ProSyn-Net: A Reliable Prototypical Synergy Network for Imbalanced Dermatologic Multimodal Ultrasound Diagnosis

Jicheng Lin[1†], Xiangning Dai[1†], Beidi Wang[1†], Juncai Feng[1], Haotian Liu[1], Chenke Yu[1], Ruimeng Liu[1], Ziwei Qin[2], Yujing Zhao[2](✉), and Ye Luo[1](✉)

[1] School of Computer Science and Technology, Tongji University, Shanghai, China
yeluo@tongji.edu.cn
[2] Shanghai Skin Disease Hospital, School of Medicine, Tongji University, Shanghai, China
zhaoyj@shskin.com

Abstract. Multimodal ultrasound, combining B-mode and color Doppler, is vital for the early screening of skin lesions. However, its automated clinical application is severely hindered by modality quality disparities, incidence-driven data imbalance, and model over-confidence. To this end, we propose ProSyn-Net, a unified multimodal diagnostic framework comprising three synergistic modules. First, a dynamic cross-attention fusion module adaptively aligns morphological and hemodynamic features while suppressing modality noise. The fused representations are then processed by a non-linear prototypical mapping module, which effectively disentangles complex, non-convex minority samples from majority clusters in a high-dimensional space. Finally, an uncertainty-aware manifold ensemble leverages the geometric diversity of these prototypes to yield reliable predictions and extract predictive uncertainty. Extensive experiments demonstrate that ProSyn-Net achieves state-of-the-art performance on the primary in-house imbalanced 6-class dataset (78.60% Macro F1 and 96.66% AUC), while showing promising transferability on a complementary independent public binary dataset (87.80% accuracy). We further analyze uncertainty-guided selective deferral for ambiguous cases. Code is available at https://github.com/EdualcLin/ProSyn-Net.

Keywords: Multimodal Ultrasound · Imbalanced Learning · Prototypical Network · Ensemble Synergy · Uncertainty Estimation.

1 Introduction

The high incidence of skin cancer and various skin lesions necessitates timely and accurate early diagnosis for effective clinical intervention [26,24]. High-frequency ultrasound (HFUS) has emerged as an indispensable, non-invasive imaging modality for dermatologic evaluation. In clinical practice, relying on a

[1] † Equal contribution. ✉ Corresponding author.

single modality is often insufficient for a definitive diagnosis. Consequently, the combination of morphological B-mode images (providing structural and boundary characteristics) and color Doppler images (capturing hemodynamic and vascularization patterns) has become the gold standard for comprehensively evaluating skin lesions [25,14].

Despite the remarkable progress of deep learning in medical image analysis, deploying automated diagnostic systems for multimodal ultrasound skin lesions faces three severe clinical hurdles. First, *modality quality disparities*: blood flow signals vary drastically across different lesions, and Doppler images of certain benign plaques may consist almost entirely of noise [24,25]. Conventional fusion methods (e.g., direct concatenation or static attention) fail to adaptively assess modality informativeness, introducing severe interference [1]. Second, *incidence-driven data imbalance*: the natural incidence of skin diseases inevitably leads to severe class imbalance, which inherently hinders the identification of critical but rare lesions [12,11,2]. While prototypical learning mitigates this to some extent, traditional linear metric approaches still impose rigid linear Voronoi decision boundaries. Under high data variance, these linear boundaries cause the non-convex, irregular feature manifolds of minority classes to be severely squeezed and misclassified. Third, *model over-confidence*: existing medical AI models typically output deterministic predictions with unjustifiably high confidence, lacking the capacity to quantify predictive uncertainty [15,10,3]. In high-stakes settings, forced predictions on ambiguous cases may introduce clinical risks.

To circumvent these bottlenecks, we propose **ProSyn-Net**, a reliable prototypical synergy network for imbalanced dermatologic multimodal ultrasound diagnosis. Specifically, a **dynamic cross-attention fusion** module first adaptively fuses morphological and hemodynamic features, suppressing interference from low-quality modalities to overcome quality disparities. The fused representations are then processed by a **non-linear prototypical mapping** module. By decoding a distance profile into a discriminative subspace, this module effectively disentangles irregular minority samples from prevalent clusters, directly addressing feature space distortion caused by extreme imbalance. Finally, to counter model over-confidence, an **uncertainty-aware manifold ensemble** leverages the topological diversity of the geometric prototypes across varying data partitions. By extracting a predictive uncertainty metric, this mechanism identifies clinically ambiguous skin lesions and supports a selective deferral strategy, allowing high-uncertainty cases to be reviewed by human experts in a risk-aware workflow [7].

The main contributions of this paper are summarized as follows: (1) We propose ProSyn-Net, a unified multimodal diagnostic framework integrating dynamic cross-attention fusion and non-linear prototypical mapping to effectively tackle modality quality disparities and extreme incidence-driven data imbalance. (2) We introduce an uncertainty-aware manifold ensemble strategy that leverages the geometric diversity of cross-fold prototypes to yield significant ensemble gain and support uncertainty-guided selective deferral analysis. (3) Extensive experiments show state-of-the-art performance on the primary in-house 6-class

dataset (78.60% Macro F1, 96.66% AUC) and complementary transferability on an independent public binary dataset (87.80% accuracy).

2 Methodology

Given paired B-mode (x_b) and color Doppler (x_d) images, ProSyn-Net processes them through three synergistic stages as illustrated in Fig. 1.

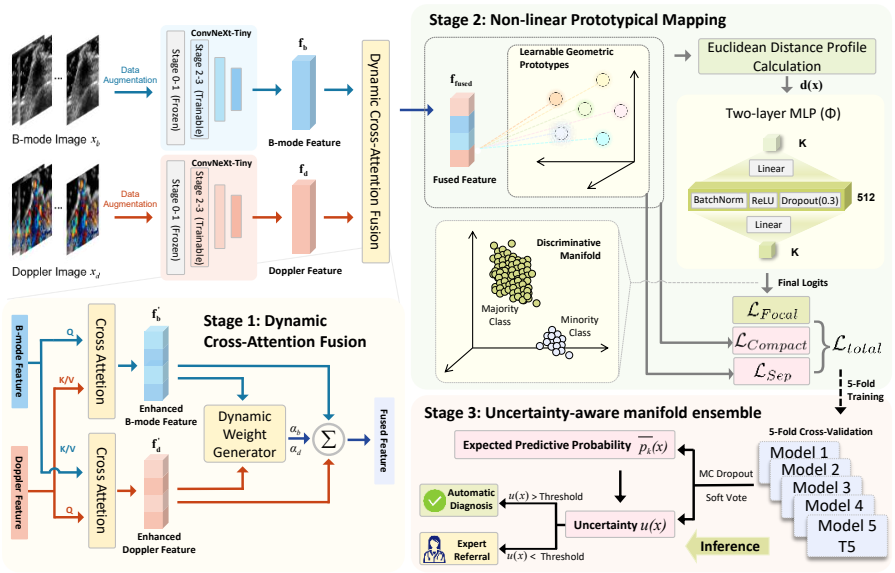

Fig. 1. The overall architecture of ProSyn-Net comprising dynamic cross-attention fusion, non-linear prototypical mapping, and uncertainty-aware manifold ensemble with predictive uncertainty ($\mathcal{U}$) for uncertainty-guided selective deferral.

2.1 Stage 1: Dynamic Cross-Attention Fusion

We employ a pre-trained ConvNeXt-Tiny [19] as a dual-branch backbone to extract features $\mathbf{f}_b, \mathbf{f}_d \in \mathbb{R}^D$, with the first two stages frozen to preserve low-level texture representations. To exploit cross-modal complementarity while suppressing modality noise, we design a dynamic cross-attention fusion module, where $\text{Attention}(Q, C)$ denotes multi-head cross-attention [22] with Q as query and C providing key-value pairs. Specifically, $\mathbf{f}_b$ serves as query against $\mathbf{f}_d$, and symmetrically vice versa. A lightweight dynamic weight generator (an MLP with batch normalization) adaptively produces normalized fusion weights $[\alpha_b, \alpha_d] = \text{Softmax}(\text{MLP}([\mathbf{f}_b \parallel \mathbf{f}_d]))$, yielding the fused representation:

$$\mathbf{f}_{fused} = \alpha_b \cdot \text{Attention}(\mathbf{f}_b, \mathbf{f}_d) + \alpha_d \cdot \text{Attention}(\mathbf{f}_d, \mathbf{f}_b) \tag{1}$$

This sample-adaptive weighting suppresses interference from low-quality or uninformative modalities.

2.2 Stage 2: Non-linear Prototypical Mapping

Incidence-driven data imbalance naturally causes the feature manifolds of minority classes to be severely compressed by prevalent classes. To address this, we define a set of learnable geometric prototypes $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_K\} \in \mathbb{R}^{K \times D}$, where K represents the total number of lesion categories. Instead of applying a fully connected classifier, we compute the squared Euclidean distance between the fused feature $\mathbf{f}_{fused}$ and all prototypes, constructing a distance profile $\mathbf{d}(x) \in \mathbb{R}^K$:

$$d_k(x) = \|\mathbf{f}_{fused} - \mathbf{P}_k\|_2^2 \tag{2}$$

Traditional prototypical networks [21] map $-d_k(x)$ directly into a Softmax function, mathematically equating to imposing strict linear Voronoi decision boundaries. Under high data variance, such rigid boundaries easily misclassify irregular minority samples. In contrast, ProSyn-Net decodes $\mathbf{d}(x)$ through a non-linear mapping network $\Phi : \mathbb{R}^K \to \mathbb{R}^K$ (a two-layer MLP with hidden dimension 512, equipped with Dropout and Batch Normalization) to produce the final logits $\mathbf{z} = \Phi(\mathbf{d}(x))$. This non-linear mapping dynamically transforms the distance profile into a discriminative subspace, effectively disentangling the non-convex feature distributions of minority classes from majority clusters.

To ensure topological discriminability, the network is optimized via a hybrid loss: $\mathcal{L}_{total} = \mathcal{L}_{Focal} + \lambda_c \mathcal{L}_{Compact} + \lambda_s \mathcal{L}_{Sep}$. The focal loss $\mathcal{L}_{Focal}$ [17] mitigates absolute numerical sample imbalance. The compact loss [23] forces the features to be clustered around their ground-truth class prototype $\mathbf{P}_y$ (where $y \in \{1, \ldots, K\}$), formulated as $\mathcal{L}_{Compact} = \|\mathbf{f}_{fused} - \mathbf{P}_y\|_2$. Notably, for the prototype separation loss $\mathcal{L}_{Sep}$, rather than using unbounded distance maximization, we formulate an exponential decay penalty:

$$\mathcal{L}_{Sep} = \frac{1}{N_{pair}} \sum_{i \neq j} \exp(-\|\mathbf{P}_i - \mathbf{P}_j\|_2) \tag{3}$$

where $N_{pair} = K(K-1)$ is the number of distinct prototype pairs. This smooth, bounded formulation effectively separates inter-class prototypes without driving them to infinity.

2.3 Stage 3: Uncertainty-Aware Manifold Ensemble

To counter over-confidence, we use an uncertainty-aware selective deferral mechanism. Since Φ is sensitive to initialization and data partitioning, prototypes converge to diverse local structures across folds [13], amplifying ensemble benefit. During inference, we ensemble all 5 models trained on different cross-validation

folds. To robustly smooth the predictive distribution, we keep Dropout active and perform $M = 50$ Monte Carlo (MC) stochastic forward passes across all ensemble models [6]. The mean predictive distribution $\bar{p}_k(x)$ is aggregated over all passes, and the predictive uncertainty is defined as $\mathcal{U}(x) = 1 - \max_k \bar{p}_k(x)$ [7], serving as an effective criterion for identifying ambiguous cases and deferring them to human experts.

3 Experimental Results and Analysis

3.1 Datasets and Implementation Details

Datasets. We comprehensively evaluated our framework on two dual-modal ultrasound datasets. The **in-house 6-class dataset** comprises 917 pathologically confirmed pairs collected from a specialized hospital. This retrospective study was approved by the Clinical Research Ethics Committee of Shanghai Skin Disease Hospital (approval No. 2023-10), with informed consent waived and all data de-identified before analysis. As shown in Table 1, the data reflects natural clinical incidence with class imbalance. To prevent data leakage, we partitioned this cohort into an 85% training/validation set (for standard 5-fold cross-validation) and a 15% independent hold-out test set. We also incorporated a **public binary dataset** [14] containing 202 pairs (133 benign and 69 malignant lesions), which was randomly split into 80% training (with 5-fold cross-validation) and 20% independent test sets. As this cohort is binary, it serves as a complementary transferability check rather than matched 6-class validation.

Table 1. Class distribution of the in-house dataset (imbalance ratio up to 4.7:1).

Category	Full Name	Pairs (B/Doppler)	Percentage (%)
BCC	Basal Cell Carcinoma	325	35.4
SCC	Squamous Cell Carcinoma	203	22.1
Paget	Paget's Disease	132	14.4
Bowen	Bowen's Disease	97	10.6
AK	Actinic Keratosis	91	9.9
SK	Seborrheic Keratosis	69	7.5
Total	—	**917**	**100.0**

Implementation Details. All images were resized to 224×224 and normalized. Online augmentations (random cropping, horizontal flipping, rotation $\pm 10^\circ$, photometric distortions) were applied. The network was trained for 80 epochs (batch size 16) with AdamW (lr 1×10^{-4}, wd 5×10^{-3}), a 5-epoch linear warmup, multi-step decay, and gradient clipping. Loss weights: $\lambda_c = 0.2$, $\lambda_s = 0.02$; focal loss label smoothing 0.05 ($\gamma = 2$). All reported metrics are macro-averaged across classes. *Fairness:* All baselines use identical augmentation protocols and 5-fold ensemble, except OvcaFinder-DL [27], which is evaluated

with its built-in 6-model heterogeneous ensemble. For task-adapted comparison, flexible baselines adopt ConvNeXt-Tiny and Focal Loss, while specialized systems retain their original designs. ProSyn-Net contains 61.0M parameters (8.93 GMACs). Disabling MC Dropout reduces single-GPU inference latency to ∼21.5 ms/sample on an NVIDIA RTX PRO 6000, with a 2.23% F1 trade-off (Table 3); the full 5-fold MC-enabled ensemble requires ∼1.25 s/sample and is more suitable when sufficient GPU resources are available.

3.2 Comparison with State-of-the-Art Methods

Baselines include four general multimodal fusion methods (Cross-ViT, CMX, EmbraceNet, Hyper-Fusion) and three medical ultrasound-specific systems (FAMF-Net, MUVF-YOLOX, OvcaFinder-DL), all cited in Table 2.

Table 2. State-of-the-art comparison on both datasets (%). **Bold**: best; underline: second best. ($^*p < 0.001$ vs. all baselines, McNemar's test.)

In-house Dataset (6-Class)							
Method	**Acc**	**AUC**	**F1**	**Sens.**	**Spec.**	**F1$_{BCC}$**	**F1$_{SK}$**
FAMF-Net [18]	60.82	90.62	53.18	54.45	90.85	71.88	42.86
Cross-ViT [4]	66.67	91.37	61.78	60.24	92.47	72.28	56.67
MUVF-YOLOX [16]	72.26	94.45	69.15	66.98	93.72	74.11	57.11
Hyper-Fusion [8]	72.30	95.15	69.04	67.31	93.77	78.90	60.88
EmbraceNet [5]	73.04	95.21	69.95	68.60	94.00	79.03	69.45
OvcaFinder-DL [27]	73.72	93.57	68.99	68.72	94.28	79.59	46.15
CMX [28]	75.00	95.89	70.81	69.50	94.44	81.07	62.14
ProSyn-Net (Ours)	**80.14***	**96.66***	**78.60***	**76.82**	**95.47**	**83.73**	**80.79**
Public Dataset (Binary)							
Method	**Acc**	**PPV**	**NPV**	**Sens.**	**Spec.**	**F1$_{ben.}$**	**F1$_{mal.}$**
Cross-ViT [4]	75.61	66.67	79.31	57.14	85.19	82.14	61.54
Hyper-Fusion [8]	78.05	69.23	82.14	64.29	85.19	83.61	66.67
EmbraceNet [5]	80.49	75.00	82.76	64.29	88.89	85.69	69.23
MUVF-YOLOX [16]	80.65	72.73	85.00	75.86	78.86	81.82	74.28
FAMF-Net [18]	80.70	79.49	80.43	73.08	73.08	76.58	76.14
CMX [28]	82.93	**81.82**	83.33	64.29	**92.59**	87.72	72.01
OvcaFinder-DL [27]	86.59	77.50	92.16	**86.51**	86.51	89.26	81.75
ProSyn-Net (Ours)	**87.80***	80.00	**92.31**	85.71	88.89	**90.55**	**82.75**

Results on the In-house 6-Class Dataset. As shown in Table 2, ProSyn-Net achieves 78.60% macro-F1 and 96.66% AUC on the unseen test set, surpassing all baselines ($p < 0.001$, McNemar's test). The per-class breakdown reveals the source of this advantage: on majority class BCC, ProSyn-Net attains 83.73% F1 (Δ+2.66% over CMX); more strikingly, on extreme minority class SK (7.5% prevalence), it achieves 80.79% F1—an 18.65-point gain over CMX (62.14%) and 11.34 points over the next-best EmbraceNet (69.45%). This balanced dis-

crimination validates the effectiveness of our non-linear prototypical mapping in reshaping the feature manifold.

Results on the Public Binary Dataset: On the complementary independent public binary cohort (Table 2), ProSyn-Net achieves the highest accuracy (87.80%) and NPV (92.31%), outperforming the strongest competitor OvcaFinder-DL (86.59% Acc). Given the task mismatch with our primary 6-class setting, this result is interpreted as complementary evidence of transferability.

3.3 Ablation Study

Table 3 validates each component of ProSyn-Net. Single-modality baselines achieve at most 72.20% F1 (Doppler Only), while our full fusion reaches 78.60%, confirming the complementary value of multimodal integration. Disabling the dynamic weight generator or cross-attention reduces F1 to 73.69%/73.45%, confirming the crucial role of adaptive modality alignment. Removing $\mathcal{L}_{Compact}$ causes the largest single-component drop (70.97% F1), while omitting $\mathcal{L}_{Sep}$ or MC Dropout yields milder degradation. Comparing the single-fold mean (69.32% F1) to the full ensemble reveals a 9.28% absolute gain, nearly twice that of all baselines, confirming that our prototypical mapping induces significantly greater model diversity across folds.

Table 3. Ablation study of ProSyn-Net on the in-house dataset (%).

Model Variants	**Acc**	**AUC**	**F1**	**Sens.**	**Spec.**
B-mode Only	56.20	84.58	46.62	45.91	90.34
Doppler Only	75.18	96.25	72.20	69.26	94.27
Ours w/o Dynamic Weight	76.89	96.32	73.69	71.51	94.74
Ours w/o Cross Att	76.89	96.50	73.45	71.35	94.73
Ours w/o Sep Loss	78.83	96.51	76.34	75.02	95.25
Ours w/o Comp Loss	74.70	95.98	70.97	68.76	94.26
Ours w/o MC Dropout	78.83	96.54	76.37	74.14	95.13
Ours - Pure Distance	76.46	96.18	72.19	70.60	94.79
Ours Single-Fold Mean	72.36	94.18	69.32	68.27	93.85
ProSyn-Net (Ours)	**80.14**	**96.66**	**78.60**	**76.82**	**95.47**

Non-linear Mapping vs. Pure Distance. Reverting to a pure distance classifier degrades F1 to 72.19%. As shown by t-SNE (Fig. 2), under extreme variance, pure distance imposes rigid linear Voronoi boundaries that entangle minority features with prevalent clusters, whereas our non-linear mapping effectively disentangles them into well-separated manifolds.

3.4 Uncertainty Estimation and Clinical Deferral

To reliably identify ambiguous cases in high-stakes clinical settings, we extract the predictive uncertainty $\mathcal{U}(x)$ from our 5-fold MC-enabled ensemble. Empirical evaluation within our framework reveals that $\mathcal{U}(x)$ consistently outperforms

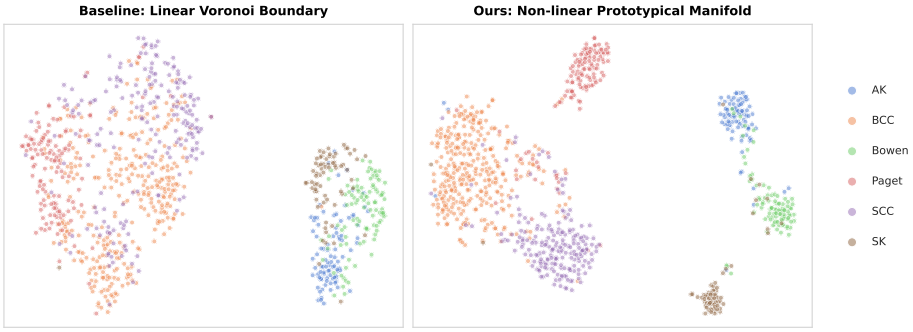

Fig. 2. t-SNE [20] visualization of a challenging data split comparing pure distance (left, entangled) vs. our non-linear mapping (right, disentangled).

Softmax Entropy [9] and MC Dropout Variance in capturing accurate relative prediction risks. As illustrated in Fig. 3, deferring the most uncertain cases to experts causes the accuracy on the remaining autonomous predictions to surge from the 80.14% baseline. At a 50% deferral rate, the autonomous accuracy on the remaining cases exceeds 91%, illustrating the potential of uncertainty-based triage; in practice, the threshold should be selected according to referral burden and expert capacity.

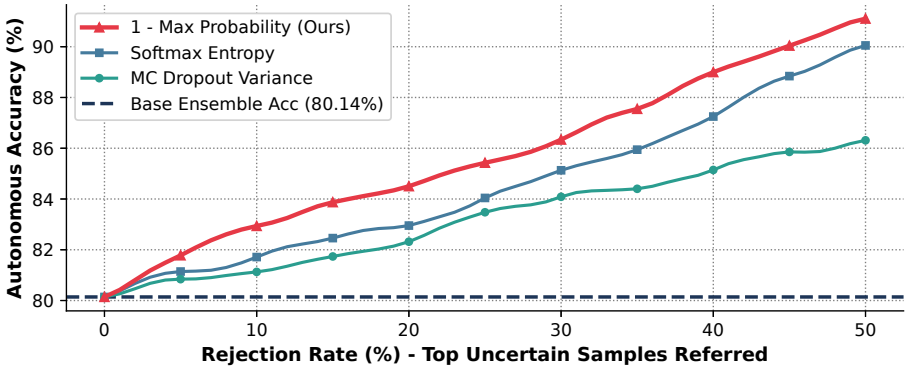

Fig. 3. Clinical deferral curve. Deferring highly uncertain samples (ranked by $\mathcal{U}$) increases the accuracy on the remaining autonomous predictions; the operating point should be selected according to the acceptable referral burden in the target workflow.

4 Conclusion

We proposed ProSyn-Net, a multimodal diagnostic framework for dermatologic ultrasound that integrates dynamic cross-modal attention, non-linear prototypical mapping, and uncertainty-aware manifold ensembling to address modality heterogeneity and incidence-driven imbalance. Experiments show state-of-the-art performance on the primary in-house 6-class cohort and promising transferability on a complementary public binary cohort. The selective deferral analysis further suggests that uncertainty can support risk-aware routing of ambiguous cases to experts under a tunable threshold.

Acknowledgments. This work was supported by the General Program of the National Natural Science Foundation of China under Grant 62276189, the Grants of Tongji University Medicine-X Interdisciplinary Research Initiative under Grant Nos. 2025-0554-YB-09, 2025-0650-YB-17, and 2026-0674-YB-05, and the Fundamental Research Funds for the Central Universities under Grant 2025-1-ZD-02.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 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 (2018)
2. Cai, Z., Wei, T., Lin, L., Chen, H., Tang, X.: Bpaco: Balanced parametric contrastive learning for long-tailed medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 383–393. Springer (2024)
3. Chen, A., Li, Y., Qian, W., Morse, K., Miao, C., Huai, M.: Modeling and understanding uncertainty in medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 557–567. Springer (2024)
4. 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)
5. Choi, J.H., Lee, J.S.: Embracenet: A robust deep learning architecture for multimodal classification. Information Fusion **51**, 259–270 (2019)
6. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
7. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. Advances in neural information processing systems **30** (2017)
8. Ha, D., Dai, A.M., Le, Q.V.: Hypernetworks. In: International Conference on Learning Representations (2017)
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)

10. Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Zhou, Q., Liu, T., Ge, Z.: Improving medical images classification with label noise using dual-uncertainty estimation. IEEE transactions on medical imaging **41**(6), 1533–1546 (2022)
11. Ju, L., Wu, Y., Wang, L., Yu, Z., Zhao, X., Wang, X., Bonnington, P., Ge, Z.: Flexible sampling for long-tailed skin lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 462–471. Springer (2022)
12. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: International Conference on Learning Representations (2020)
13. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
14. Laverde-Saad, A., Jfri, A., García, R., Salgüero, I., Martínez, C., Cembrero, H., Roustán, G., Alfageme, F.: Discriminative deep learning based benignity/malignancy diagnosis of dermatologic ultrasound skin lesions with pretrained artificial intelligence architecture. Skin Research and Technology **28**(1), 35–39 (2022)
15. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Scientific reports **7**(1), 1–14 (2017)
16. Li, J., Huang, H., Ni, D., Xue, W., Zhu, D., Cheng, J.: Muvf-yolox: a multi-modal ultrasound video fusion network for renal tumor diagnosis. In: International conference on medical image computing and computer-assisted intervention. pp. 642–651. Springer (2023)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
18. Liu, Y., Li, J., Zhao, C., Zhang, Y., Chen, Q., Qin, J., Dong, L., Wang, T., Jiang, W., Lei, B.: Famf-net: Feature alignment mutual attention fusion with region awareness for breast cancer diagnosis via imbalanced data. IEEE Transactions on Medical Imaging **44**(3), 1153–1167 (2024)
19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
20. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
21. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30**, I (2017)
23. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. pp. 499–515. Springer (2016)
24. Wortsman, X.: Common applications of dermatologic sonography. Journal of Ultrasound in Medicine **31**(1), 97–111 (2012)
25. Wortsman, X., Alfageme, F., Roustan, G., Arias-Santiago, S., Martorell, A., Catalano, O., Scotto di Santolo, M., Zarchi, K., Bouer, M., Gonzalez, C., et al.: Guidelines for performing dermatologic ultrasound examinations by the dermus group. Journal of Ultrasound in Medicine **35**(3), 577–580 (2016)

26. Wortsman, X., Wortsman, J.: Clinical usefulness of variable-frequency ultrasound in localized lesions of the skin. Journal of the American Academy of Dermatology **62**(2), 247–256 (2010)
27. Xiang, H., Xiao, Y., Li, F., Li, C., Liu, L., Deng, T., Yan, C., Zhou, F., Wang, X., Ou, J., et al.: Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis. Nature Communications **15**(1), 2681 (2024)
28. Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R.: Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. IEEE Transactions on intelligent transportation systems **24**(12), 14679–14694 (2023)