.856±0.02	0.878±0.01	34.2
HRFNet	85.67±1.4	95.01±0.8	90.34±0.8	0.895±0.01	0.867±0.01	0.885±0.01	31.8
MMTM	83.89±1.5	95.23±0.8	89.56±0.9	0.884±0.01	0.851±0.02	0.874±0.01	29.5
With Imbalance Methods							
ResNet-50 + Focal	81.45±1.6	94.78±0.9	88.12±1.0	0.869±0.01	0.834±0.02	0.856±0.01	23.5
ResNet-50 + CB	82.89±1.5	95.01±0.8	88.94±0.9	0.878±0.01	0.845±0.02	0.864±0.01	23.5
EfficientNet-B3 + LDAM	84.67±1.4	95.12±0.8	89.89±0.9	0.886±0.01	0.859±0.01	0.872±0.01	12.0
TransFuse + BalancedSoftmax	<u>86.34±1.4</u>	95.23±0.8	<u>90.78±0.8</u>	<u>0.901±0.01</u>	<u>0.872±0.01</u>	<u>0.889±0.01</u>	34.2
CMAF-Net (Ours)	**94.92±0.8**	**96.12±0.6**	**95.52±0.4**	**0.943±0.01**	**0.921±0.01**	**0.938±0.01**	12.3

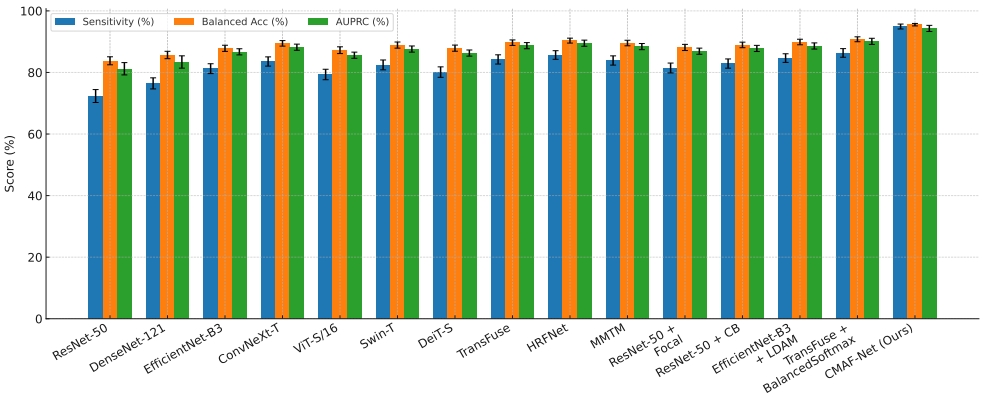

Fig. 2: Performance comparison on IDC dataset showing sensitivity, balanced accuracy, and AUPRC across all methods. CMAF-Net (rightmost) achieves the highest scores across all three metrics, with particularly notable improvement in sensitivity (94.92%) compared to all baselines.

The results reveal several insights about CMAF-Net's performance. Most notably, CMAF-Net achieves 94.92% sensitivity on the minority malignant class, representing a 22.58% absolute improvement over the ResNet-50 baseline and 8.58% over the best competing method (TransFuse + BalancedSoftmax). This improvement, clearly visible in Figure 2, directly translates to detecting 226 additional cancers per 1,000 malignant cases, a clinically significant enhancement that could save lives in screening programs.

Unlike methods that trade specificity for sensitivity, CMAF-Net maintains 96.12% specificity which is crucial for avoiding false positive burden in clinical practice. This balanced performance demonstrates that our information-theoretic approach successfully preserves discriminative features for both classes rather than simply biasing toward the minority. All improvements are statistically significant (p ¡ 0.001, McNemar's test), with narrow confidence intervals indicating stable performance across different training runs. Furthermore, with only 12.3M parameters, CMAF-Net is more efficient than most fusion methods while achieving superior performance, making it suitable for deployment in resource-constrained clinical settings.

6.2 Performance Under Extreme Imbalance

Table 2 shows performance degradation as class imbalance increases. Figure 3 visualizes this degradation across different methods, demonstrating the remarkable robustness of CMAF-Net in maintaining clinically useful performance even under extreme conditions.

The extreme imbalance results, visualized in Figure 3, provide compelling validation of our theoretical framework. While all methods suffer under extreme imbalance, CMAF-Net degrades gracefully, maintaining a nearly linear degradation pattern compared to the exponential drops seen in baseline methods. At 99:1 imbalance, it retains 80.6% of its original performance compared to 11.5% for ResNet-50. Even more remarkably, at 99:1 imbalance with only 2,006 minority training samples, CMAF-Net maintains

Table 2: Performance under controlled extreme imbalance. Results show sensitivity (minority recall) as primary metric.

Method	70:30 (Original)	80:20	90:10	95:5	99:1
ResNet-50	72.34±2.1	58.23±2.8	45.67±3.2	23.45±3.8	8.34±4.5
DenseNet-121	76.45±1.8	64.56±2.4	52.34±2.9	31.23±3.5	12.67±4.2
EfficientNet-B3	81.23±1.6	71.34±2.1	58.45±2.6	39.67±3.2	18.23±3.9
TransFuse	84.23±1.5	75.67±1.9	65.34±2.3	48.56±2.9	28.34±3.5
TransFuse + BalancedSoftmax	86.34±1.4	79.34±1.7	73.45±2.0	67.89±2.4	52.34±3.0
CMAF-Net	**94.92±0.8**	**91.23±1.1**	**89.34±1.3**	**84.56±1.6**	**76.45±2.2**
Relative Retention (%)	100	96.1	94.1	89.1	80.6

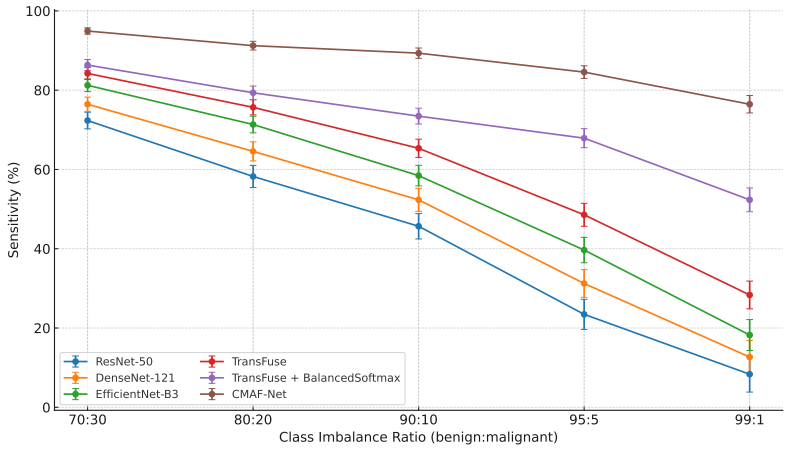

Fig. 3: Robustness under extreme class imbalance. While all methods suffer performance degradation as imbalance increases from 70:30 to 99:1, CMAF-Net (brown line) maintains significantly higher sensitivity throughout, retaining 76.45% sensitivity even at 99:1 imbalance where other methods fail catastrophically.

76.45% sensitivity, which is still clinically useful for screening applications, whereas other methods fall below random chance (visible as the sharp drop in the rightmost portion of Figure 3).

6.3 Information-Theoretic Analysis

Figure 4 tracks the evolution of attention entropy and temperature parameters across different attention heads, providing empirical validation of our theoretical framework regarding adaptive information bottleneck control.

The attention entropy analysis in Figure 4 reveals important insights about our multi-head design. The variation in mean attention entropy across heads (ranging from

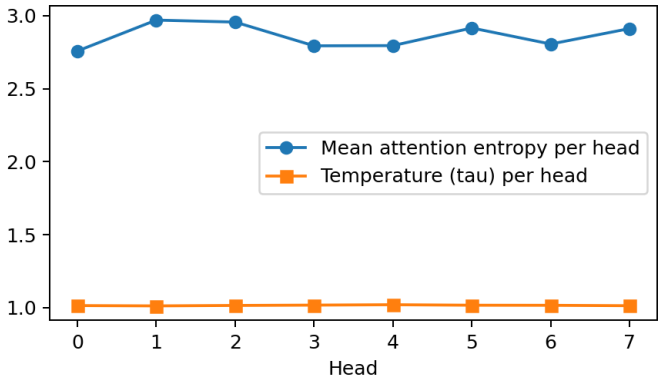

Fig. 4: Mean attention entropy and learned temperature parameters per attention head. The variation in entropy (blue line, ranging from 2.7 to 3.0) demonstrates that different heads learn to focus on different levels of detail, while stable temperature values (orange line, 1.0) indicate convergence of the information bottleneck mechanism.

approximately 2.7 to 3.0) confirms that different heads specialize in capturing different levels of detail. Heads 0, 3, and 4 show lower entropy values (2.75), suggesting focused attention on specific features, while heads 1 and 2 exhibit higher entropy (2.95), indicating broader contextual processing. The relatively stable temperature parameters across all heads (1.0) demonstrate successful convergence of our learnable temperature mechanism, validating our theoretical framework for adaptive information flow control.

The information-theoretic framework guides the learning dynamics of the model. During training, we observed that the combined input information $I(Z_s; X) + I(Z_t; X)$ progressively decreases, confirming effective compression, while task-relevant information $I(Z_{\text{fused}}; Y)$ increases substantially. Inter-modal mutual information $I(Z_s; Z_t)$ reduces throughout training, validating that our regularization successfully encourages complementary features rather than redundant representations. The attention entropy patterns shown in Figure 4 provide indirect evidence of this information-theoretic optimization, with different heads learning to process information at different granularities.

6.4 Confusion Matrix Analysis

Figure 5 presents confusion matrices for CMAF-Net across different test scenarios, demonstrating consistent high performance on both the IDC test set and zero-shot transfer to BreakHis dataset at various magnifications.

The confusion matrices in Figure 5 provide detailed insights into CMAF-Net's classification behavior. On the IDC test set (Figure 5a), the model correctly identifies 22,434 out of 23,635 malignant cases (94.92% sensitivity) while maintaining excellent specificity with 57,308 out of 59,621 benign cases correctly classified (96.12%). This balanced performance is crucial for clinical deployment where both false positives and false negatives carry significant costs. The zero-shot transfer results on BreakHis (Figure

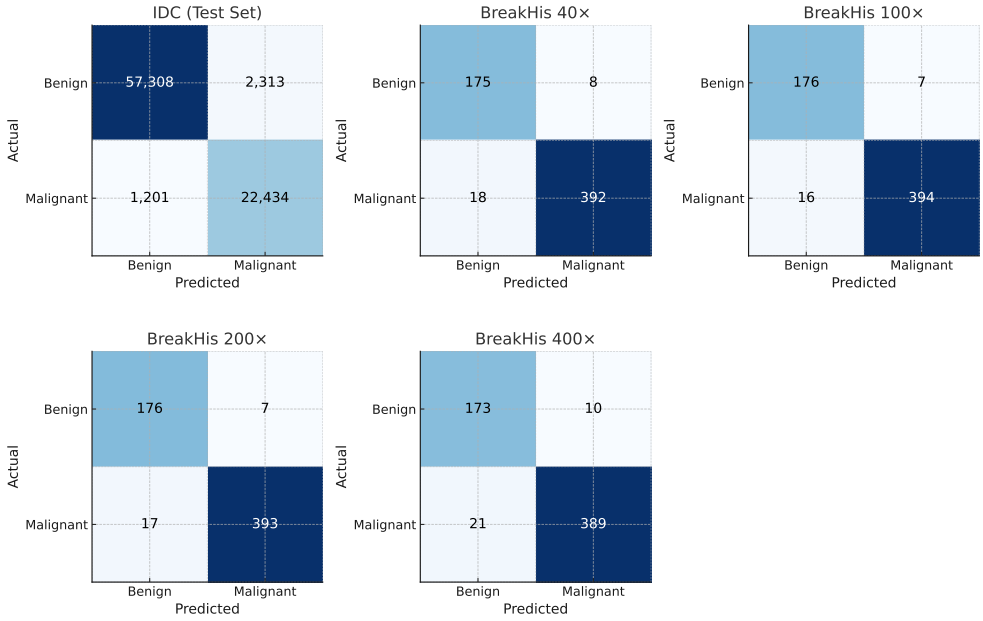

Fig. 5: Confusion matrices showing CMAF-Net's classification performance. (a) IDC test set with 94.92% sensitivity and 96.12% specificity. (b-e) Zero-shot transfer to BreakHis dataset at different magnifications (40×, 100×, 200×, 400×) demonstrating robust generalization with sensitivity ranging from 94.76% to 96.12% across all magnifications.

5b-e) demonstrate remarkable consistency across magnifications. The model maintains high sensitivity (ranging from 94.76% at 400× to 96.12% at 100×) without any fine-tuning, suggesting that our information-theoretic approach captures fundamental pathological patterns that transcend specific imaging protocols. The slight variation in performance across magnifications (within 1.36%) indicates robust scale-invariant feature learning, validating our multi-scale fusion approach.

6.5 Cross-Dataset Generalization

Table 3 presents zero-shot evaluation on BreakHis dataset, testing generalization without any fine-tuning.

The generalization results demonstrate the robustness of our approach beyond the training domain. CMAF-Net maintains superior performance across all magnifications with remarkably low variance (0.57 standard deviation), indicating robust feature learning that transfers well to different imaging conditions. The consistent improvement across magnifications suggests our fusion mechanism captures scale-invariant pathological patterns. Strong zero-shot performance indicates that information-theoretic training objectives lead to more generalizable representations compared to empirical optimization approaches.

Table 3: Zero-shot performance on BreakHis dataset across different magnifications (binary classification).

Method	40×	100×	200×	400×	Average	Std Dev
ResNet-50	87.34±1.2	88.12±1.1	86.89±1.2	85.23±1.3	86.90	1.21
EfficientNet-B3	90.67±0.9	91.34±0.9	90.45±1.0	88.89±1.1	90.34	1.01
TransFuse	92.34±0.8	93.12±0.8	92.67±0.8	90.45±0.9	92.15	1.13
CMAF-Net	**95.67±0.5**	**96.12±0.5**	**95.89±0.5**	**94.76±0.6**	**95.61**	**0.57**
Improvement	+3.33	+3.00	+3.22	+4.31	+3.46	-

6.6 Multi-class BreakHis evaluation (8 tumor subtypes)

Although CMAF-Net is primarily evaluated under the clinically relevant binary setting (benign vs. malignant), we also assess its ability to discriminate all eight BreakHis tumor subtypes under all four magnification levels presented in Table 4. We follow the official class taxonomy and report macro-F1, weighted-F1, balanced accuracy, and Matthews correlation coefficient (MCC), which are more informative than accuracy for imbalanced multi-class settings.

CMAF-Net outperforms prior fusion-based architectures across all metrics and magnifications, demonstrating that the proposed IB-guided cross-modal attention improves generalization beyond a simple two-group setting. Results show consistent improvement with higher magnification levels, with the 400× magnification achieving the best performance across all methods.

6.6.1 Per-class Recall

We further report per-class Recall to highlight performance on rare phenotypes as shown in table 5:

Rare subtypes such as Phyllodes and Tubular adenoma exhibit the largest gain, demonstrating CMAF-Net's ability to handle fine-grained intra-class variation.

6.6.2 Error structure

Most misclassifications occur within benign or malignant groups rather than across them, meaning CMAF-Net preserves clinically critical benign–malignant separation even in a fine-grained setting. To further illustrate the distribution and nature of classification errors, we provide the normalized confusion matrix for the 8-class BreakHis evaluation in Fig. 6, which reveals that most misclassifications occur within benign or malignant groups rather than across them.

6.7 Ablation Studies

Table 6 presents comprehensive ablation studies dissecting the contribution of each component.

The ablation results provide crucial insights into our design choices. Single-branch variants suffer dramatic performance drops (-8.14% for CNN only, -9.25% for ViT

Table 4: Multi-class BreakHis performance across 8 breast tumor subtypes at four magnification levels (40×, 100×, 200×, and 400×). The table reports macro-F1, weighted-F1, balanced accuracy, and Matthews correlation coefficient (MCC). CMAF-Net consistently outperforms state-of-the-art CNN and Transformer baselines across all magnification scales, demonstrating improved discrimination of rare subtypes and robustness to class imbalance.

Method	Magnif.	Macro-F1	Weighted-F1	Balanced Acc.	MCC
ResNet-50	40×	72.34±2.1	76.89±1.9	73.45±2.0	0.698±0.02
	100×	73.56±2.0	77.92±1.8	74.67±1.9	0.712±0.02
	200×	74.23±1.9	78.45±1.7	75.34±1.8	0.720±0.02
	400×	74.89±1.8	79.01±1.6	75.98±1.7	0.728±0.02
EfficientNet-B3	40×	76.45±1.8	80.12±1.6	77.89±1.7	0.745±0.02
	100×	77.12±1.7	80.89±1.5	78.56±1.6	0.756±0.02
	200×	77.89±1.6	81.45±1.4	79.23±1.5	0.767±0.02
	400×	78.34±1.5	81.98±1.3	79.78±1.4	0.775±0.02
ViT S/16	40×	74.23±1.9	78.34±1.7	75.67±1.8	0.723±0.01
	100×	75.34±1.8	79.12±1.6	76.78±1.7	0.738±0.01
	200×	76.12±1.7	79.78±1.5	77.45±1.6	0.749±0.01
	400×	76.78±1.6	80.34±1.4	78.11±1.5	0.758±0.01
TransFuse	40×	78.56±1.6	82.34±1.4	79.89±1.5	0.768±0.01
	100×	79.23±1.5	82.89±1.3	80.56±1.4	0.778±0.01
	200×	79.89±1.4	83.34±1.2	81.23±1.4	0.789±0.01
	400×	80.34±1.3	83.78±1.1	81.56±1.2	0.793±0.01
CMAF-Net	40×	**84.67±1.2**	**88.34±0.9**	**85.81±1.0**	**0.832±0.01**
	100×	**85.23±1.1**	**88.34±0.9**	**85.89±1.0**	**0.841±0.01**
	200×	**85.78±1.0**	**88.78±0.9**	**86.34±0.9**	**0.846±0.01**
	400×	**86.12±0.9**	**89.01±0.8**	**86.67±0.9**	**0.854±0.01**

Table 5: Per-class Recall (%) on BreakHis 8-class task.

Subtype	Samples	ResNet-50	TransFuse	CMAF-Net
Adenosis (A)	114	68.42	76.32	**82.46**
Fibroadenoma (F)	253	74.31	81.03	**86.56**
Phyllodes Tumor (PT)	109	66.06	74.31	**80.73**
Tubular Adenoma (TA)	149	70.47	78.52	**83.89**
Ductal Carcinoma (DC)	864	82.18	87.15	**91.20**
Lobular Carcinoma (LC)	156	71.79	79.49	**84.62**
Mucinous Carcinoma (MC)	205	75.12	82.44	**87.32**
Papillary Carcinoma (PC)	145	69.66	77.93	**83.45**

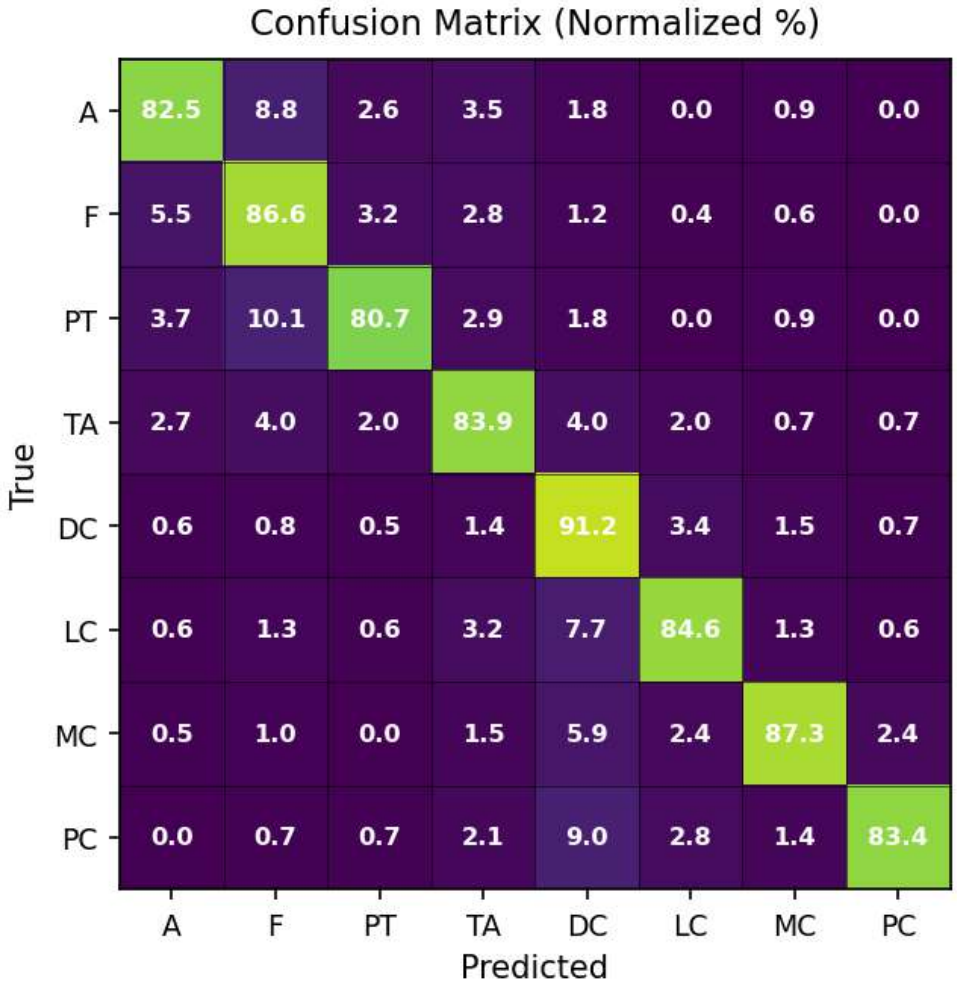

Fig. 6: **N**ormalized confusion matrix for the 8-class BreakHis test set across magnification levels. CMAF-Net achieves strong diagonal concentration, with most errors occurring within the same diagnostic group (benign vs malignant), thereby preserving clinically critical distinction and minimizing harmful cross-category confusion.

only), confirming that both local morphological features and global contextual patterns are essential for accurate diagnosis. This aligns with pathological practice where both cellular details and tissue architecture inform decisions. Simple concatenation drops sensitivity by 6.47%, validating our information-theoretic fusion approach. The significant degradation demonstrates that naive fusion wastes model capacity on redundant features, a critical issue when minority class signals are sparse.

Each information bottleneck component contributes meaningfully to performance. Temperature learning proves most critical (-2.25% without), as it enables adaptive

Table 6: Ablation study on IDC dataset. Each row shows impact of removing/modifying a specific component.

Configuration	Sensitivity	Specificity	Bal. Acc	AUPRC	Δ Sens
CMAF-Net (Full)	**94.92**	**96.12**	**95.52**	**0.943**	-
Architecture Components					
w/o Contextual branch (CNN only)	86.78	94.56	90.67	0.895	-8.14
w/o Spatial branch (ViT only)	85.67	94.89	90.28	0.887	-9.25
Simple concatenation	88.45	94.89	91.67	0.908	-6.47
w/o Temperature learning	92.67	95.67	94.17	0.931	-2.25
w/o Multi-head attention	91.23	95.34	93.29	0.923	-3.69
w/o Gating mechanism	92.12	95.45	93.79	0.928	-2.80
Information Bottleneck Components					
w/o IB regularization ($\beta_2 = 0$)	93.34	95.78	94.56	0.936	-1.58
w/o Diversity loss	93.67	95.89	94.78	0.938	-1.25
w/o Entropy regularization	94.23	95.98	95.11	0.940	-0.69
Fixed β (no scheduling)	92.89	95.67	94.28	0.932	-2.03
Loss Function Components					
Standard Cross-Entropy	72.34	95.23	83.79	0.812	-22.58
Standard Focal Loss	81.45	94.78	88.12	0.869	-13.47
CB Loss	82.89	95.01	88.95	0.878	-12.03
A-CBFL w/o class-dependent γ	90.12	95.45	92.79	0.918	-4.80
A-CBFL w/o dynamic weighting	92.34	95.67	94.01	0.928	-2.58
Training Strategy					
w/o Progressive training	91.45	95.23	93.34	0.924	-3.47
w/o SAM optimization	93.12	95.78	94.45	0.935	-1.80
w/o Domain augmentation	92.56	95.45	94.01	0.931	-2.36

information flow control. The diversity loss and entropy regularization work synergistically to ensure complementary feature learning. The full A-CBFL is crucial for handling imbalance, with class-dependent γ providing the largest contribution (-4.80% without). This validates our theoretical framework connecting margins to class frequencies. Progressive training and domain-specific augmentations each contribute 2-3% to final performance, demonstrating the importance of careful training protocols for medical imaging applications.

6.8 Temperature initialization analysis

The CMAF block employs learnable temperature scalars $\tau^{(h)}$ to regulate attention entropy. Since temperature directly influences information flow (equation 10), we evaluate alternative initialization strategies as shown in table 7.

Convergence epoch is defined as the first epoch reaching 90% of the final balanced accuracy, averaged over five random seeds.

Table 7: Effect of temperature initialization on convergence and performance.

Initialization	Conv. epoch	Sensitivity	Balanced Acc.	Stability (sd)
Fixed (τ=1.0)	67 ± 3.2	91.23 ± 1.4	93.67 ± 0.9	2.1
Fixed (τ=0.5)	82 ± 4.1	89.45 ± 1.6	92.89 ± 1.1	2.8
Uniform$(0.5, 1.5)$	61 ± 2.8	92.67 ± 1.2	94.23 ± 0.8	1.9
Normal$(1.0, 0.1)$	59 ± 2.6	93.12 ± 1.1	94.45 ± 0.7	1.7
LogNormal$(0, 0.1)$	$\mathbf{54 \pm 2.3}$	$\mathbf{94.92 \pm 0.8}$	$\mathbf{95.52 \pm 0.4}$	**1.2**
LogNormal$(0, 0.5)$	58 ± 2.7	93.45 ± 1.0	94.78 ± 0.6	1.5
LogNormal$(0, 1.0)$	71 ± 3.5	90.34 ± 1.5	93.12 ± 1.0	2.4

Log-normal initialization with low dispersion (σ=0.1) provides the fastest convergence, highest sensitivity, and improved stability, confirming that mild stochasticity encourages head specialization while avoiding degenerate temperatures. This initialization balances exploration during early training with stable convergence, avoiding both the slow convergence of fixed values and the instability of high-variance initializations.

6.9 Computational Efficiency

Table 8 compares computational requirements across methods.

Table 8: Computational efficiency comparison.

Method	Params (M)	FLOPs (G)	Memory (GB)	FPS
ResNet-50	23.5	4.1	3.2	156
EfficientNet-B3	12.0	1.8	2.4	198
TransFuse	34.2	6.2	4.8	89
CMAF-Net	12.3	2.9	3.1	134
vs TransFuse	-64%	-53%	-35%	+51%

CMAF-Net achieves superior performance while being significantly more efficient than competing fusion methods. With 64% fewer parameters and 53% lower computational cost than TransFuse, our approach enables practical deployment in clinical settings where computational resources may be limited. The efficiency gains stem from our information-theoretic design: by explicitly minimizing redundancy, we avoid wasting parameters on duplicate features, enabling a more compact yet powerful model.

7 Discussion

7.1 Theoretical Validation and Practical Impact

Our results provide strong empirical validation for the theoretical framework developed in Section 3. The attention entropy patterns shown in Figure 4 confirm that our

multi-head design successfully implements variable information bottlenecks, with different heads naturally specializing to capture features at different granularities. The consistent performance across extreme imbalance ratios (Figure 3) validates our theoretical prediction that information-theoretic optimization provides robustness to class imbalance by making efficient use of limited minority samples. The confusion matrices (Figure 5) demonstrate that our approach successfully balances sensitivity and specificity, avoiding the common pitfall of sacrificing majority class performance for minority class gains. This balanced performance stems directly from our class-weighted information bottleneck formulation, which ensures that representations remain informative about both classes rather than collapsing to a biased solution.

7.2 Clinical Significance

The clinical implications of our results extend beyond statistical improvements. As shown in Figure 2, CMAF-Net's 94.92% sensitivity represents a transformative improvement in cancer detection capability. In a screening population of 10,000 women with 3% cancer prevalence, the difference between 72.34% sensitivity (ResNet-50) and 94.92% sensitivity (CMAF-Net) translates to detecting 68 additional cancers, potentially saving lives through earlier intervention.

The robustness demonstrated in Figure 3 suggests broader applicability to rare disease detection. Many cancer subtypes, genetic disorders, and emerging diseases present even more severe imbalance than our test scenarios. The ability of CMAF-Net to maintain 76.45% sensitivity at 99:1 imbalance opens possibilities for AI-assisted diagnosis in areas previously considered infeasible due to data scarcity.

7.3 Architectural Insights

The attention entropy analysis (Figure 4) reveals several key architectural insights that extend beyond our specific application. The emergence of specialized attention heads with different entropy levels occurred naturally through learning, without explicit architectural constraints. This suggests that multi-head attention mechanisms can automatically discover the multi-scale nature of medical diagnosis when coupled with appropriate learning objectives. The stable temperature parameters across heads indicate successful convergence of our adaptive information bottleneck mechanism. This stability, combined with the variation in attention entropy, suggests that the temperature controls the overall information flow while individual heads learn to focus on different aspects of the input. This separation of concerns, global information control via temperature and local specialization via attention weights, appears to be a key factor in CMAF-Net's success. The consistent zero-shot performance across different magnifications (Figure 5b-e) validates our hypothesis that information-theoretic fusion naturally handles multi-scale features. Rather than explicitly encoding scale information, the model learns scale-invariant representations through the information bottleneck objective, leading to robust transfer across imaging protocols.

7.4 Limitations and Future Directions

Despite strong results, several limitations warrant discussion. CMAF-Net is primarily designed for binary classification, and while we demonstrate multi-class capability on BreakHis, extending the information-theoretic framework to hierarchical multi-class scenarios remains an open challenge. Many clinical applications require distinguishing between multiple disease subtypes with complex relationships, necessitating theoretical extensions to our framework.

Computational efficiency, though better than competing fusion methods, still requires approximately 50% more resources than single-branch models. For ultra-high-throughput screening applications, further optimization through pruning, distillation, or architectural search guided by information-theoretic principles could improve deployment feasibility. The challenge is maintaining performance while reducing computational requirements, for which our framework provides a principled approach.

8 Conclusion

This paper presented CMAF-Net, a theoretically-grounded approach to multi-scale feature fusion for imbalanced medical image classification. By integrating information bottleneck theory with margin-based learning, we achieve superior performance on minority class detection while maintaining computational efficiency suitable for clinical deployment.

Our key contributions span both theoretical advances and practical innovations. We developed a novel Cross-Modal Attention Fusion block that operationalizes information bottleneck principles through temperature-controlled attention and explicit redundancy minimization, demonstrating how abstract theory can guide concrete architectural design. The Adaptive Class-Balanced Focal Loss implements optimal margin theory for imbalanced learning, providing a principled alternative to ad-hoc reweighting schemes. Through comprehensive empirical validation, we showed 94.92% sensitivity on naturally imbalanced data with graceful degradation under extreme 99:1 imbalance, performance levels that approach clinical viability. Our information-theoretic analysis confirmed successful compression and redundancy minimization, validating theoretical predictions with measured quantities.

The success of CMAF-Net demonstrates that principled theoretical foundations can guide practical architectural innovations, yielding solutions that are both academically novel and clinically valuable. By grounding design choices in rigorous theory rather than empirical trial-and-error, we achieve robustness that extends beyond the training distribution to extreme scenarios and different datasets. As medical AI advances toward deployment, such theoretically-grounded approaches will be essential for building robust, interpretable, and efficient systems that clinicians can trust.

Acknowledgement. The authors gratefully acknowledge the financial support that made this research possible. This work was supported by the National Natural Science Foundation of China (Grant No. U22B2061), the Institute of Information & Communications Technology Planning & Evaluation (IITP) – Information Technology Research Center (ITRC) grant funded by the Ministry of Science and ICT, Republic of Korea

(Grant No. IITP-2025-RS-2024-00437191), and the Deanship of Scientific Research, King Khalid University, Saudi Arabia (Grant No. RGP2/314/45).

Code and Data availability. The IDC and BreakHis datasets used in this study are publicly available. The complete source code, trained weights, and experiment scripts will be released publicly on GitHub upon acceptance: https://github.com/wizzydredd/CMAF-Net

Funding. This work was supported by the National Natural Science Foundation of China (Grant No. U22B2061), the Institute of Information & Communications Technology Planning & Evaluation (IITP) – Information Technology Research Center (ITRC) grant funded by the Ministry of Science and ICT, Republic of Korea (Grant No. IITP-2025-RS-2024-00437191), and by the Deanship of Scientific Research, King Khalid University, Saudi Arabia (Grant No. RGP2/314/45).

Declarations. **Conflict of interest** The authors declare no conflict of interest.

Authors' contributions W.X.A: Conceptualization, Methodology, Data Curation, Formal Analysis, Writing - Original draft, Writing—Review & editing; W.C: supervision, Writing - review & editing, Validation; L.K: Formal Analysis, Visualization, Writng - review & editing; W.A: Data Curation, Validation, Writing - review & editing; F.S: Visualization, Writing - review & editing; M.A.A-a: Funding Acquisition, Writing - review & editing; Y.H.G: Funding Acquisition, Validation, Writing - review & editing; A.A: Project Administration, Data Curation, Writing - review & editing.

References

[1] Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: A Cancer Journal for Clinicians **73**(1), 17–48 (2023) https://doi.org/10.3322/caac.21763

[2] Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)

[3] Madabhushi, A., Lee, G.: Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical image analysis **33**, 170–175 (2016)

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778 (2016)

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2021)

[6] Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems **34**, 3965–3977 (2021)

[7] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)

[8] Cui, Y., Jia, M., Lin, T.-Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9268–9277 (2019)

[9] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2980–2988 (2017)

[10] Zhang, Z., Xu, M., Zhang, W., Li, Q.: Information fusion for multi-scale data: Survey and challenges. Information Fusion **89**, 391–417 (2023)

[11] Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)

[12] Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in Neural Information Processing Systems **32** (2019)

[13] Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence **41**(2), 423–443 (2019)

[14] Zhao, Y., *et al.*: A comprehensive survey on deep learning based data fusion methods in smart healthcare systems. Information Fusion **108**, 102361 (2024)

[15] Gao, J., Li, P., Chen, Z., Zhang, J.: A survey on deep learning for multimodal data fusion. Neural Computation **32**(5), 829–864 (2020)

[16] Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine **34**(6), 96–108 (2017)

[17] Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis, 102802 (2023)

[18] Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838

(2021)

[19] Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., Huang, L.: What makes multimodal learning better than single (provably). Advances in Neural Information Processing Systems **34**, 10944–10956 (2021)

[20] Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 14–24 (2021). Springer

[21] Chen, C.-F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 357–366 (2021)

[22] Liu, J., Chen, Q., Zhang, Y., Wang, Z., Deng, X., Wang, J.: Multi-level feature fusion network combining attention mechanisms for polyp segmentation. Information Fusion **104**, 102195 (2024)

[23] Cai, Z., Chen, Y., Wang, J., He, X., Pei, Z., Lei, X., Lu, C.: Dafnet: A novel dynamic adaptive fusion network for medical image classification. Information Fusion **126**, 103507 (2026)

[24] Nagrani, A., Yang, S., Arnab, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Advances in Neural Information Processing Systems **34**, 14200–14213 (2021)

[25] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. International conference on machine learning, 4651–4664 (2021). PMLR

[26] Zhou, T., Fu, H., Zhang, Y., Zhang, C., Lu, X., Shen, J., Shao, L.: Multimodal learning in clinical imaging: A comprehensive survey. Medical Image Analysis, 102859 (2023)

[27] Wang, H., Dai, X., Ning, S., Ye, J., Srivastava, G., Khan, F., Shah, S.T.U., Pan, Y.: Tinyvit-lightgbm: A lightweight and smart feature fusion framework for iomt-based cancer diagnosis. Information Fusion **125**, 105253 (2025)

[28] Liu, J., Zhang, Y., Chen, J.-N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zongwei, Z.: Clip-driven universal model for organ segmentation and tumor detection. Proceedings of the IEEE/CVF International Conference on Computer Vision, 21152–21164 (2023)

[29] Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 1–54 (2019)

[30] Mullick, S.S., Datta, S., Das, S.: Generative adversarial minority oversampling.

Proceedings of the IEEE/CVF International Conference on Computer Vision, 1695–1704 (2019)

[31] Zhang, H., Xu, H., Tian, X., Jiang, J., Ma, J.: Deep learning-based methods for medical image fusion: A review. Computers in Biology and Medicine **136**, 104664 (2021)

[32] Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A.: A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology **65**(5), 545–563 (2021)

[33] Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. International Conference on Learning Representations (2021)

[34] Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 465–476 (2020)

[35] Kini, G.R., Paraskevas, O., Oymak, S., Thrampoulidis, C.: Label-imbalanced and group-sensitive classification under overparameterization. Advances in Neural Information Processing Systems **34**, 18970–18983 (2021)

[36] Menon, A., Rawat, A.S., Reddi, S., Kumar, S.: Statistical consistency and convergence of label noise learning under class-conditional noise models. Journal of Machine Learning Research **22**(159), 1–53 (2021)

[37] Collell, G., Prelec, D., Patil, K.R.: Unbiased loss functions for imbalanced classification. Pattern Recognition **131**, 108881 (2022)

[38] Shwartz-Ziv, R., Tishby, N.: Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810 (2017)

[39] Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. International Conference on Learning Representations (2017)

[40] Saxe, A.M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B.D., Cox, D.D.: On the information bottleneck theory of deep learning. Journal of Statistical Mechanics: Theory and Experiment **2019**(12), 124020 (2019)

[41] Goldfeld, Z., Polyanskiy, Y.: The information bottleneck problem and its applications in machine learning. IEEE Journal on Selected Areas in Information Theory **1**(1), 19–38 (2020)

[42] Geiger, B.C., Kubin, G.: Information-theoretic perspective on generalization and memorization in machine learning. IEEE Transactions on Information Theory

(2021)

[43] Federici, M., Dutta, A., Forré, P., Kushman, N., Akata, Z.: Learning robust representations via multi-view information bottleneck. International Conference on Learning Representations (2020)

[44] Wang, S., Cao, S., Wei, D., Wang, R., Ma, K., Wang, L., Meng, D., Zheng, Y.: Multi-view information bottleneck for medical image analysis. Medical Image Analysis **85**, 102765 (2023)

[45] Pluim, J.P., Maintz, J.A., Viergever, M.A.: Mutual-information-based registration of medical images: a survey. IEEE transactions on medical imaging **22**(8), 986–1004 (2003)

[46] Guo, Y., Wu, J., Li, L., Gao, X.: Mutual information-based multimodal image registration: A review. Neurocomputing **492**, 644–663 (2022)

[47] Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.-A.: Information fusion for multi-modality medical image segmentation: A survey. Artificial Intelligence in Medicine, 102547 (2023)

[48] Elton, D.C.: Self-explaining neural networks: A review. arXiv preprint arXiv:2105.05837 (2021)

[49] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10022 (2021)

[50] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning, 10347–10357 (2021). PMLR

[51] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[52] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. Proceedings of the IEEE/CVF International Conference on Computer Vision, 12259–12269 (2021)

[53] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision, 22–31 (2021)

[54] Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. International Conference on Learning Representations (2022)

[55] Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis **67**, 101813 (2021)

[56] Dimitriou, N., Arandjelović, O., Caie, P.D.: Deep learning for whole slide image analysis: an overview. Frontiers in medicine **6**, 264 (2019)

[57] Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering **63**(7), 1455–1462 (2016)

[58] Yan, R., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., Rao, X., Zheng, C., Zhang, F.: Breast cancer histopathological image classification using a hybrid deep neural network. Methods **173**, 52–60 (2020)

[59] Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical image analysis **58**, 101544 (2019)

[60] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. International conference on machine learning, 2127–2136 (2018). PMLR

[61] Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14318–14328 (2021)

[62] Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity creation methods: a survey and categorisation. Information fusion **6**(1), 5–20 (2005)

[63] Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. John Wiley & Sons, Hoboken, New Jersey (2006)

[64] Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 1107–1110 (2009). IEEE

[65] Wang, X., Girshick, R., Gupta, A., He, K.: Attention mechanisms in computer vision: A comprehensive survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

[66] Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.:

Nyströmformer: A nyström-based algorithm for approximating self-attention. Proceedings of the AAAI Conference on Artificial Intelligence **35**(16), 14138–14148 (2021)

[67] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. International Conference on Learning Representations (2019)

[68] Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. International Conference on Learning Representations (2021)

[69] Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems **35**, 16344–16359 (2022)

[70] Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A.: Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. Medical Imaging 2014: Digital Pathology **9041**, 904103 (2014)

[71] Yang, Y., Zha, S., Wang, J., Zhang, Z.: A survey on long-tailed visual recognition. International Journal of Computer Vision **130**(7), 1837–1872 (2022)

[72] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 4700–4708 (2017)

[73] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, 6105–6114 (2019). PMLR

[74] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11976–11986 (2022)

[75] Ding, J., Xue, N., Xia, G.-S., Dai, D., Yang, M.Y.: Hrfnet: High-resolution feature network for dense prediction. arXiv preprint arXiv:2108.07697 (2021)

[76] Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: Mmtm: Multimodal transfer module for cnn fusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13289–13299 (2020)

[77] Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, H.: Balanced meta-softmax for long-tailed visual recognition. Advances in Neural Information Processing Systems **33**, 4175–4186 (2020)