BUCAN: Bayesian Uncertainty-aware Classification with Attention Networks for Medical Images

Abhinav Sagar

Vrije Universiteit Brussel (VUB)
abhinav.sagar@vub.be

Abstract

Accurate and reliable medical image classification is critical for clinical decision-making across diverse imaging modalities, including X-ray, CT, and MRI. Traditional convolutional neural networks often produce overconfident predictions, limiting their clinical trustworthiness. In this work, we propose an uncertainty-aware, attention-augmented neural network that integrates multi-scale SwirlAttention and FeedBackAttention modules with a Bayesian probabilistic classifier. This framework enables robust feature extraction, interpretable attention maps, and principled estimation of epistemic uncertainty. We evaluate our approach on four diverse datasets, including Diabetic Retinopathy, Kvasir, Skin Cancer, and fused multi-focal Oocyte images, covering a wide range of pathological and morphological variations. Extensive experiments demonstrate that our method outperforms state-of-the-art CNN and transformer-based baselines in terms of accuracy, calibration, and interpretability. Grad-CAM visualizations highlight clinically relevant regions, while uncertainty estimates provide actionable insights for ambiguous cases, making the framework suitable for reliable deployment in real-world clinical settings.

Introduction

Medical image classification plays a critical role in modern clinical workflows by enabling automated diagnosis, early disease detection, and treatment planning across a wide range of imaging modalities. Despite significant advances with deep learning, conventional convolutional neural networks (CNNs) often operate under deterministic assumptions, producing overconfident predictions when confronted with noisy, ambiguous, or out-of-distribution medical images. This limitation is particularly concerning in safety-critical applications, where diagnostic errors can have severe clinical consequences.

Recent research has shown that incorporating attention mechanisms can improve both predictive accuracy and interpretability. Channel and spatial attention modules, such as SE (Squeeze-and-Excitation) (Hu, Shen, and Sun 2018)and CBAM (Convolutional Block Attention Module) (Woo et al. 2018), enable networks to adaptively emphasize informative feature maps and focus on the most relevant regions within an image. By jointly modeling inter-channel relationships and spatial dependencies, these mechanisms enhance the network's ability to discriminate between visually similar classes and provide more interpretable feature representations for clinical decision-making. However, existing attention-based methods still face several challenges. Many architectures tend to overfit on small medical datasets due to their high model complexity, while others struggle to generalize across imaging modalities with varying contrast, resolution, and noise characteristics. Furthermore, most attention mechanisms provide deterministic outputs without quantifying uncertainty, limiting their reliability in critical diagnostic scenarios where confidence estimation is essential.

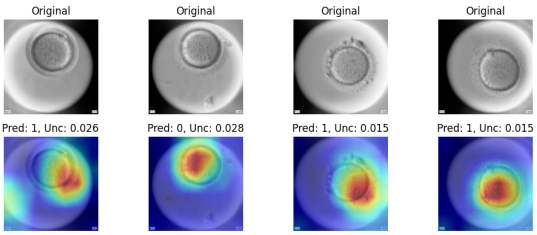

Figure 1: Illustration of GradCAM-based attention maps along with predicted labels and uncertainty quantification using the Oocytes dataset.

In this study, we propose a novel framework that integrates multi-scale attention with a Bayesian probabilistic classifier for robust medical image classification. Our approach is evaluated on four diverse datasets spanning multiple organs and imaging modalities: Diabetic Retinopathy fundus images, Kvasir endoscopic images, Skin Cancer dermoscopic images, and a fused multi-focal Oocyte dataset. Extensive experiments demonstrate that our method not only achieves state-of-the-art classification performance but also provides interpretable attention maps and meaningful uncertainty estimates, making it particularly suitable for clinical deployment.

To further explore the interpretability of our framework, we visualize the Grad-CAM-based attention maps alongside uncertainty quantification for oocyte images, as illustrated in Fig. 1. The highlighted regions in the attention maps represent discriminative features that significantly contribute to the model's decision-making process. For oocyte classifica-

tion, the model primarily focuses on morphological structures and texture patterns that indicate variations in oocyte quality.

The main contributions of this work are summarized as follows:

1. We introduce a novel attention-augmented residual network that integrates SwirlAttention and FeedBackAttention modules to capture multi-scale spatial and channel-wise dependencies in medical images.
2. We incorporate a Bayesian probabilistic classifier using Bayes-by-Backprop, enabling reliable epistemic uncertainty estimation alongside discriminative predictions.
3. We conduct comprehensive experiments on four diverse medical imaging datasets, demonstrating improved accuracy, calibration, and interpretability compared to state-of-the-art CNN and transformer-based baselines.
4. We provide qualitative insights using Grad-CAM visualizations, showing that the network consistently attends to clinically relevant regions and highlighting cases where predictions may be less reliable.

Overall, this work presents a unified framework for robust, interpretable, and uncertainty-aware medical image classification, addressing key limitations of conventional deterministic approaches while providing practical tools for real-world clinical applications.

Related Work

Medical image classification has witnessed rapid advancements with the adoption of deep learning, ranging from convolutional neural networks (CNNs) to vision transformers (ViTs) and hybrid architectures. Early works focused on CNNs for various modalities, demonstrating strong performance in tasks such as retinal fundus analysis, dermoscopic lesion classification, and brain tumor detection (Tajbakhsh et al. 2016; Swati et al. 2019). Residual and densely connected architectures further improved feature representation and gradient flow, with models such as ResGANet (Cheng et al. 2022) and DenseNet (Huang et al. 2017) achieving state-of-the-art results on multiple datasets.

Recent studies have emphasized attention mechanisms to enhance feature discriminability and interpretability. Lesion-aware networks (Fang et al. 2019), multi-scale pyramid fusion models (Wen et al. 2025), and hierarchical attentive fusion approaches (Abdar et al. 2022) have shown that guiding the network to focus on clinically relevant regions improves both accuracy and reliability. Concurrently, uncertainty-aware methods, including BARF (Abdar et al. 2021) and BayTTA (Sherkatghanad et al. 2025), have highlighted the importance of estimating epistemic and aleatoric uncertainty for robust clinical deployment.

Transformers and hybrid CNN-Transformer architectures have also gained traction in medical imaging due to their capacity for capturing long-range dependencies. Models such as MedViT (Manzari et al. 2023), MedMamba (Yue and Li 2024), TransMed (Dai, Gao, and Liu 2021), and MedTransNet (Shaik et al. 2024) leverage self-attention for global context modeling, while hybrid designs like EFFResNet-ViT (Hussain et al. 2025) and DBCvT (Li, Feng, and Xia 2024) combine convolutional feature extraction with transformer-based token modeling for improved performance on multi-modal datasets.

Several works have focused on model calibration and reliable uncertainty estimation to mitigate overconfident predictions, particularly in class-imbalanced scenarios (Rajaraman, Ganesan, and Antani 2022; Liang et al. 2020; Ju et al. 2022). Dynamic fusion networks such as DAFNet (Cai et al. 2025) and multi-task Mamba variants (Wu and Gou 2025) further enhance robustness by adaptively integrating multi-scale and multi-modal information.

Despite these advances, challenges remain in simultaneously achieving high classification accuracy, interpretable attention, and reliable uncertainty estimation across diverse medical imaging modalities. The proposed framework addresses these gaps by integrating multi-scale attention modules with a Bayesian probabilistic classifier, offering both strong discriminative capability and trustworthy uncertainty quantification. Our method is evaluated across four distinct datasets, including fundus, dermatology, endoscopy, and fused multi-focal microscopy images, demonstrating broad applicability and improved reliability compared to existing approaches.

Methodology

Problem Definition

Medical image classification aims to automatically categorize medical scans or images into clinically meaningful classes, such as disease presence, tissue type, or anatomical region, based on their visual and structural characteristics. Formally, let the dataset be defined as

$$\mathcal{D} = \{(\mathbf{I}_i, y_i)\}_{i=1}^{N},$$

where $\mathbf{I}_i \in R^{3\times H\times W}$ or $\mathbf{I}_i \in R^{1\times H\times W}$ represents a medical image (e.g., MRI, CT, or histopathology image) and $y_i \in \mathcal{Y} = \{1, 2, \ldots, C\}$ denotes its ground-truth diagnostic label among C possible disease categories. The goal of the model is to learn a mapping function.

$$f_\theta : \mathbf{I} \mapsto \hat{y} = \arg\max_{c\in\mathcal{Y}} p(y = c \mid \mathbf{I}, \theta),$$

parameterized by θ, that accurately predicts the correct class label $\hat{y}$ for each input image.

Traditional CNNs achieve high discriminative power but operate under deterministic weight assumptions, which often yield overconfident predictions in the presence of noisy or out-of-distribution medical data. This limitation is particularly critical in clinical settings where predictive confidence directly affects diagnostic reliability and downstream decision-making.

To mitigate this issue, we introduce a Bayesian uncertainty-aware framework that models the weight parameters w of the classification layer as probability distributions rather than fixed values. The predictive distribution for a new unseen image $\mathbf{I}^*$ is expressed as:

$$p(y^* \mid \mathbf{I}^*, \mathcal{D}) = \int p(y^* \mid \mathbf{I}^*, w)\, p(w \mid \mathcal{D})\, dw,$$

Where $p(w \mid \mathcal{D})$ represents the posterior over the weights given the training data. Since this integral is intractable, it is approximated through a variational posterior $q(w \mid \theta)$ using the Bayes-by-Backprop principle.

Thus, the learning objective is to infer both discriminative parameters and uncertainty estimates by minimizing the negative evidence lower bound (ELBO):

$$\mathcal{L} = E_{q(w|\theta)}[-\log p(y|\mathbf{I}, w)] + \beta\, D_{\mathrm{KL}}\big(q(w|\theta) \,\|\, p(w)\big),$$

Where the first term corresponds to the expected classification loss and the second term regularizes the approximate posterior toward the prior distribution.

In summary, the problem of medical image classification is reformulated as learning a robust, uncertainty-aware mapping.

$$f_\theta^* : \mathbf{I} \to \{p(y \mid \mathbf{I}),\, \mathrm{Var}(y \mid \mathbf{I})\},$$

That outputs both diagnostic probabilities and corresponding uncertainty estimates. This design ensures reliable and interpretable predictions, a crucial property for safety-critical medical decision-making.

Network Architecture

The proposed architecture integrates deterministic and probabilistic reasoning through a hierarchical attention-based residual backbone coupled with a Bayesian classification head. The overall design aims to achieve robust feature representation under uncertainty while maintaining efficient spatial–channel interactions.

1) Swirl Attention. To enhance spatial contextual reasoning, we introduce the **Swirl Attention** module, which performs multi-scale depthwise convolutions with kernel sizes $\{3, 5, 7, 9\}$. These convolutions emulate a "swirling" expansion of the receptive field, enabling the model to capture both fine and coarse spatial dependencies. The outputs from each scale are concatenated and fused through a 1×1 convolution followed by batch normalization:

$$\mathbf{y} = \mathrm{BN}\Big(\mathrm{Conv}_{1\times1}([\mathrm{Conv}_3, \mathrm{Conv}_5, \mathrm{Conv}_7, \mathrm{Conv}_9](\mathbf{x}))\Big) + \mathbf{x}.$$

This residual formulation ensures stable feature propagation while enriching contextual awareness.

2) FeedBack Attention. We propose a novel **FeedBack Attention** mechanism to jointly refine spatial and channel-wise activations. It consists of (i) a channel attention path based on global average pooling and two-layer bottleneck transformation, and (ii) a spatial attention path computed from the concatenation of average- and max-pooled channel responses. Unlike conventional dual-branch attention (e.g., CBAM (Woo et al. 2018)), the FeedBack Attention reuses the spatially modulated features to re-estimate the channel weights, thus forming a feedback loop that progressively refines salient feature activations. The joint attention is applied as:

$$\mathbf{y} = \mathbf{x} \odot \sigma_c(\mathbf{x}_{\mathrm{sp}}) \odot \sigma_s(\mathbf{x}_{\mathrm{ch}}),$$

Where σ_c and σ_s denote channel and spatial attention maps, respectively.

3) Residual Attention Blocks. Each residual unit integrates both the Swirl and FeedBack attention modules, providing adaptive feature recalibration and multi-scale context aggregation. When downsampling is required, a 1×1 convolutional skip connection is employed to match dimensions. The block formulation is given by:

$$\mathbf{y} = \mathrm{ReLU}\big(\mathrm{BN}_2(\mathrm{Conv}_2(\mathrm{BN}_1(\mathrm{Conv}_1(\mathbf{x})))) + \mathbf{x}_{\mathrm{skip}}\big),$$

Where intermediate activations are modulated by the two attention branches.

4) Bayesian Classification Head. To quantify epistemic uncertainty, the deterministic backbone is followed by a Bayesian inference layer using the Bayes-by-Backprop framework. The final fully connected layer is replaced by a **BayesLinear** module, where each weight and bias is modeled as a Gaussian distribution parameterized by (μ, ρ), with $\sigma = \log(1 + \exp(\rho))$ ensuring positivity. During training, weights are sampled as:

$$w = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

And optimized via the evidence lower bound (ELBO), incorporating a Kullback–Leibler divergence term against a standard normal prior. This probabilistic head enables calibrated uncertainty estimation at inference.

5) Overall Architecture. The complete model begins with a convolutional stem followed by four hierarchical Residual Attention Blocks with progressive channel expansion (32→64→128→256). A global average pooling layer and dropout regularization precede the Bayesian classifier. The integrated attention hierarchy and probabilistic inference enable robust and interpretable decision-making under data noise and domain variability.

The complete network architecture diagram is shown in Figure 2.

Loss Function

The proposed framework employs a Bayesian variational learning strategy to jointly optimize deterministic feature representations and probabilistic model parameters. Given the Bayes-by-Backprop formulation of the final classification layer, each weight and bias is represented by a variational distribution parameterized by its mean μ and log-scale parameter ρ. This allows the model to capture epistemic uncertainty through stochastic sampling during training.

Formally, let $q(w \mid \theta)$ denote the approximate posterior distribution over the network parameters w, where $\theta = \{\mu, \rho\}$, and let $p(w)$ represent a standard normal prior. The learning objective is derived from the maximization of the evidence lower bound (ELBO), which can be equivalently expressed as a minimization problem:

$$\mathcal{L}_{\text{total}} = E_{q(w|\theta)}[-\log p(y \mid \mathbf{I}, w)] + \beta\, D_{\mathrm{KL}}\big(q(w|\theta) \,\|\, p(w)\big),$$

where: - The first term, $E_{q(w|\theta)}[-\log p(y \mid \mathbf{I}, w)]$, corresponds to the expected negative log-likelihood of the correct class and serves as the **classification loss**. - The second term, $D_{\mathrm{KL}}(q(w|\theta) \,\|\, p(w))$, represents the **Kullback–Leibler (KL) divergence** between the approximate

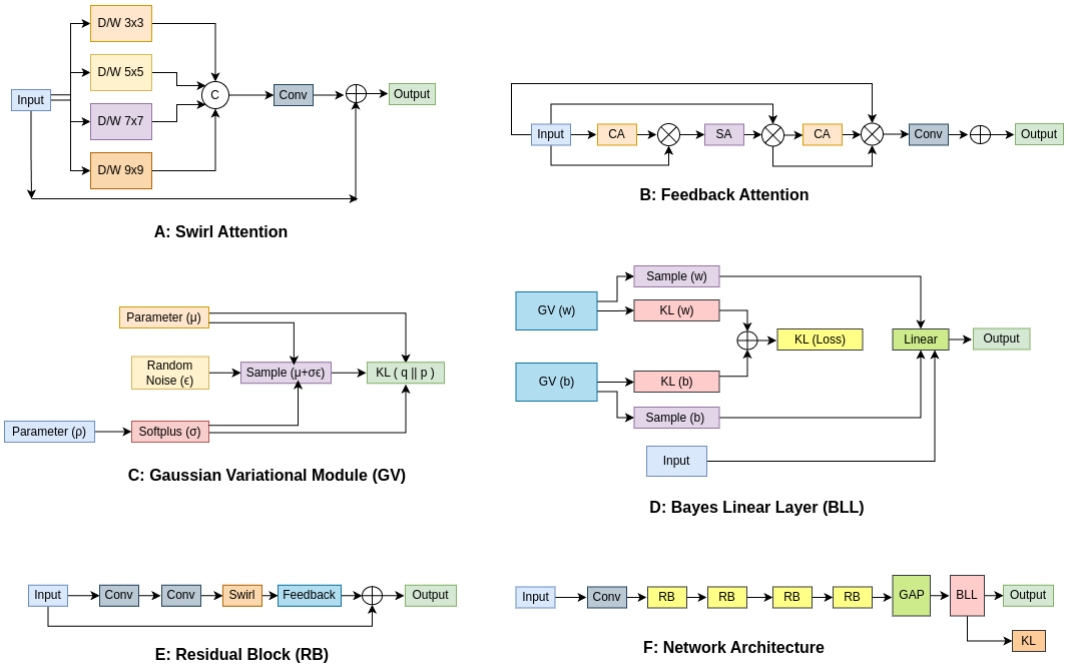

Figure 2: Illustration of the proposed network architecture and its components - A: Swirl Attention. B: Feedback Attention. C: Gaussian Variational Module. D: Bayes Linear Layer. E: Residual Block. F: Complete Network Architecture. Here, D/W, Conv, CA, SA, KL, Linear, and GAP denote Depthwise Convolution, Regular Convolution, Channel Attention, Spatial Attention, KL Divergence, Fully-Connected Layer, and Global-Average-Pooling Layer, respectively.

posterior and the prior, acting as a regularizer that penalizes overconfident posterior distributions. - The scalar β balances predictive accuracy and uncertainty regularization; empirically, $\beta \in [10^{-4}, 10^{-2}]$ yields stable convergence.

For medical image classification, the likelihood term is modeled using the categorical cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log \hat{p}(y = c \mid \mathbf{I}_i, w),$$

where $y_{i,c}$ denotes the one-hot encoded ground-truth label and $\hat{p}(y = c \mid \mathbf{I}_i, w)$ is the predicted class probability. The final loss function can thus be written as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \beta\, \mathcal{L}_{\text{KL}},$$

with

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\big(q(w|\theta) \,\|\, p(w)\big).$$

During training, the reparameterization trick

$$w = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

It is used to enable gradient-based optimization through stochastic sampling. The model parameters θ are updated via backpropagation to minimize $\mathcal{L}_{\text{total}}$.

This combined objective ensures that the model simultaneously learns discriminative features and captures epistemic uncertainty. As a result, the proposed loss formulation encourages reliable predictions with calibrated confidence—an essential property for robust decision-making in medical image analysis.

Experiments

Dataset

To comprehensively evaluate the effectiveness and generalizability of the proposed medical image classification framework, four publicly available and custom-prepared datasets were utilized. These datasets span multiple imaging modalities, organs, and diagnostic tasks, ensuring robust validation across diverse clinical applications.

1. **Diabetic Retinopathy (DR):** The Diabetic Retinopathy dataset contains high-resolution retinal fundus images annotated with five severity levels: *No DR*, *Mild*, *Moderate*, *Severe*, and *Proliferative DR*. Specifically, the dataset includes 1,805 images without DR, 370 mild cases, 999 moderate cases, 193 severe cases, and 295 proliferative DR cases (Asia Pacific Tele-Ophthalmology Society (APTOS) 2019).
2. **Kvasir:** The Kvasir dataset consists of endoscopic gastrointestinal (GI) tract images annotated for both anatomical landmarks and pathological findings. It contains 1,000 images per category across eight classes: dyed-lifted polyps, dyed-resection margins, esophagitis, normal cecum, normal pylorus, normal z-line, polyps, and ulcerative colitis (Pogorelov et al. 2017).
3. **Skin Cancer (HAM10000):** The HAM10000 dataset comprises dermoscopic images of skin lesions spanning nine diagnostic categories, covering both benign and malignant conditions. Specifically, it includes 130 images of actinic keratosis, 392 basal cell carcinoma, 111 der-

matofibroma, 454 melanoma, 373 nevus, 478 pigmented benign keratosis, 80 seborrheic keratosis, 197 squamous cell carcinoma, and 142 vascular lesions (Codella et al. 2019).

4. **Oocyte Multi-Focal Fusion Dataset:** A specialized private oocyte imaging dataset was constructed by fusing **11 focal plane images** into a single high-information composite using an adaptive fusion strategy. The dataset is categorized into four biologically meaningful classes representing distinct oocyte maturation stages.

Implementation Details

Dataset and Preprocessing. All images were resized to 224×224 pixels and normalized using standard ImageNet statistics. For cross-validation experiments, a 5-fold stratified splitting strategy was employed to preserve class distributions across folds.

Training. Models were optimized using the Adam optimizer with an initial learning rate of 1×10^{-4}. Training was conducted for 100 epochs per fold in the cross-validation setting using a batch size of 4. During inference, Monte Carlo Dropout with 30 stochastic forward passes was utilized to estimate epistemic uncertainty.

Software and Hardware. The framework was implemented in Python 3.10 using PyTorch 2.1, torchvision, and scikit-learn. All training and evaluation were performed on a workstation equipped with an NVIDIA A100 GPU (CUDA 11.8), 32 GB RAM, and an Intel Xeon CPU.

Evaluation Metrics. Model performance was assessed using accuracy, precision, recall, F1-score, Matthews correlation coefficient (MCC), and ROC-AUC. Calibration quality was evaluated using the Brier score, Expected Calibration Error (ECE), and reliability diagrams. Confusion matrices, ROC curves, and precision-recall curves were generated for qualitative assessment. Additionally, Grad-CAM was applied to visualize attention maps and interpret model decision-making.

Comparison Approaches

To evaluate the effectiveness of the proposed uncertainty-aware, attention-augmented framework, we benchmark it against several state-of-the-art models for medical image classification. The selected baselines encompass a broad range of architectures, including conventional convolutional neural networks (CNNs), lightweight mobile networks, and modern transformer-based approaches, providing a comprehensive performance comparison.

Quantitative Performance Quantitative evaluation was conducted on four diverse medical imaging datasets: Diabetic Retinopathy (Asia Pacific Tele-Ophthalmology Society (APTOS) 2019), Kvasir (Pogorelov et al. 2017), Skin Cancer (HAM10000) (Codella et al. 2019), and the private Oocyte Multi-Focal Fusion dataset. The comparison approaches include the following representative models:

- **CNN-based models:** DenseNet121 (Huang et al. 2017), ResNet50 (He et al. 2016), ResNeXt101 (Xie et al. 2017), RepVGG (Ding et al. 2021), ShuffleNetV2 (Ma et al. 2018), MobileNetV3 (Howard et al. 2019), EfficientNetB0 (Tan and Le 2019), and InceptionNext-Tiny (Yu et al. 2024).
- **Transformer-based and hybrid models:** SimpleViT (Beyer, Zhai, and Kolesnikov 2022), RegionViT (Chen, Panda, and Fan 2021), VAN (Guo et al. 2023), EfficientFormerL1 (Li et al. 2022), MobileViTXXS (Mehta and Rastegari 2021), MLPMixer (Tolstikhin et al. 2021), and ConvMixer (Trockman and Kolter 2022).

All models were trained and evaluated under the same preprocessing, data augmentation, and cross-validation protocols, ensuring a fair comparison.

Qualitative Performance Figure 3 offer qualitative insights into the model's performance. Figure 3 A depicts the confusion matrix using the Diabetic-Retinopathy dataset, showing that the classifier achieves high true positive rates across most classes with only a few misclassifications. Figure 3 B illustrates the ROC–AUC curves using the skin cancer dataset, indicating strong discriminative power between classes, as most curves approach the ideal top-left region. Th right diagram in Figure 3 C presents the precision–recall curves using the Kvasir dataset, demonstrating that the model sustains high precision even at elevated recall levels.

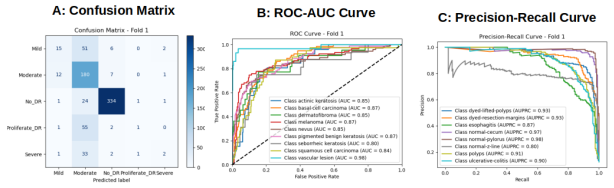

Figure 3: From left to right - A: Confusion matrix using the diabetic retinopathy dataset, B: ROC-AUC curve using the skin cancer dataset, and C: Precision-Recall curve using the Kvasir dataset

Attention Visualization

To provide insights into the decision-making process of the proposed model, Grad-CAM (Selvaraju et al. 2017) was employed to generate visual explanations of class-specific activations. For each input image, heatmaps were overlaid on the original images to highlight regions that contributed most strongly to the predicted class.

The Grad-CAM visualizations reveal that the model consistently attends to clinically relevant regions. For example, in retinal images, the model highlights microaneurysms, hemorrhages, and exudates corresponding to diabetic retinopathy severity as in Figure 4. Similarly, in dermoscopic images, attention is concentrated on lesion boundaries and abnormal tissue structures as in Figure 5.

These qualitative results demonstrate that the attention mechanisms integrated into the network, coupled with Grad-CAM analysis, provide interpretable and clinically meaningful explanations for predictions. They also help identify cases where the model may rely on spurious features, guid-

A: Diabetic-Retinopathy Dataset												
Method	AUC ↑	Accuracy ↑	Precision ↑	Recall↑	F1 Score ↑	MCC ↑	Brier ↓	ECE ↓	Params ↓	Flops ↓	Time ↓	FPS ↓
ConvMixer (Trockman and Kolter 2022)	0.8541	0.7428	0.5591	0.5220	0.5262	0.6090	0.1675	0.1901	3.7596	3.8609	0.0953	10.4885
DenseNet121 (Huang et al. 2017)	**0.9143**	**0.7829**	0.6273	**0.6040**	**0.6085**	**0.6713**	**0.1420**	**0.1578**	6.9590	2.8960	0.0199	50.1820
EfficientFormerL1 (Li et al. 2022)	0.8555	0.7206	0.5184	0.4929	0.4996	0.5725	0.1752	0.2342	11.3891	1.3095	0.0414	24.1817
EfficientNetB0 (Tan and Le 2019)	0.8682	0.7346	0.5432	0.5008	0.5099	0.5920	0.1702	0.2029	4.0140	0.4139	0.0107	93.5750
InceptionNextTiny (Yu et al. 2024)	0.8394	0.6996	0.5428	0.4385	0.4348	0.5445	0.2091	0.2594	25.7556	4.1982	0.0095	105.2280
Med-Mamba-Tiny (Yue and Li 2024)	0.8489	0.7324	0.5280	0.4935	0.4984	0.5899	0.1822	0.2457	13.2614	2.0244	0.0172	58.2478
Med-ViT-Small (Manzari et al. 2023)	0.8749	0.7545	0.5898	0.5205	0.5339	0.6223	0.1528	0.1722	31.1416	5.9286	0.0323	30.9432
Med-ViT-V2-Tiny (Manzari et al. 2025)	0.8158	0.7059	0.5184	0.4845	0.4880	0.5517	0.1722	0.2490	2.7448	1.4942	0.1897	5.2702
MLPMixer (Tolstikhin et al. 2021)	0.8508	0.7130	0.5072	0.4825	0.4875	0.5604	0.1906	0.2858	5.4226	1.3170	**0.0043**	**230.3711**
MobileViTXXS (Mehta and Rastegari 2021)	0.8536	0.7100	0.5100	0.5021	0.5023	0.5629	0.2005	0.3152	1.0131	0.2731	0.0101	98.7030
MobileNetV3 (Howard et al. 2019)	0.8382	0.7086	0.5035	0.4733	0.4789	0.5558	0.1678	0.2474	1.5230	**0.0615**	0.0091	109.6083
RegionViT (Chen, Panda, and Fan 2021)	0.8416	0.6958	0.4819	0.4642	0.4649	0.5373	0.1825	0.2460	12.2360	2.2633	0.0600	16.6629
RepVGG (Ding et al. 2021)	0.8819	0.7616	0.5990	0.5467	0.5604	0.6356	0.1689	0.2049	10.2591	2.8682	0.0418	23.9452
Resnet50 (He et al. 2016)	0.8646	0.7354	0.5504	0.5243	0.5287	0.5975	0.1693	0.2079	23.5183	4.1317	0.0106	94.7223
ResNeXt101 (Xie et al. 2017)	0.8460	0.7127	0.5080	0.4857	0.4874	0.5616	0.1784	0.2564	42.1389	8.0529	0.1494	6.6929
ShuffleNetV2 (Ma et al. 2018)	0.8871	0.7589	0.5852	0.5520	0.5596	0.6324	0.1621	0.1958	0.9217	0.1473	0.0086	116.7053
SimpleViT (Beyer, Zhai, and Kolesnikov 2022)	0.8308	0.7081	0.4920	0.4696	0.4709	0.5540	0.1666	0.2179	**0.3808**	0.0768	0.0056	177.2153
VAN (Guo et al. 2023)	0.8580	0.7286	0.5293	0.4891	0.4951	0.5817	0.1905	0.2400	3.8469	0.8706	0.0120	83.2445
Ours	0.9029	0.7810	**0.6275**	0.5842	0.5902	0.6669	0.1636	0.2040	3.3463	4.4537	0.0095	105.7026
B: Kvasir Dataset												
Method	AUC ↑	Accuracy ↑	Precision ↑	Recall↑	F1 Score ↑	MCC ↑	Brier ↓	ECE ↓	Params ↓	Flops ↓	Time ↓	FPS ↓
ConvMixer (Trockman and Kolter 2022)	0.9748	0.8180	0.8197	0.8180	0.8167	0.7926	0.0348	0.1280	3.7612	3.8609	0.0949	10.5350
DenseNet121 (Huang et al. 2017)	**0.9877**	0.8714	0.8736	0.8714	0.8707	0.8535	**0.0198**	**0.0899**	6.9621	2.8960	0.0201	49.7727
EfficientFormerL1 (Li et al. 2022)	0.9667	0.7416	0.7452	0.7416	0.7379	0.7062	0.0736	0.1748	11.3918	1.3095	0.0475	21.0739
EfficientNetB0 (Tan and Le 2019)	0.9796	0.8316	0.8349	0.8316	0.8300	0.8084	0.0506	0.1816	4.0178	0.4139	0.0105	94.7937
InceptionNextTiny (Yu et al. 2024)	0.9573	0.7070	0.7103	0.7070	0.7019	0.6670	0.0865	0.2529	25.7625	4.1982	0.0098	101.6899
Med-Mamba-Tiny (Yue and Li 2024)	0.9748	0.7879	0.7885	0.7879	0.7867	0.7580	0.0378	0.0994	13.2638	2.0244	0.0209	47.8656
Med-ViT-Small (Manzari et al. 2023)	0.9697	0.7686	0.7779	0.7686	0.7623	0.7380	0.0307	0.1638	31.1446	5.9286	0.0324	30.8993
Med-ViT-V2-Tiny (Manzari et al. 2025)	0.9679	0.7652	0.7694	0.7652	0.7626	0.7330	0.0384	0.1483	2.7459	1.4942	0.2102	4.7580
MLPMixer (Tolstikhin et al. 2021)	0.9603	0.7344	0.7333	0.7344	0.7323	0.6969	0.0466	0.1222	5.4242	1.3170	**0.0036**	**279.0365**
MobileViTXXS (Mehta and Rastegari 2021)	0.9779	0.8231	0.8255	0.8231	0.8217	0.7986	0.0686	0.1974	1.0140	0.2731	0.0118	84.8797
MobileNetV3 (Howard et al. 2019)	0.9799	0.8096	0.8188	0.8096	0.8071	0.7844	0.0447	0.0990	1.5261	**0.0615**	0.0091	109.6956
RegionViT (Chen, Panda, and Fan 2021)	0.9645	0.7209	0.7245	0.7209	0.7183	0.6822	0.0805	0.2275	12.2376	2.2633	0.0528	18.9278
RepVGG (Ding et al. 2021)	0.9842	0.8479	0.8497	0.8479	0.8469	0.8267	0.0283	0.1169	10.2606	2.8682	0.0431	23.2149
Resnet50 (He et al. 2016)	0.9808	0.8206	0.8226	0.8206	0.8192	0.7957	0.0530	0.1582	23.5244	4.1317	0.0109	91.4189
ResNeXt101 (Xie et al. 2017)	0.9696	0.7925	0.7936	0.7925	0.7896	0.7637	0.0395	0.0916	42.1451	8.0529	0.1419	7.0477
ShuffleNetV2 (Ma et al. 2018)	0.9864	0.8654	0.8659	0.8654	0.8650	0.8463	0.0413	0.1520	0.9246	0.1473	0.0084	118.7920
SimpleViT (Beyer, Zhai, and Kolesnikov 2022)	0.9657	0.7426	0.7441	0.7426	0.7416	0.7064	0.0651	0.2000	**0.3810**	0.0768	0.0039	253.2973
VAN (Guo et al. 2023)	0.9671	0.7564	0.7563	0.7564	0.7548	0.7220	0.0605	0.1695	3.8477	0.8706	0.0113	88.2147
Ours	0.9873	**0.8746**	**0.8776**	**0.8746**	**0.8743**	**0.8572**	0.0241	0.0965	3.3463	4.4537	0.0094	106.2179
C: Oocytes Dataset												
Method	AUC ↑	Accuracy ↑	Precision ↑	Recall↑	F1 Score ↑	MCC ↑	Brier ↓	ECE ↓	Params ↓	Flops ↓	Time ↓	FPS ↓
ConvMixer (Trockman and Kolter 2022)	0.6227	0.4055	**0.3812**	0.3763	**0.3569**	0.1926	0.2683	0.4227	3.7089	3.8095	0.0115	86.6638
DenseNet121 (Huang et al. 2017)	**0.6386**	0.4007	0.3003	0.3476	0.2947	0.1779	0.3046	0.4954	6.9517	2.8174	0.0205	48.8486
EfficientFormerL1 (Li et al. 2022)	0.5689	0.3385	0.3030	0.3105	0.3026	0.0935	0.4236	0.5691	11.3878	1.3041	0.0067	148.3002
EfficientNetB0 (Tan and Le 2019)	0.5811	0.3369	0.3039	0.3089	0.3011	0.0921	0.2884	0.5488	4.0121	0.4066	0.0101	98.8973
InceptionNextTiny (Yu et al. 2024)	0.5438	0.3001	0.2650	0.2634	0.2079	0.0191	0.3722	0.5946	25.7502	4.1886	0.0096	103.7895
Med-Mamba-Tiny (Yue and Li 2024)	0.5460	0.3352	0.2975	0.3040	0.2994	0.0905	0.4034	0.6758	13.2576	2.0148	0.0198	50.6119
Med-ViT-Small (Manzari et al. 2023)	0.5643	0.3344	0.2951	0.3057	0.2970	0.0870	0.2815	0.4005	31.1394	5.9142	0.0305	32.7587
Med-ViT-V2-Tiny (Manzari et al. 2025)	0.5767	0.3573	0.3240	0.3285	0.3059	0.1255	**0.2647**	**0.3864**	2.7433	1.4797	0.1838	5.4407
MLPMixer (Tolstikhin et al. 2021)	0.5189	0.2837	0.2702	0.2693	0.2684	0.0243	0.2995	0.3870	5.1600	1.2656	0.0039	255.4177
MobileViTXXS (Mehta and Rastegari 2021)	0.5470	0.3140	0.2822	0.2894	0.2821	0.0647	0.3483	0.5783	1.0125	0.2695	0.0102	97.9007
MobileNetV3 (Howard et al. 2019)	0.5795	0.3327	0.3165	0.3145	0.3085	0.0979	0.2964	0.4583	1.5217	**0.0578**	0.0082	121.9475
RegionViT (Chen, Panda, and Fan 2021)	0.5274	0.3115	0.2854	0.2725	0.2130	0.0408	0.3692	0.5164	12.1270	2.2312	0.0202	49.4871
RepVGG (Ding et al. 2021)	0.5992	0.3581	0.3312	0.3294	0.3235	0.1184	0.2685	0.4467	10.2573	2.8521	0.0050	198.1665
Resnet50 (He et al. 2016)	0.6278	0.4162	0.3403	0.3674	0.3107	0.2015	0.3524	0.4984	23.5100	4.0530	0.0107	93.8753
ResNeXt101 (Xie et al. 2017)	0.6220	0.4031	0.3270	0.3598	0.3138	0.1836	0.2795	0.3966	42.1306	7.9742	0.0239	41.9054
ShuffleNetV2 (Ma et al. 2018)	0.5948	0.3467	0.3170	0.3222	0.3109	0.1126	0.3682	0.5683	0.9203	0.1419	0.0077	130.3309
SimpleViT (Beyer, Zhai, and Kolesnikov 2022)	0.5343	0.2984	0.2686	0.2724	0.2690	0.0390	0.3819	0.6419	**0.3469**	0.0700	**0.0029**	**349.4394**
VAN (Guo et al. 2023)	0.5684	0.3344	0.2985	0.3067	0.2980	0.0872	0.3122	0.5931	3.8435	0.8608	0.0106	94.1333
Ours	0.6267	**0.4350**	0.3605	**0.3805**	0.3237	**0.2343**	0.2653	0.4042	3.3457	4.4248	0.0095	105.1254
D: Skin-Cancer Dataset												
Method	AUC ↑	Accuracy ↑	Precision ↑	Recall↑	F1 Score ↑	MCC ↑	Brier ↓	ECE ↓	Params ↓	Flops ↓	Time ↓	FPS ↓
ConvMixer (Trockman and Kolter 2022)	0.8819	0.5643	0.4881	0.4954	0.4843	0.4894	**0.0723**	0.2491	3.7617	3.8609	0.0849	11.7729
DenseNet121 (Huang et al. 2017)	0.8531	0.5384	0.4740	0.4798	0.4628	0.4636	0.0924	0.2489	6.9631	2.8960	0.0199	50.2135
EfficientFormerL1 (Li et al. 2022)	0.8435	0.4370	0.3635	0.3720	0.3597	0.3411	0.1444	0.3865	11.3927	1.3095	0.0334	29.8995
EfficientNetB0 (Tan and Le 2019)	0.8387	0.5082	0.4365	0.4285	0.4268	0.4221	0.1341	0.3334	4.0191	0.4139	0.0113	88.8317
InceptionNextTiny (Yu et al. 2024)	0.8450	0.5028	0.4407	0.3805	0.3639	0.4152	0.1585	0.4376	25.7648	4.1983	0.0091	109.8458
Med-Mamba-Tiny (Yue and Li 2024)	0.8425	0.5044	0.4277	0.4198	0.4153	0.4169	0.1548	0.3709	13.2645	2.0244	0.0199	50.1691
Med-ViT-Small (Manzari et al. 2023)	0.8540	0.5028	0.4372	0.4269	0.4254	0.4155	0.0999	0.1869	31.1457	5.9286	0.0306	32.6488
Med-ViT-V2-Tiny (Manzari et al. 2025)	0.8444	0.4816	0.4141	0.4207	0.4096	0.3936	0.1205	0.4157	2.7463	1.4942	0.1920	5.2088
MLPMixer (Tolstikhin et al. 2021)	0.8540	0.5168	0.4284	0.4207	0.4209	0.4304	0.0958	0.3269	5.4247	1.3170	**0.0039**	**254.3940**
MobileViTXXS (Mehta and Rastegari 2021)	0.8220	0.4752	0.3993	0.3893	0.3869	0.3822	0.1245	0.4363	1.0144	0.2731	0.0108	92.5576
MobileNetV3 (Howard et al. 2019)	0.8173	0.4913	0.4024	0.3933	0.3901	0.4000	0.0952	0.3168	1.5271	**0.0615**	0.0090	110.6497
RegionViT (Chen, Panda, and Fan 2021)	0.8303	0.4828	0.3964	0.3989	0.3963	0.3921	0.1406	0.4275	12.2381	2.2633	0.0528	18.9483
RepVGG (Ding et al. 2021)	0.8897	0.5613	0.4836	0.4820	0.4760	0.4872	0.1025	0.3311	10.2611	2.8682	0.0315	31.7892
Resnet50 (He et al. 2016)	0.8263	0.4909	0.4128	0.4148	0.4041	0.4049	0.1090	0.3215	23.5265	4.1317	0.0099	100.7933
ResNeXt101 (Xie et al. 2017)	0.8428	0.4849	0.4387	0.4197	0.4185	0.3985	0.1017	0.3588	42.1471	8.0529	0.1411	7.0866
ShuffleNetV2 (Ma et al. 2018)	0.8719	0.5592	0.4872	0.4807	0.4717	0.4841	0.0995	0.3121	0.9256	0.1473	0.0070	142.5641
SimpleViT (Beyer, Zhai, and Kolesnikov 2022)	0.8348	0.4662	0.4109	0.3785	0.3819	0.3713	0.1484	0.4513	**0.3811**	0.0768	0.0055	180.6023
VAN (Guo et al. 2023)	0.8667	0.5189	0.4393	0.4415	0.4338	0.4337	0.1316	0.3641	3.8479	0.8706	0.0106	94.7108
Ours	**0.9089**	**0.6020**	**0.5326**	**0.5419**	**0.5254**	**0.5368**	0.0818	**0.0971**	3.3463	4.4537	0.0099	100.8756

Table 1: A: Quantitative comparison on the Diabetic Retinopathy dataset. B: Quantitative comparison on the Kvasir dataset. C: Quantitative comparison on the Oocytes dataset. D: Quantitative comparison on the Skin Cancer dataset. The best values are highlighted in bold, while in cases where our model performs the second best, we highlight with the color red.

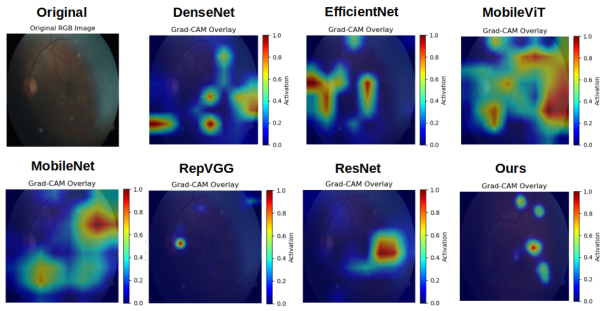

Figure 4: Grad-CAM visual comparisons between our proposed model and other state-of-the-art image classification approaches on the Diabetic Retinopathy dataset.

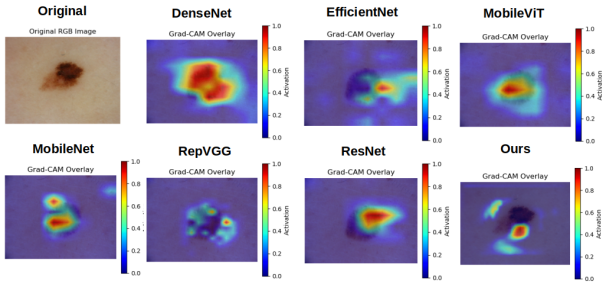

Figure 5: Grad-CAM visual comparisons between our proposed model and other state-of-the-art image classification approaches on the Skin Cancer dataset.

ing future refinement of both architecture and training strategies for enhanced reliability.

Reliability Diagram:

Figure 6 illustrates the reliability diagrams for multiple datasets, including Diabetic-Retinopathy, Kvasir, Oocytes, and Skin-Cancer. Each plot compares the model's predicted confidence (X-axis) with the corresponding empirical accuracy (Y-axis), providing insights into calibration quality. Ideally, a well-calibrated model's reliability curve aligns closely with the diagonal reference line, indicating that predicted probabilities accurately reflect true likelihoods. The diagrams show that, across most datasets, the model's predictions exhibit reasonable calibration, though slight deviations from the diagonal suggest instances of overconfidence or underconfidence in certain confidence intervals.

Top Uncertain and Certain Samples:

To gain deeper insights into the reliability of the proposed model, we further examine its predictive uncertainty on the Diabetic-Retinopathy dataset. Figure 7 presents the five most uncertain and the five most certain cases as estimated by our framework. The most certain cases exhibit high-confidence predictions consistent with the ground-truth labels, underscoring the model's robustness in recognizing clear diagnostic patterns. In contrast, the most uncertain cases involve challenging samples—such as those with ambiguous lesion boundaries or visually similar inter-class features—where the model demonstrates lower confidence. This analysis highlights not only the value of incorporating

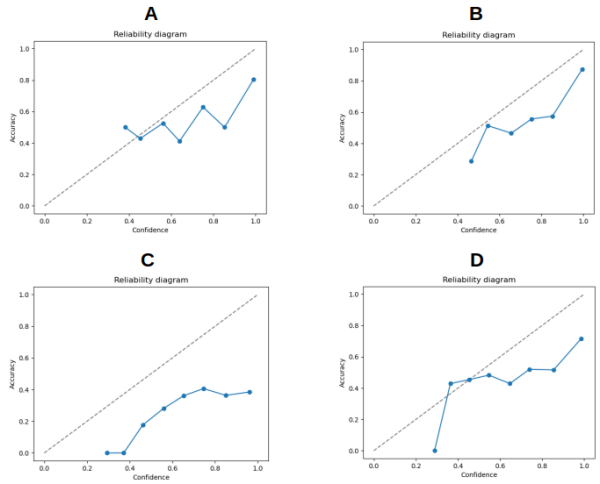

Figure 6: Reliability Diagram using datasets - A: Diabetic-Retinopathy, B: Kvasir, C: Oocytes, and D: Skin-Cancer. The Y-axis represents accuracy, and the X-axis represents confidence.

uncertainty estimation in medical image classification but also its potential as a diagnostic tool for flagging cases that may warrant additional expert assessment.

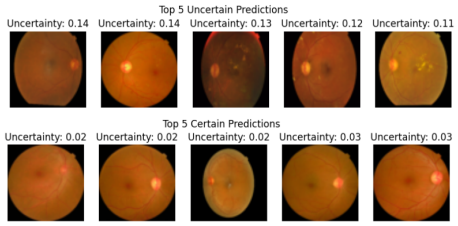

Figure 7: Top row shows the 5 most uncertain samples, while the bottom row shows the 5 most certain samples using the diabetic-retinopathy dataset.

Ablation Study

Table 2 presents an extensive ablation study conducted on two medical datasets—Diabetic-Retinopathy (A) and Skin-Cancer (B)—to evaluate the impact of different architectural components on model performance. Across both datasets, the proposed model ("Ours") consistently achieves superior or highly competitive results in nearly all key metrics, including AUC, Accuracy, F1 Score, and MCC, while also demonstrating the lowest computational overhead in terms of inference time and the highest FPS. These results highlight the synergistic benefit of integrating Swirl Attention, Feedback Attention, and Bayesian fully connected (BayesFC) mechanisms, yielding improved predictive reliability, computational efficiency, and uncertainty calibration across diverse medical imaging tasks.

Table 3 presents the ablation analysis comparing the proposed loss function against the conventional Cross-Entropy (CE) loss on the Diabetic-Retinopathy and Skin-Cancer

A: Diabetes Dataset												
Method	AUC ↑	Accuracy ↑	Precision ↑	Recall↑	F1 Score ↑	MCC ↑	Brier ↓	ECE ↓	Params ↓	Flops ↓	Time ↓	FPS ↓
Swirl → Conv	0.8903	0.7564	0.5890	0.5741	0.5685	0.6337	0.1795	0.1854	3.3463	4.4537	0.0286	35.0081
Feedback → Conv	0.8934	0.7736	0.6166	0.5652	0.5792	0.6553	0.1738	**0.1718**	3.2692	4.4486	0.0242	41.3499
Swirl → Conv, Feedback → Conv	0.8873	0.7534	0.5875	0.5591	0.5558	0.6317	0.2263	0.2185	2.6949	3.5623	0.0117	85.2971
Swirl → SA	0.8973	0.7735	0.6189	0.5793	0.5879	0.6613	0.2496	0.4912	2.4741	3.1674	0.0132	75.9204
Feedback → CA	0.8752	0.7299	0.5940	0.5004	0.4901	0.5957	0.2846	0.5081	**2.3978**	3.1674	0.0134	74.5348
Swirl → SA, Feedback → CA	0.8829	0.7591	0.5974	0.5225	0.5275	0.6347	0.2247	0.3228	2.4675	**3.1546**	0.0138	72.3772
BayesFC→FC	**0.9058**	0.7772	**0.6373**	0.5786	0.5842	0.6624	0.4114	0.5627	3.3476	4.4537	0.0283	35.2949
Ours	0.9029	**0.7810**	0.6275	**0.5842**	**0.5902**	**0.6669**	**0.1636**	0.2040	3.3463	4.4537	**0.0095**	**105.7026**
B: Skin-Cancer Dataset												
Method	AUC ↑	Accuracy ↑	Precision ↑	Recall↑	F1 Score ↑	MCC ↑	Brier ↓	ECE ↓	Params ↓	Flops ↓	Time ↓	FPS ↓
Swirl → Conv	0.8691	0.5482	0.4685	0.4545	0.4450	0.4668	0.0852	**0.0747**	2.7720	3.5674	0.0173	57.6629
Feedback → Conv	0.8955	0.5910	**0.5694**	0.5138	0.5125	0.5238	0.2063	0.6050	3.2692	4.4486	0.0244	40.9963
Swirl → Conv, Feedback → Conv	0.8555	0.5414	0.4525	0.4502	0.4381	0.4602	0.1468	0.2235	2.6949	3.5623	0.0117	85.3893
Swirl → SA	0.9002	0.6024	0.5274	0.5403	0.5179	0.5384	0.1457	0.3297	2.4741	3.1674	0.0131	76.5227
Feedback → CA	0.8757	0.5609	0.4752	0.4601	0.4510	0.4827	0.0996	0.1059	**2.3978**	3.1674	0.0136	73.5819
Swirl → SA, Feedback → CA	0.8856	0.5736	0.5237	0.4841	0.4791	0.4985	0.1802	0.4520	2.4675	**3.1546**	0.0135	73.8911
BayesFC→FC	0.9050	**0.6037**	0.5374	0.5410	0.5227	**0.5385**	0.1613	0.5337	3.3486	4.4537	0.0287	34.8153
Ours	**0.9089**	0.6020	0.5326	**0.5419**	**0.5254**	0.5368	**0.0818**	0.0971	3.3463	4.4537	**0.0099**	**100.8756**

Table 2: Ablation study using different components in the network architecture. A: Diabetic-Retinopathy dataset. B: Skin-Cancer dataset. Here, Conv, SA, CA, and FC denotes Convolutional layer, Spatial Attention, Channel Attention, and Fully-Connected layer, respectively. The best values are highlighted in bold.

datasets. The results clearly demonstrate the effectiveness of the proposed loss formulation in improving both predictive accuracy and calibration quality. These consistent improvements across datasets highlight the robustness and calibration benefits of the proposed loss, effectively balancing classification accuracy and uncertainty awareness in medical diagnosis tasks.

A: Diabetes Dataset								
Method	Acc ↑	Pre ↑	Rec ↑	F1 ↑	AUC ↑	MCC ↑	Brier ↓	ECE ↓
CE	0.8962	0.7796	**0.6484**	0.5811	0.5852	**0.6694**	0.3686	0.6111
Ours	**0.9029**	**0.7810**	0.6275	**0.5842**	**0.5902**	0.6669	**0.1636**	**0.2040**
B: Skin-Cancer Dataset								
Method	Acc ↑	Pre ↑	Rec ↑	F1 ↑	AUC ↑	MCC ↑	Brier ↓	ECE ↓
CE	0.8975	0.5948	**0.5563**	0.5239	0.5117	0.5285	0.1236	0.1749
Ours	**0.9089**	**0.6020**	0.5326	**0.5419**	**0.5254**	**0.5368**	**0.0818**	**0.0971**

Table 3: Ablation study using loss function. A: Diabetic-Retinopathy dataset. B: Skin-Cancer dataset. Here CE denotes the Cross-Entropy loss. The best values are highlighted in bold.

Failure Cases

Despite the overall strong performance of the proposed uncertainty-aware, attention-augmented framework, certain failure cases were observed across the evaluated datasets. In the Diabetic Retinopathy and Skin-Cancer datasets, intermediate stages or borderline morphological variations were occasionally misclassified due to overlapping feature representations between adjacent classes (e.g., Mild vs. Moderate DR). In dermoscopic images from the Skin Cancer dataset, lesions with atypical pigmentation or small size were occasionally misclassified, highlighting the challenge of rare or underrepresented patterns.

We observe instances of both False Positives and False Negatives in Figure 8. False Positives occur when the model's predictions are correct, but Grad-CAM reveals that the model is focusing on irrelevant or inaccurate regions of the image. False Negatives, on the other hand, occur when the model's predictions are incorrect, even though Grad-CAM highlights the correct or semantically relevant regions of the image.

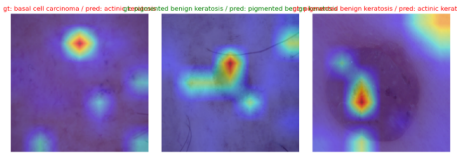

Figure 8: Illustration of some failure cases using GradCAM-based attention maps along with true and predicted labels using the Skin-Cancer dataset.

These failure cases underscore the importance of high-quality imaging, sufficient class representation, and careful preprocessing. They also indicate potential avenues for future work, including multi-modal data integration, improved attention-guided feature extraction, and uncertainty-driven active learning to better handle ambiguous or visually challenging samples.

Conclusions

In this work, we propose an uncertainty-aware, attention-augmented neural network for medical image classification that combines multi-scale SwirlAttention and FeedBack-Attention modules with a Bayes-by-Backprop probabilistic classifier. The framework enables both discriminative feature learning and reliable uncertainty estimation, addressing the overconfidence of conventional deterministic models. Extensive experiments on four diverse datasets—including Diabetic Retinopathy, Kvasir, Skin Cancer, and fused multi-focal Oocyte images—demonstrate improved predictive accuracy, calibration, and interpretability compared to state-of-the-art CNN and transformer-based models. The results highlight the potential of integrating attention mechanisms with Bayesian inference for robust and clinically reliable medical image analysis.

References

Aalishah, R.; Navardi, M.; and Mohsenin, T. 2025. MedMambaLite: Hardware-Aware Mamba for Medical Image Classification. *arXiv preprint arXiv:2508.05049*.

Abdar, M.; Fahami, M. A.; Chakrabarti, S.; Khosravi, A.; Pławiak, P.; Acharya, U. R.; Tadeusiewicz, R.; and Nahavandi, S. 2021. BARF: A new direct and cross-based binary residual feature fusion with uncertainty-aware module for medical image classification. *Information Sciences*, 577: 353–378.

Abdar, M.; Fahami, M. A.; Rundo, L.; Radeva, P.; Frangi, A. F.; Acharya, U. R.; Khosravi, A.; Lam, H.-K.; Jung, A.; and Nahavandi, S. 2022. Hercules: Deep hierarchical attentive multilevel fusion model with uncertainty quantification for medical image classification. *IEEE Transactions on Industrial Informatics*, 19(1): 274–285.

Ansari, S. A.; Agrawal, A. P.; Wajid, M. A.; Wajid, M. S.; and Zafar, A. 2024. MetaV: A pioneer in feature augmented meta-learning based vision transformer for medical image classification. *Interdisciplinary Sciences: Computational Life Sciences*, 16(2): 469–488.

Asia Pacific Tele-Ophthalmology Society (APTOS). 2019. APTOS 2019 Blindness Detection. https://www.kaggle.com/competitions/aptos2019-blindness-detection. Kaggle dataset.

Beyer, L.; Zhai, X.; and Kolesnikov, A. 2022. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*.

Cai, Z.; Chen, Y.; Wang, J.; He, X.; Pei, Z.; Lei, X.; and Lu, C. 2025. DAFNet: A novel Dynamic Adaptive Fusion Network for medical image classification. *Information Fusion*, 103507.

Chen, C.-F.; Panda, R.; and Fan, Q. 2021. Regionvit: Regional-to-local attention for vision transformers. *arXiv preprint arXiv:2106.02689*.

Cheng, J.; Tian, S.; Yu, L.; Gao, C.; Kang, X.; Ma, X.; Wu, W.; Liu, S.; and Lu, H. 2022. ResGANet: Residual group attention network for medical image classification and segmentation. *Medical Image Analysis*, 76: 102313.

Chowdary, G. J.; and Yin, Z. 2024. Med-former: A transformer based architecture for medical image classification. In *International conference on medical image computing and computer-assisted intervention*, 448–457. Springer.

Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M. E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Dai, Y.; Gao, Y.; and Liu, F. 2021. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8): 1384.

Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; and Sun, J. 2021. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13733–13742.

Djoumessi, K.; Mensah, S. O.; and Berens, P. 2025. A Hybrid Fully Convolutional CNN-Transformer Model for Inherently Interpretable Medical Image Classification. *arXiv preprint arXiv:2504.08481*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, L.; Wang, C.; Li, S.; Rabbani, H.; Chen, X.; and Liu, Z. 2019. Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification. *IEEE transactions on medical imaging*, 38(8): 1959–1970.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; and Hu, S.-M. 2023. Visual attention network. *Computational visual media*, 9(4): 733–752.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, X.; Deng, Y.; Fang, L.; and Peng, Q. 2021. Multi-modal retinal image classification with modality-specific attention network. *IEEE transactions on medical imaging*, 40(6): 1591–1602.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Hussain, T.; Shouno, H.; Hussain, A.; Hussain, D.; Ismail, M.; Mir, T. H.; Hsu, F. R.; Alam, T.; and Akhy, S. A. 2025. EFFResNet-ViT: A fusion-based convolutional and vision transformer model for explainable medical image classification. *IEEE Access*.

Ju, L.; Wang, X.; Wang, L.; Mahapatra, D.; Zhao, X.; Zhou, Q.; Liu, T.; and Ge, Z. 2022. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE transactions on medical imaging*, 41(6): 1533–1546.

Li, J.; Feng, M.; and Xia, C. 2024. DBCvT: Double Branch Convolutional Transformer for Medical Image Classification. *Pattern Recognition Letters*, 186: 250–257.

Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; and Ren, J. 2022. Efficientformer: Vision

transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35: 12934–12949.

Liang, G.; Zhang, Y.; Wang, X.; and Jacobs, N. 2020. Improved trainable calibration method for neural networks on medical imaging classification. *arXiv preprint arXiv:2009.04057*.

Liu, S.; Wang, L.; and Yue, W. 2024. An efficient medical image classification network based on multi-branch CNN, token grouping Transformer and mixer MLP. *Applied Soft Computing*, 153: 111323.

Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 116–131.

Manzari, O. N.; Ahmadabadi, H.; Kashiani, H.; Shokouhi, S. B.; and Ayatollahi, A. 2023. MedViT: a robust vision transformer for generalized medical image classification. *Computers in biology and medicine*, 157: 106791.

Manzari, O. N.; Asgariandehkordi, H.; Koleilat, T.; Xiao, Y.; and Rivaz, H. 2025. Medical image classification with kan-integrated transformers and dilated neighborhood attention. *arXiv preprint arXiv:2502.13693*.

Mehta, S.; and Rastegari, M. 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*.

Nickparvar, M. 2021. Brain Tumor MRI Dataset.

Okolo, G. I.; Katsigiannis, S.; and Ramzan, N. 2022. IEViT: An enhanced vision transformer architecture for chest X-ray image classification. *Computer Methods and Programs in Biomedicine*, 226: 107141.

Pogorelov, K.; Randel, K. R.; Griwodz, C.; Eskeland, S. L.; de Lange, T.; Johansen, D.; Spampinato, C.; Dang-Nguyen, D.-T.; Lux, M.; Schmidt, P. T.; et al. 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, 164–169.

Raghu, M.; Zhang, C.; Kleinberg, J.; and Bengio, S. 2019. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32.

Rajaraman, S.; Ganesan, P.; and Antani, S. 2022. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PloS one*, 17(1): e0262838.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shaik, N. S.; Cherukuri, T. K.; Veeranjaneulu, N.; and Bodapati, J. D. 2024. Medtransnet: advanced gating transformer network for medical image classification. *Machine Vision and Applications*, 35(4): 73.

Shastri, S.; Kansal, I.; Kumar, S.; Singh, K.; Popli, R.; and Mansotra, V. 2022. CheXImageNet: a novel architecture for accurate classification of Covid-19 with chest x-ray digital images using deep convolutional neural networks. *Health and technology*, 12(1): 193–204.

Sherkatghanad, Z.; Abdar, M.; Bakhtyari, M.; Pławiak, P.; and Makarenkov, V. 2025. BayTTA: Uncertainty-aware medical image classification with optimized test-time augmentation using Bayesian model averaging. *Knowledge-Based Systems*, 114123.

Swati, Z. N. K.; Zhao, Q.; Kabir, M.; Ali, F.; Ali, Z.; Ahmed, S.; and Lu, J. 2019. Brain tumor classification for MR images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, 75: 34–46.

Tajbakhsh, N.; Shin, J. Y.; Gurudu, S. R.; Hurst, R. T.; Kendall, C. B.; Gotway, M. B.; and Liang, J. 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5): 1299–1312.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34: 24261–24272.

Trockman, A.; and Kolter, J. Z. 2022. Patches are all you need? *arXiv preprint arXiv:2201.09792*.

Wen, Y.; Chen, B.; Shi, W.; Feng, D.; Cao, W.; and Wu, S. 2025. MSPFM: Multi-Scale Pyramid Fusion Mamba for Medical Image Classification: Y. Wen et al. *The Visual Computer*, 1–16.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, X.; and Gou, G. 2025. Uncertainty bidirectional guidance of multi-task mamba network for medical image classification and segmentation. *Signal, Image and Video Processing*, 19(1): 29.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Xue, C.; Dou, Q.; Shi, X.; Chen, H.; and Heng, P.-A. 2019. Robust learning at noisy labeled medical images: Applied to skin lesion classification. In *2019 IEEE 16th International symposium on biomedical imaging (ISBI 2019)*, 1280–1283. IEEE.

Yu, W.; Zhou, P.; Yan, S.; and Wang, X. 2024. Inceptionnext: When inception meets convnext. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 5672–5683.

Yue, Y.; and Li, Z. 2024. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*.