# Low-Resource Diabetic Retinopathy Screening via AUM-ST with Vision Transformer

**Praneeth Rikka[1], S. M. Saiful Islam Badhon[1], Mohammad Adibuzzaman[2], Abu Saleh Mohammad Mosa[3], Ana D. Cleveland[1], Junhua Ding[4], K. S. M. Tozammel Hossain[1,4]**

[1]Department of Information Science, University of North Texas, Denton, TX, USA
[2]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, OR, USA
[3]Department of Biomedical Informatics & Data Science, University of Alabama at Birmingham, USA
[4]Anuradha and Vikas Sinha Department of Data Science, University of North Texas, Denton, TX, USA

***Abstract*—Diabetic Retinopathy (DR) is a leading cause of preventable blindness globally, yet early diagnosis faces significant challenges, including a lack of annotated data and limited access to ophthalmic expertise. While machine learning based screening systems (e.g., Convolutional Neural Networks) show promise for DR prediction using retinal image analysis, they typically require a large number of labeled images, an expensive and time-intensive demand that is unsuitable for resource-constrained settings. Semi-supervised learning (SSL) offers an alternative by harnessing abundant unlabeled images alongside limited labeled data. However, existing SSL methods for DR detection struggle with unreliable pseudo-labels. We propose a novel SSL framework that adapts area under margin self-training (AUM-ST), originally developed for natural language processing, to medical imaging for DR detection. Our approach addresses unreliable pseudo-labelling by leveraging training dynamics to assess label quality. We replace the text-based BERT encoder with BEiT (Bidirectional Encoder representation from Image Transformers) to process retinal fundus images while preserving AUM-ST's core reliability assessment mechanism. Additionally, we implement saliency-preserving augmentation strategies that avoid aggressive transformations such as random rotations and crops, thereby retaining critical lesion features essential for accurate diagnosis. Experimental validation on a real-world dataset of 757 retinal fundus images demonstrates that the proposed method outperforms established baselines, including ResNet-18, ResNet-50, DenseNet-121, MobileNet-v2, and EfficientNet-B0. Our approach achieves substantial improvements in accuracy and F1 scores when training with fewer than 50 labeled samples per class, an advancement over traditional semi-supervised methods. The results demonstrate the practical viability of the proposed method for medical imaging applications and its potential for DR screening in resource-limited environments, including rural healthcare facilities and developing regions.**

***Index Terms*—Diabetic retinopathy, AUM, vision transformers, self-training, low resource**

## I. Introduction

Diabetic Retinopathy (DR) is a diabetes-induced disease that damages blood vessels in the eye. More than one-third of people with diabetes develop some form of DR during their lifetime [1], [2]. This disease is a leading cause of blindness worldwide, but it is preventable if detected early. The prevalence of DR could rise significantly as global diabetes rates are expected to increase [1], [2]. Detecting DR in its early stages is challenging, as it often shows no symptoms initially. Hence, timely detection and treatment are essential to prevent progression to more severe forms, such as proliferative DR and permanent vision loss.

Early detection of DR can be extremely beneficial to patients, as there are treatments available. Clinicians usually manually inspect retinal fundus images and identify DR stages. However, it can be difficult to identify blood vessels in the image and to distinguish the first three stages of DR from a normal eye. With advances in machine-learning-based image analysis, there is growing interest in developing automated DR detection methods that can classify various DR types. Most of these methods use a supervised learning (SL) setting. For example, convolutional neural networks (CNNs) have demonstrated performance comparable to that of expert ophthalmologists in detecting DR from fundus images [3], [4]. More recent studies use complex architectures and vision transformers to further improve DR diagnostic accuracy. However, these SL-based methods require a considerable amount of labeled instances to train the model. Acquiring such large-scale labeling is time-consuming, expensive, and dependent on clinicians' expertise, making it infeasible in many real-world, resource-limited healthcare settings.

Semi-supervised learning (SSL) is a promising approach for detecting DR with a limited number of labeled instances alongside a large pool of unlabeled ones. SSL methods such as self-training, FixMatch, and FlexMatch generate pseudo-labels for unlabeled data and iteratively retrain the model using high-confidence predictions [5], [6]. Although these approaches show strong performance in computer vision tasks, their effectiveness in medical imaging remains limited. Retinal lesions associated with early-stage DR are often subtle and easily misclassified. E.g., pathological features such as microaneurysms and mild bleeding occupy small regions and exhibit low contrast. These subtleties in images may lead SSL-methods to generate noisy pseudo-labels, which could significantly degrade model performance. As a remedy, we can discard noisy pseudo-labels. However, filtering out unreliable pseudo-labels remains a challenge, and current SSL methods lack robust mechanisms to assess pseudo-label quality.

To address the above limitation in SSL methods for DR detection, we propose a novel semi-supervised learning framework that adapts the area under margin self-training (AUM-ST) method, which is developed for text classification

tasks [7]. Unlike confidence-based methods [5], [6], AUM-ST evaluates pseudo-label quality using training dynamics by measuring the area under the margin between the predicted class and competing classes across training epochs. This temporal perspective enables more robust identification of reliable pseudo-labels. To support retinal image representation, we replace the BERT backbone used in the original AUM-ST formulation with BEiT (Bidirectional Encoder Representation from Image Transformers), a transformer-based architecture pre-trained for image understanding [8] (see Fig 1). In addition, we employ saliency-preserving augmentation strategies to protect clinically important lesion features and avoid any transformations that could distort pathological regions. The contributions of this study are as follows:

- We propose a semi-supervised framework for early DR detection that adapts AUM-based pseudo-label filtering originally designed for text classification to fundus image analysis. To our knowledge, this is the first application of AUM-guided self-training to medical image classification.
- We perform extensive experiments using a real-world fundus image dataset. The proposed method outperforms the state-of-the-art semi-supervised baselines, including ResNet-18, ResNet-50, DenseNet-121, MobileNetV2, and EfficientNet-B0.
- We show that our approach is efficient with a fewer number labeled instances, which is a critical advantage for clinical deployment in a low-resource setting.

## II. Literature Review

### A. Semi-Supervised Learning

Semi-supervised learning (SSL) aims to improve model performance by leveraging a small set of labeled data alongside a large pool of unlabeled data. One of the earliest SSL paradigms is self-training, introduced by Yarowsky [9], where a classifier trained on limited labeled data is iteratively refined by incorporating confidently predicted unlabeled samples. This foundational work influenced a wide range of SSL applications. Rosenberg et al. [10] demonstrate the applicability of self-training in object detection tasks, showing that iterative label expansion can improve real-world performance. Grandvalet and Bengio [11] propose entropy regularization to encourage confident predictions on unlabeled data, enhancing the reliability of pseudo-label selection.

Pseudo-labeling was further formalized by Lee [12], who shows that assigning labels based on maximum predicted probability could effectively incorporate unlabeled data while reducing noise. This method significantly improved deep learning performance in low-label regimes and became a cornerstone of modern SSL pipelines.

Consistency regularization introduces another significant advancement. Miyato et al. [13] propose Virtual Adversarial Training (VAT), enforcing prediction consistency under adversarial perturbations in embedding space. Clark et al. [14] discuss cross-view training, where models learn consistent predictions from partial and complete views of the same input. Xie et al. [15] extend consistency-based learning by applying strong data augmentations to unlabeled data, improving generalization.

Sohn et al. [5] present FixMatch, which combines confidence-based pseudo-labeling with consistency regularization, achieving state-of-the-art performance across multiple benchmarks. Mukherjee and Awadallah [16] incorporate uncertainty estimation into self-training, improving robustness during iterative teacher–student learning. Zhang et al. [6] addressed class imbalance in FixMatch by introducing class-specific confidence thresholds. More recently, Sosea [17] proposes AUM-ST, using the Area Under the Margin metric to refine pseudo-label quality and stabilize training dynamics.

### B. Semi-Supervised Learning in Medical Imaging

The scarcity of labeled medical data and the abundance of unlabeled scans make SSL particularly valuable in medical imaging. Enguehard et al. [18] present SSLDEC, which integrates deep embedded clustering with SSL, achieving competitive results on both natural image datasets and infant brain MRI segmentation tasks using limited annotations.

In ophthalmology, Bengani et al. [19] use convolutional autoencoders pretrained on unlabeled fundus images to improve optic disc segmentation with minimal labeled data. Generative adversarial networks (GANs) have also been explored; Wu et al. [20] propose a semi-supervised GAN framework for fundus image enhancement, improving downstream segmentation accuracy.

Contrastive learning has recently gained traction in medical SSL. Huang et al. [21] propose a saliency-guided semi-supervised contrastive learning framework for DR grading, achieving strong performance with only 10% labeled data. Multi-task SSL has also been effective; Ullah et al. [22] propose SSMD-UNet, a semi-supervised multi-task architecture for DR lesion segmentation, which outperforms single-task models on the FGADR and IDRiD datasets.

## III. Methods

### A. Proposed Methodology

Our proposed methodology is based on AUM-ST, a semi-supervised learning framework introduced by Sosea et al. [17]. AUM-ST enhances model performance in *text classification tasks* by leveraging the training dynamics of unlabeled data via the area under the margin (AUM). This approach analyzes model behavior during training to identify and filter out unreliable pseudo-labels, thereby improving the quality of supervision derived from unlabeled data.

We adapt the AUM-ST framework for image classification with important modifications. While preserving the core idea of using training dynamics to guide self-training, we modify the implementation to suit the characteristics of visual data better. In the following subsections, we first describe the original AUM-ST framework and then present the adaptations introduced for medical image classification.

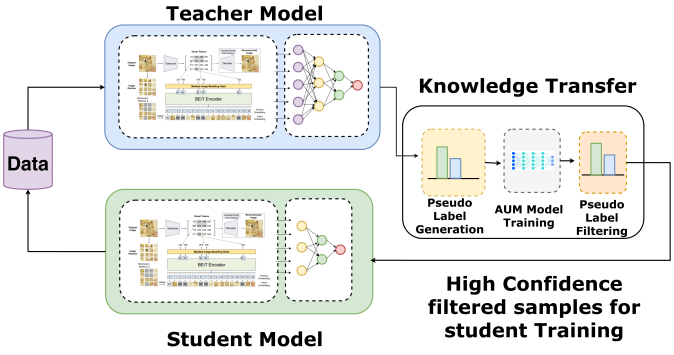


Fig. 1. **From noisy unlabeled data to reliable supervision via SSL.** The diagram illustrates the proposed AUM-ST based framework, where a teacher model generates pseudo-labels, an AUM module filters high-confidence samples using training dynamics, and the refined supervision is transferred to train the student model.

### B. Original AUM-ST Method

The foundation of the area under margin self-training (AUM-ST ) framework lies in the use of the area under margin (AUM), a data quality metric introduced by Pleiss et al. [23]. AUM enables the identification and filtering of noisy pseudo-labels in a semi-supervised learning setting by analyzing training dynamics across epochs. This metric characterizes training examples based on their contribution to model generalization.

Let $T$ denote the total number of training epochs. The margin for an example $(x, y)$ at epoch $t \in T$ is defined as:

$$M_t(x, y) = z_y - \max_{i \neq y}(z_i) \tag{1}$$

where $z_y$ represents the logit corresponding to the true class $y$, and $\max_{i \neq y}(z_i)$ denotes the maximum logit among all incorrect classes.

The AUM score for an example is computed as the average margin across all training epochs:

$$\text{AUM}(x, y) = \frac{1}{T} \sum_{t=1}^{T} M_t(x, y) \tag{2}$$

Pleiss et al. [23] demonstrate that training examples with consistently low AUM scores are more likely to be mislabeled and can negatively affect model generalization. Consequently, such examples are suitable candidates for exclusion during training.

The AUM-ST framework consists of the following components:

1) **Teacher Model Training:** A teacher model is trained using weakly labeled data.
2) **Pseudo-label Generation:** The trained teacher model generates pseudo-labels for unlabeled data.
3) **Training Dynamics Monitoring:** During training, the model tracks the training dynamics of pseudo-labeled examples.
4) **AUM-based Filtering:** Unlabeled examples are evaluated using AUM scores, and samples with low AUM values are filtered out.
5) **Student Model Training:** A student model is trained using a combined objective consisting of:
   - Cross-entropy loss on labeled examples,
   - Cross-entropy loss on pseudo-labeled, unlabeled examples with high AUM scores.
6) **Iterative Self-Training:** The student model is promoted to the role of the teacher, and the process is repeated until convergence.

### C. Proposed Adaptation for Image Classification

Our proposed approach builds upon the AUM-ST framework [17] for image classification tasks. The original AUM-ST approach is designed for text classification and employs BERT as both the teacher and student model. Since BERT is inherently limited to text, we replace it with BEiT (Bidirectional Encoder Representation from Transformers), a transformer-based architecture specifically designed for vision tasks.

BEiT maintains architectural similarities with BERT, making it a suitable replacement that preserves methodological consistency while enabling image-based learning. We deliberately adopt the *base* BEiT model to align with the original study, which employs base BERT rather than larger variants. This choice ensures that any observed performance differences arise from the domain shift from text to images, rather than increased model capacity.

Another key modification concerns data augmentation. While the original AUM-ST framework applied weak and strong augmentations to unlabeled data, we exclude such augmentations in our implementation. Common image augmentations such as flipping, cropping, rotation, or noise injection can distort fine-grained visual features critical to medical imaging tasks, particularly for diabetic retinopathy detection. Even subtle perturbations in retinal images may obscure clinically relevant patterns and negatively affect model learning.

### D. Baselines

We evaluate the proposed method against a set of baselines. All baseline models are trained under a self-training framework using a fixed confidence threshold of 0.95 for pseudo-label selection. Applying the same threshold across all baselines ensures a fair and controlled comparison.

*1) ResNet-18:* ResNet-18 is a lightweight convolutional neural network from the Residual Network (ResNet) family [24]. The method consists of 18 layers and employs residual connections with identity shortcuts to mitigate the vanishing gradient problem.

*2) ResNet-50:* ResNet-50 is a deeper variant of the ResNet architecture, comprising 50 layers organized using bottleneck residual blocks [24]. The bottleneck design enables efficient training of deep networks by reducing computational overhead while preserving representational power.

*3) DenseNet-121:* DenseNet-121 is a member of the Dense Convolutional Network (DenseNet) family, proposed by Huang et al [25]. This architecture consists of 121 layers with dense connectivity, where each layer receives feature maps from all preceding layers.

*4) MobileNetV2:* MobileNetV2, introduced by Sandler et al. [26], is an efficient neural network architecture designed for mobile and embedded vision applications. It utilizes depthwise separable convolutions and inverted residual blocks with linear bottlenecks to significantly reduce model size and computational cost while maintaining competitive accuracy. Its lightweight design makes it particularly suitable for semi-supervised learning in low-resource settings.

*5) EfficientNet-B0:* EfficientNet-B0 is the baseline architecture of the EfficientNet family, developed by Tan and Le [27]. This method uses a compound scaling strategy that uniformly scales the network's depth, width, and input resolution. EfficientNet-B0 achieves strong accuracy with fewer parameters and floating-point operations (FLOPs), making it well-suited for scenarios involving limited labeled data and constrained computational resources.

## IV. Results

We design our experiments to address the following research questions:

- **RQ1:** How does the proposed method perform compared to the baseline models?
- **RQ2:** How does region-focused augmentation affect the performance of the proposed AUM-ST framework?
- **RQ3:** What qualitative insights can be drawn from the model predictions?

### A. Dataset

We use the publicly available *dataset from Fundus Images for the Study of Diabetic Retinopathy* curated by Verónica et al. [28]. This dataset is a standard benchmark for diabetic retinopathy (DR) image analysis and consists of 757 color fundus images acquired at the Department of Ophthalmology, Hospital de Clínicas, Facultad de Ciencias Médicas (FCM), Universidad Nacional de Asunción (UNA), Paraguay.

Each image has a resolution of $2124 \times 2056$ pixels and a field of view of $45°$ (see 2). The dataset contains seven classes: no DR signs, non-proliferative diabetic retinopathy (NPDR) with varying severity levels, proliferative diabetic retinopathy (PDR), and advanced PDR. For the purpose of binary classification, all DR-related categories were merged into a single *DR* class, while images with no pathological signs were retained as the *No DR* class. The final dataset contains 187 negative (No DR) and 570 positive (DR) samples.

### B. Experimental Setup

We partition the dataset into training, validation, testing, and unlabeled subsets. A total of 250 images were selected for the training pool, ensuring class balance between DR and No-DR. From the remaining samples, 100 images were reserved for evaluation (50 for validation and 50 for testing), and the remaining 407 images were treated as unlabeled data (see Fig. 3). The data distribution is summarized as follows:

- **Training set:** 250 samples (balanced, $\sim$33%)
- **Validation + Test set:** 100 samples (imbalanced, $\sim$13%)
- **Unlabeled set:** 407 samples (imbalanced, $\sim$53%)

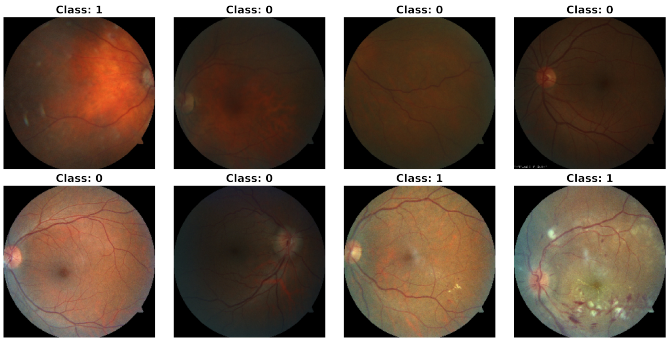


Fig. 2. **Sample Data.** Representative retinal fundus samples from both classes are shown, highlighting the diversity in appearance and subtle visual cues that motivate robust semi-supervised learning.

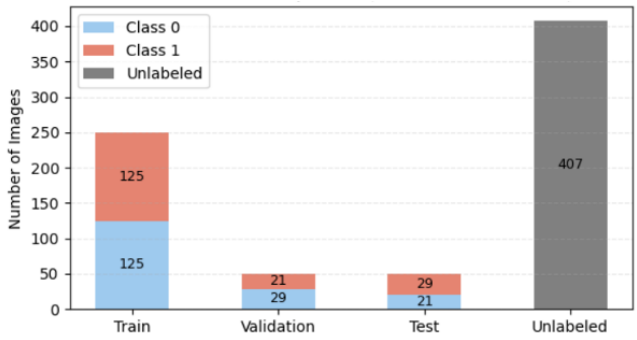


Fig. 3. **Class-balanced labeled data alongside a large unlabeled pool.** The bar chart summarizes the distribution of samples across training, validation, and test splits, emphasizing the balanced labeled sets and the substantially larger unlabeled set leveraged by the proposed semi-supervised framework.

Following the original AUM-ST protocol, we evaluate model performance under varying levels of label scarcity by selecting 5, 10, 15, 20, 50, and 100 labeled samples per class. The remaining samples from the training pool are treated as weakly unlabeled data. The additional unlabeled dataset is used as strongly unlabeled data, although no data augmentation is applied, as justified in the methodology section.

### C. Estimating the performance of the proposed method (RQ1)

We compare the proposed method against baselines in terms of accuracy and F1 scores. The rationale for using these two metrics is rooted in the nature of the test and validation datasets, which are imbalanced (see Dataset). All experiments are conducted twice, and the average accuracy and F1-score values are reported to ensure the reliability and consistency of the results (see Table I). Based on the experimental results, the proposed AUM-ST-based approach consistently outperforms the baseline models across most labeled data regimes. The only exception occurs in the 100-sample-per-class setting, where EfficientNet-B0 achieves marginally higher performance. This behavior can be attributed to the strong inductive biases and scaling strategy of EfficientNet-B0, which may be particularly effective when sufficient labeled data is available.

However, the primary objective of the proposed method is to achieve competitive performance under conditions of limited

TABLE I
PERFORMANCE COMPARISON (ACCURACY AND F1-SCORE) UNDER VARYING NUMBERS OF LABELED SAMPLES PER CLASS. BEST RESULTS ARE BOLD-FACED.

| Models | Number of Samples per Class | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5 | | 10 | | 15 | | 20 | | 50 | | 100 | |
| | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| **ResNet-18** | 0.64 | 0.61 | 0.42 | 0.24 | 0.72 | 0.65 | 0.66 | 0.65 | 0.70 | 0.68 | 0.86 | 0.86 |
| **ResNet-50** | 0.64 | 0.58 | 0.54 | 0.47 | 0.42 | 0.24 | 0.60 | 0.55 | 0.66 | 0.65 | 0.58 | 0.42 |
| **DenseNet-121** | 0.72 | 0.69 | 0.68 | 0.64 | 0.70 | 0.69 | 0.70 | 0.69 | 0.70 | 0.69 | 0.84 | 0.84 |
| **MobileNetV2** | 0.58 | 0.42 | 0.78 | 0.76 | 0.58 | 0.42 | 0.70 | 0.70 | 0.78 | 0.76 | 0.84 | 0.84 |
| **EfficientNet-B0** | 0.62 | 0.50 | 0.56 | 0.56 | 0.48 | 0.41 | 0.78 | 0.78 | 0.82 | 0.81 | **0.92** | **0.91** |
| **AUM-ST (Proposed)** | **0.81** | **0.76** | **0.88** | **0.81** | **0.81** | **0.72** | **0.88** | **0.84** | **0.86** | **0.85** | 0.88 | 0.87 |

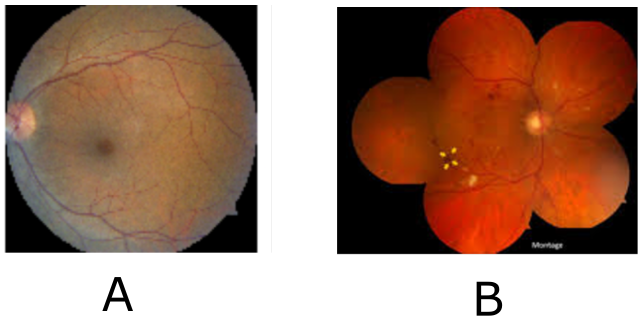


Fig. 4. **Comparison of fundus images**: (a) original image without region-focused augmentation and (b) segmented retinal ROI image obtained using the 7-circle segmentation strategy.

labeled data. In low-resource settings, our approach demonstrates robust, stable performance, achieving 81% accuracy with as few as five labeled samples per class. In contrast, fully supervised models typically require thousands of labeled samples per class to achieve comparable accuracy [29].

### *D. Impact of Region-Focused Augmentation on AUM-ST Performance (RQ2)*

In RQ1, the proposed AUM-ST framework is evaluated using retinal fundus images without data augmentation or region-based preprocessing. This ensured fair comparison across all models but required the network to process the entire image uniformly, which may dilute clinically important lesion information due to background regions.

To examine the robustness of AUM-ST and its ability to leverage region-focused inputs, we conduct additional experiments using segmented retinal regions. Since diabetic retinopathy-related pathological features primarily occur in the retinal region, emphasizing this area can improve discriminative feature learning. We adopted a seven-circle retinal segmentation strategy that highlights concentric retinal regions while suppressing irrelevant background content, allowing the model to focus on anatomically meaningful structures.

Figure 4 illustrates a comparison between an original fundus image and its corresponding segmented retinal region produced using the seven circle segmentation method.

The quantitative results obtained using ROI-segmented images as input to the AUM-ST framework are summarized in Table II. The same experimental protocol and evaluation metrics (Accuracy and F1-score) are maintained to ensure a fair comparison with the non-augmented setting.

As shown in Table II, incorporating segmented ROI inputs consistently and substantially improves performance across all labeled data regimes. The augmented AUM-ST model outperforms both the baseline architectures and the original AUM-ST configuration trained on full, unsegmented images. These gains are particularly significant in low-data scenarios, demonstrating that region-focused augmentation enhances the model's ability to effectively leverage limited labeled data.

### *E. Qualitative Insights from Model Predictions (RQ3)*

To gain insights into the model's decision-making process and validate its clinical relevance, Gradient-weighted Class Activation Mapping (Grad-CAM) is used to visualize regions in retinal images (see Fig. 5).

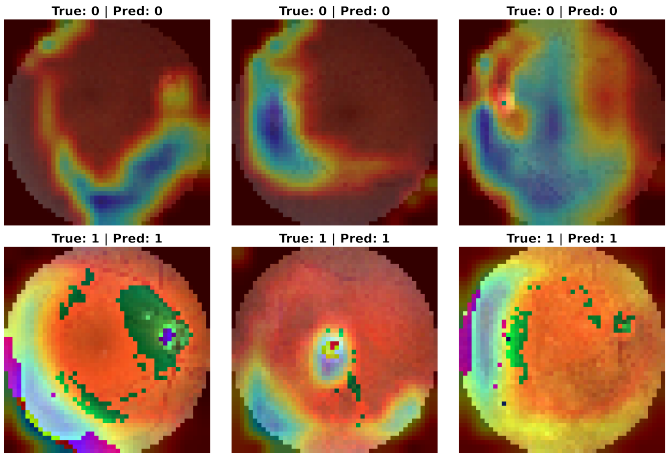


Fig. 5. **Model decisions grounded in clinically relevant retinal regions.** Correct predictions show diffuse attention for normal cases and localized activation for pathological regions, reflecting clinically meaningful focus.

The Grad-CAM visualizations demonstrate that the model exhibits spatial awareness rather than relying on random image regions for classification. For true positive cases, the model concentrates attention on specific anatomical features that correspond to pathological indicators. As for the true negative cases, attention patterns are notably more diffuse, distributed

TABLE II
PERFORMANCE COMPARISON OF AUM-ST WITH AND WITHOUT AUGMENTATION UNDER DIFFERENT LABELED SAMPLE SIZES. BEST RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN YELLOW.

| **Model** | **Number of Samples Per Class** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **5** | | **10** | | **15** | | **20** | | **50** | | **100** | |
| | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| **AUM-ST (No Augmentation)** | 0.81 | 0.76 | **0.88** | **0.81** | 0.81 | 0.72 | 0.88 | 0.84 | **0.86** | **0.85** | 0.88 | 0.87 |
| **AUM-ST (Augmented / ROI)** | **0.88** | **0.87** | 0.82 | 0.81 | **0.86** | **0.85** | **0.90** | **0.89** | 0.84 | 0.84 | **0.96** | **0.95** |

across the retinal surface without strong focal concentrations. This indicates the model has learned to recognize the absence of pathological markers as a diagnostic criterion.

## V. CONCLUSION

We successfully adapt an improved semi-supervised learning (SSL) framework, originally designed for text classification, to medical image analysis. The proposed approach demonstrates improved performance, providing strong evidence that SSL techniques can be effectively leveraged for medical imaging tasks under low-resource conditions. The results of this study highlight the potential of training-dynamics-driven SSL frameworks to reduce dependence on large labeled datasets, thereby enabling scalable and cost-efficient medical image analysis solutions in settings where annotated data and clinical expertise are limited. However, the current evaluation is limited to binary classification with a single dataset. In our future work, we will assess the framework's performance on multiple datasets in both binary and multiclass settings to further validate its generalizability and robustness across diverse medical imaging scenarios.

## REFERENCES


[1] J. W. Yau, S. L. Rogers, R. Kawasaki, E. L. Lamoureux, J. W. Kowalski, T. Bek, S. J. Chen, J. M. Dekker, A. Fletcher, J. Grauslund *et al.*, "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes Care*, vol. 35, no. 3, pp. 556–564, 2012.

[2] N. Cheung, P. Mitchell, and T. Y. Wong, "Diabetic retinopathy," *The Lancet*, vol. 376, no. 9735, pp. 124–136, 2010.

[3] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[4] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Y. Hamzah, R. Garcia-Franco, I. San Yeo, S. Y. Lee *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.

[5] K. Sohn, D. Berthelot, C.-L. Li *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *NeurIPS*, 2020.

[6] B. Zhang *et al.*, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *NeurIPS*, 2021.

[7] T. Sosea and C. Caragea, "Leveraging training dynamics and self-training for text classification," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4750–4762. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.350

[8] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," in *International Conference on Learning Representations (ICLR)*, 2022.

[9] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," *Proceedings of the 33rd Annual Meeting of the ACL*, 1995.

[10] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," *Proceedings of the IEEE Workshop on Object Tracking*, 2005.

[11] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in Neural Information Processing Systems*, 2004.

[12] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop on Challenges in Representation Learning*, 2013.

[13] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[14] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Semi-supervised sequence modeling with cross-view training," *EMNLP*, 2018.

[15] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," *CVPR*, 2020.

[16] S. Mukherjee and A. H. Awadallah, "Uncertainty-aware self-training," *NeurIPS*, 2020.

[17] T. Sosea, "Aum-st: Improving self-training with area under the margin," *ICLR*, 2022.

[18] J. Enguehard *et al.*, "Semi-supervised deep embedded clustering for medical image analysis," *IEEE Transactions on Medical Imaging*, 2019.

[19] J. Bengani *et al.*, "Automatic segmentation of optic disc in retinal fundus images using semi-supervised deep learning." *multimedia tools and applications*, 2021.

[20] H.-T. Wu *et al.*, "Fundus image enhancement via semi-supervised gan and anatomical structure preservation," *IEEE Access*, 2024.

[21] Y. Huang, J. Lyu, P. Cheng, R. Tam, and X. Tang, "Ssit: Saliency-guided self-supervised image transformer for diabetic retinopathy grading," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 5, pp. 2806–2817, May 2024.

[22] A. Ullah *et al.*, "Ssmd-unet: Semi-supervised multi-task learning for diabetic retinopathy lesion segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[23] G. Pleiss, T. Zhang, E. R. Elenberg, and K. Q. Weinberger, "Identifying mislabeled data using the area under the margin," *Advances in Neural Information Processing Systems*, 2020.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *CVPR*, 2017.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *CVPR*, 2018.

[27] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *ICML*, 2019.

[28] V. E. e. a. Castillo Benítez, "Dataset from fundus images for the study of diabetic retinopathy," *Data in Brief*, vol. 36, p. 107068, Jun. 2021.

[29] G. Mushtaq and F. Siddiqui, "Detection of diabetic retinopathy using deep learning methodology," *ICRIET*, 2020.