

LABEL TRANSFER HYPOTHESIS: A CLINICAL PRIOR KNOWLEDGE-GUIDED APPROACH FOR DISEASE DIAGNOSIS

Anonymous authors

Paper under double-blind review

ABSTRACT

Deep learning in medical image analysis requires large-scale, high-quality annotated datasets that are expensive and time-consuming to obtain due to extensive expert involvement. Most existing approaches rely on supervised learning, severely limiting practical deployment given annotation scarcity.

To address this limitation, we propose the Label Transfer Hypothesis (LTH), a theoretical framework for tackling annotation scarcity. The core hypothesis holds that when diseases present characteristic pathological features, precise lesion segmentation guided by clinical diagnosis and treatment logic can act as “implicit diagnostic labels” for disease classification—this enables knowledge transfer from segmentation to classification tasks. This approach not only reduces annotation requirements while retaining the advantages of supervised classification, but also leverages the combination of the label transfer method and clinical diagnosis and treatment logic to obtain more reliable diagnoses.

We validate LTH on diabetic macular edema (DME) and retinal vein occlusion (RVO) classification tasks. Results demonstrate that LTH achieves performance comparable or superior to supervised methods while requiring significantly fewer labeled data. The code will be released after acceptance.

This work contributes: (1) A pioneering theoretical framework bridging segmentation and classification through clinical knowledge integration; (2) Demonstrated feasible knowledge transfer maintaining competitive performance with reduced supervision; (3) A scalable solution for resource-constrained healthcare settings, particularly beneficial for medically underserved regions.

1 INTRODUCTION

Artificial intelligence technologies, particularly deep learning, have demonstrated tremendous potential in medical image analysis, achieving remarkable progress in automated screening and diagnosis of fundus diseases (He et al., 2023; Chen et al., 2021). However, clinical deployment of these AI-driven diagnostic systems faces a fundamental bottleneck: the stringent requirement for large-scale, high-quality expert annotations.

Medical image annotation presents unique challenges distinguishing it from conventional computer vision tasks. Accurate diagnostic labeling necessitates extensive clinical expertise, making the annotation process inherently time-consuming and expensive. The scarcity of qualified specialists capable of providing reliable annotations creates significant resource constraints, while inter-annotator variability introduces inconsistencies that compromise model reliability. These factors collectively create substantial barriers to acquiring the extensive labeled datasets required for robust supervised learning models, severely limiting the scalability and clinical translation of AI technologies in ophthalmic diagnosis.

Existing research has primarily explored weakly-supervised and unsupervised learning paradigms to address annotation scarcity (Kumari & Singh, 2024). However, these approaches exhibit critical limitations in clinical contexts. Weakly-supervised methods often lack the precision necessary for reliable medical diagnosis, while unsupervised approaches suffer from insufficient clinical interpretability due to the absence of explicit pathological basis in model decision-making, and sub-

optimal diagnostic accuracy resulting from the disconnect between data-driven feature learning and established clinical diagnostic logic. Current research lacks a systematic theoretical framework that can effectively reduce annotation dependency while preserving both diagnostic accuracy and clinical interpretability.

In clinical practice, experienced ophthalmologists rely on pattern recognition and prior knowledge when making diagnostic decisions. Rather than processing fundus images holistically, clinicians systematically identify and evaluate specific pathological manifestations—characteristic lesions serving as diagnostic indicators. For instance, when diagnosing Diabetic Macular Edema (DME), ophthalmologists focus on identifying and assessing the spatial distribution of microaneurysms, hard exudates, and intraretinal fluid accumulation in the macular region. This observation-based diagnostic process suggests that pathological features themselves contain rich diagnostic information that could potentially serve as surrogate supervisory signals.

Inspired by clinical diagnostic reasoning, we propose the Label Transfer Hypothesis, an innovative theoretical framework that addresses the annotation bottleneck through indirect label substitution. The core hypothesis posits that when diseases exhibit characteristic imaging manifestations, precise lesion segmentation guided by clinical prior knowledge can serve as “implicit diagnostic labels” for disease classification, enabling knowledge transfer from pixel-level annotations to image-level diagnosis.

This paradigm shift offers several theoretical innovations: (1) Epistemological redefinition: reconceptualizing diagnostic “labels” from externally assigned categorical identifiers to intrinsic pathological feature representations derivable from image content itself; (2) Methodological bridge: establishing formal theoretical connections between segmentation and classification tasks, creating a new paradigm for cross-task knowledge transfer; (3) Clinical alignment: organically integrating established medical diagnostic logic into algorithmic design, significantly enhancing model interpretability and clinical credibility.

To systematically validate the effectiveness and generalizability of our Label Transfer Hypothesis framework, we selected Diabetic Macular Edema (DME) and Retinal Vein Occlusion (RVO) as representative test cases. This selection is strategically motivated by three factors: First, both conditions represent high-prevalence diseases with significant public health impact, ensuring clinical relevance. Second, despite distinct pathophysiological mechanisms, both diseases manifest through clearly identifiable characteristic lesions in fundus imagery, providing ideal testing grounds for our lesion-based approach. Third, well-established clinical guidelines for both conditions provide robust medical foundations for objective lesion definition and validation.

Through comprehensive experiments, this study demonstrates the feasibility and clinical value of the Label Transfer Hypothesis in fundus disease diagnosis while establishing a theoretically grounded and practically viable solution to the persistent data annotation bottleneck constraining AI deployment in healthcare applications.

2 RELATED WORK

2.1 CURRENT STATE OF OPHTHALMIC DISEASE DIAGNOSIS: FROM MANUAL ASSESSMENT TO AUTOMATED ANALYSIS

Retinal diseases, including Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), and Retinal Vein Occlusion (RVO), constitute leading causes of irreversible vision loss worldwide. Current clinical diagnosis relies on manual examination of fundus images by ophthalmologists, who systematically identify individual pathological biomarkers—microaneurysms, hemorrhages, cotton-wool spots, exudates, and drusen—before synthesizing these observations into comprehensive diagnostic assessments.

This manual approach faces critical limitations: the process is labor-intensive with substantial inter-observer variability, while the global shortage of qualified specialists, particularly in underserved regions, restricts access to timely screening. These challenges necessitate automated retinal image analysis systems capable of providing scalable, consistent, and objective diagnosis while maintaining alignment with clinical reasoning processes.

2.2 DEEP LEARNING IN RETINAL IMAGE ANALYSIS: PROGRESS AND PERSISTENT CHALLENGES

Deep learning has achieved significant advances in automated retinal disease diagnosis, with approaches broadly categorized into end-to-end disease classification and lesion segmentation paradigms. Despite notable progress, fundamental challenges persist in bridging computational methods with clinical diagnostic logic.

The annotation-knowledge gap. Supervised models exhibit profound dependence on labeled data, prompting exploration of alternative learning paradigms. Unsupervised approaches have demonstrated promise—Yousefi et al. employed clustering for glaucoma progression monitoring (Yousefi et al., 2014), while Yu et al. developed frameworks for dry eye disease stratification (Matta et al., 2022). However, these methods typically operate on holistic image representations without mechanisms for integrating structured clinical knowledge. They identify patterns but fail to explicitly recognize individual pathological manifestations (hemorrhages, exudates, cotton-wool spots) that clinicians systematically evaluate, creating a fundamental disconnect between data-driven pattern discovery and the sequential, lesion-based reasoning employed in clinical practice.

Architectural approaches to lesion integration. Recent work has attempted to incorporate lesion information through architectural innovations. (?) proposed MSGDA-Net, utilizing lesion segmentation as an auxiliary task to generate regional prior knowledge for DR grading through attention mechanisms. Similarly, (?) developed a multi-view framework combining fundus images with lesion snapshots via heterogeneous convolution blocks. While these approaches recognize the importance of lesion information, they primarily treat lesions as supplementary features for enhancing classification performance rather than as the fundamental basis of diagnosis. This architectural focus diverges from clinical practice, where physicians first identify and characterize individual lesion types before integrating these observations into a final diagnosis—a sequential, evidence-based process that current methods fail to emulate.

The critical limitation remains: existing methods do not systematically mirror the clinical diagnostic workflow of observing individual pathological features and synthesizing these observations into diagnostic conclusions. This misalignment between computational approaches and medical reasoning constitutes a principal barrier to clinical trust and adoption.

2.3 THE LABEL TRANSFER HYPOTHESIS: EMULATING CLINICAL DIAGNOSTIC LOGIC THROUGH THEORETICAL FRAMEWORK

Current literature reveals a fundamental gap: the absence of a theoretical framework that faithfully reproduces the clinical diagnostic process—from individual lesion observation to comprehensive disease assessment. We propose the *Label Transfer Hypothesis* (LTH), a theoretical framework that fundamentally reconceptualizes automated diagnosis to align with clinical reasoning patterns.

Clinical alignment as core principle. Unlike existing methods that leverage lesions as auxiliary information (?) or additional input modalities (?), LTH explicitly models the two-stage clinical diagnostic process: (1) systematic identification and characterization of individual pathological manifestations (hemorrhages, microaneurysms, cotton-wool spots, hard exudates), and (2) synthesis of these lesion patterns into diagnostic conclusions. This approach treats lesion segmentation not as a means to enhance features but as the primary diagnostic evidence—mirroring how clinicians derive diagnoses directly from observed pathological patterns.

From feature enhancement to diagnostic reasoning. While architectural approaches focus on improving classification accuracy through lesion-guided attention or multi-view fusion, LTH establishes a formal theoretical connection between lesion patterns and disease states. The framework posits that precise lesion segmentation can serve as “implicit diagnostic labels,” enabling the model to learn the mapping from pathological evidence to diagnosis—exactly as clinicians are trained. This represents a paradigm shift from using lesions to improve black-box classifiers to building inherently interpretable systems where every diagnostic decision traces back to specific, clinically-defined pathological features.

Theoretical foundation for clinical interpretability. The key innovation of LTH lies not in architectural design but in its theoretical grounding: by formally defining disease labels as derivable from spatial configurations of characteristic lesions, the framework ensures that model reasoning

inherently aligns with medical logic. This alignment is not post-hoc or superficial but fundamental to the model’s operation—each diagnosis emerges from systematic evaluation of individual lesions and their relationships, providing natural interpretability that clinicians can verify against their own diagnostic process.

This theoretical framework offers a unified solution by:

- **Reducing annotation dependency through clinical logic:** Leveraging the objectivity of lesion identification—which follows established clinical criteria—as the foundation for diagnosis, eliminating the need for subjective disease-level labels while maintaining diagnostic accuracy.
- **Ensuring inherent interpretability:** Every diagnostic decision directly corresponds to specific lesion patterns, allowing clinicians to trace and validate the model’s reasoning against their own diagnostic process, fostering clinical trust through transparency.
- **Preserving diagnostic workflow:** The two-stage process of lesion identification followed by pattern synthesis faithfully reproduces clinical reasoning, making the system intuitive for medical professionals and facilitating seamless integration into clinical practice.

By establishing this formal theoretical foundation that mirrors clinical diagnostic logic, LTH advances beyond architectural innovations to provide a principled approach for developing AI systems that emulate clinical reasoning—systematically observing individual pathological features before synthesizing them into diagnostic conclusions.

3 APPROACH

Our research employs the Label Transfer Hypothesis framework, which enables disease-level classification from pixel-level lesion annotations through knowledge transfer. This leverages the intrinsic mapping between lesion patterns and disease types—specific pathological combinations correspond to distinct disease categories.

This approach mirrors clinical diagnosis, where ophthalmologists identify diseases by recognizing pathological manifestations (hemorrhages, exudates, neovascularization), providing clinical validation for our methodology.

We developed a three-stage diagnostic pipeline: (1) pathological segmentation generates pixel-level lesion masks from fundus images; (2) feature extraction transforms segmentation results into clinically-relevant representations; (3) pattern recognition performs disease classification based on extracted features, achieving accurate diagnosis (as illustrated in Figure 1).

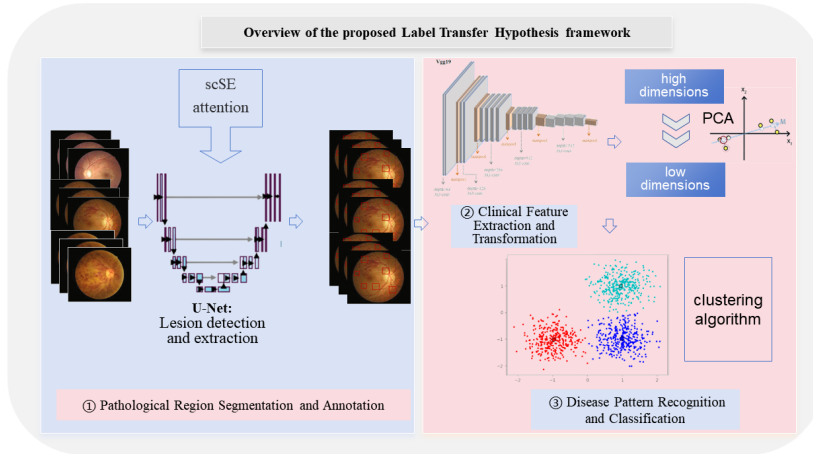


Figure 1: Overview of the proposed Label Transfer Hypothesis framework architecture

3.1 PATHOLOGICAL REGION SEGMENTATION

Precise segmentation of pathological regions constitutes the theoretical foundation for label transfer, as accurate identification and localization of various lesions are essential for providing reliable “implicit labels” for subsequent disease classification. From a clinical perspective, ophthalmologists initially identify various pathological manifestations in fundus images, each exhibiting distinct morphological, chromatic, luminance, and textural characteristics—hemorrhages present as dark red patches, hard exudates manifest as yellowish-white punctate lesions, while neovascularization appears as filamentous structures.

To emulate this clinical recognition process, we employ a modified U-Net (Ronneberger et al., 2015) architecture integrated with attention mechanisms (Figure 2). Through systematic evaluation of backbone networks including ResNet (He et al., 2016), VGG (Simonyan & Zisserman, 2014), and DenseNet (Huang et al., 2017), we identified optimal feature extraction architectures for each lesion type. The specific backbone selections and their performance metrics are detailed in Table 4 in Appendix A.1.

The decoder incorporates the squeeze-and-excitation (scSE) attention mechanism (Roy et al., 2018), which adaptively weights both channel and spatial dimensions to simulate clinicians’ cognitive process of focusing attention on pathological regions, effectively enhancing segmentation accuracy and clinical relevance.

Considering the typically small proportion of lesion regions in fundus images, we designed a composite loss function combining Dice Loss (Li et al., 2019) and Focal Loss (Lin et al., 2017) to address class imbalance challenges. The detailed loss function formulation and hyperparameter settings are provided in Appendix A.4.

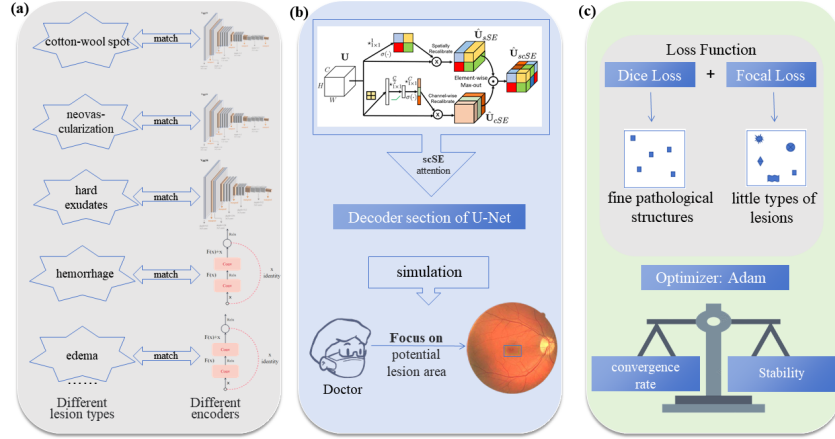


Figure 2: Architecture of the Pathological Region Segmentation and Annotation Stage. (a) Lesion-specific encoder selection strategy. (b) U-Net decoder with scSE attention mechanism. (c) Composite loss function design combining Dice Loss and Focal Loss.

3.2 CLINICAL FEATURE EXTRACTION AND TRANSFORMATION

Feature extraction transforms low-level image features into high-level clinical concepts, serving as the crucial bridge between image processing and clinical diagnosis. From a clinical perspective, different lesion types possess varying diagnostic significance—in diabetic retinopathy diagnosis, neovascularization often indicates progression to the proliferative stage, while the quantity of microaneurysms and hemorrhages reflects disease severity.

Based on this clinical understanding, we utilize a pre-trained VGG19 network as the feature extractor, transforming input images into $512 \times 7 \times 7$ feature maps corresponding to 25,088-dimensional feature vectors. We apply importance-weighted aggregation to the extracted features, with weight

combinations optimized through grid search. The specific optimization process and resulting weight values are detailed in Appendix A.4.

Finally, Principal Component Analysis (PCA) (Abdi & Williams, 2010) reduces the high-dimensional features to 50 dimensions, preserving critical diagnostic information while improving computational efficiency (Figure 3).

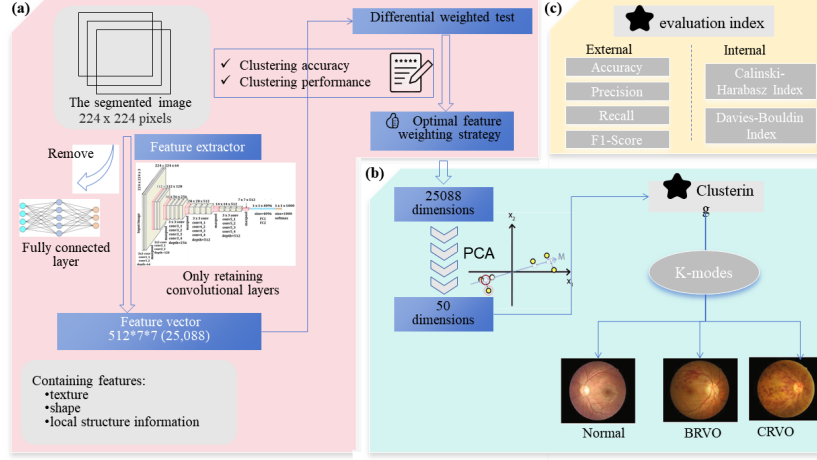


Figure 3: Architecture of the Clinical Feature Extraction and Disease Classification Stages. (a) Feature extraction with importance-weighted aggregation. (b) PCA dimensionality reduction and clustering. (c) Multi-dimensional clustering evaluation metrics.

3.3 DISEASE PATTERN RECOGNITION AND CLASSIFICATION

Disease pattern recognition derives from a fundamental assumption in medical image analysis: similar lesion patterns typically correspond to similar disease types, enabling unsupervised classification based on lesion features. From a clinical perspective, physicians identify disease types by recognizing specific lesion combination patterns—the co-occurrence of microaneurysms, hemorrhages, and hard exudates commonly indicates diabetic retinopathy, while geographic atrophy and drusen suggest age-related macular degeneration.

To implement this pattern recognition, considering the discrete distribution characteristics of fundus lesion features, we employ the k-modes (Huang & Ng, 1999) clustering algorithm, which updates cluster centers based on frequency, making it more suitable for categorical or mixed-feature data characteristic of medical image features. The detailed clustering algorithm parameters and evaluation metrics are provided in Appendix A.4.

For clustering effectiveness evaluation, we adopt a multi-dimensional assessment framework combining external metrics (accuracy, precision, recall, F1 score) with internal cluster quality metrics (Calinski-Harabasz index (Caliński & Harabasz, 1974) and Davies-Bouldin index (Davies & Bouldin, 2009)).

3.4 SYSTEM INTEGRATION AND CLINICAL WORKFLOW ALIGNMENT

The three stages operate synergistically to implement the Label Transfer Hypothesis paradigm: pathological region segmentation provides structured lesion representations, clinical feature transformation incorporates medical expertise through importance weighting, and disease pattern recognition demonstrates unsupervised diagnostic capability. This architecture realizes knowledge transfer by exploiting the intrinsic relationship between lesion patterns and disease manifestations, guided by established clinical diagnostic criteria. The system design aligns with clinical diagnostic workflows—physicians systematically identify characteristic lesions, evaluate their clinical significance, and make diagnostic decisions based on pattern recognition, thereby achieving both high diagnostic accuracy and enhanced clinical interpretability. This design ensures the entire diagnostic process re-

mains transparent and comprehensible to clinicians, facilitating the establishment of human-machine collaborative diagnostic paradigms.

Table 1: Overview of Publicly Available Medical Image Datasets Used in This Study

Dataset	Samples	Classes/Task
LCFP-14M	13,718,610	10 Classes
IDRiD Dataset	506	5 Classes
Retinal Lesions Dataset	198	Segmentation
Retinal Vessel Segmentation Combined	104	Segmentation
REFUGE2	1,200	Segmentation
Final Clean Haemorrhage Dataset	3,483	Segmentation

4 EXPERIMENTS

4.1 DATASETS AND EXPERIMENTAL SETUP

To evaluate the efficacy and generalizability of our proposed Label Transfer Hypothesis, we assembled three complementary datasets encompassing the spectrum of fundus pathology analysis tasks.

Diabetic Macular Edema (DME) Classification Dataset: This dataset comprises 302 high-resolution retinal fundus images from six public repositories (Rocamora, 2023; Scientist, 2022; Lemos, 2024; Patel, 2020; SHIROUQSHAWKY16, 2024; Decenci re et al., 2014) (Table 1). Each image features pixel-level annotations for microhemorrhages, macular edema, cotton wool spots, vascular structures, and lipid exudates. The dataset maintains balanced class distribution with 151 DME-positive and 151 DME-negative cases.

Retinal Vein Occlusion (RVO) Classification Dataset: The RVO dataset contains 374 fundus images from the aforementioned six databases (Table 1), comprising 187 RVO-positive cases and 187 controls.

Multi-lesion Segmentation Dataset: We constructed pixel-level ground-truth annotations by integrating multiple public sources. This dataset provides binary masks for hemorrhagic lesions, edematous regions, lipid deposits, neovascular formations, cotton wool spots, and other retinal pathologies essential for segmentation model training.

Experimental Setup: All experiments were conducted on a distributed GPU cluster. The detailed hardware configuration and training hyperparameters are provided in Appendix A.4.

4.2 COMPARATIVE EXPERIMENTS

To rigorously assess the efficacy and clinical viability of our proposed Label Transfer Hypothesis paradigm, we conducted systematic benchmarking experiments against established supervised learning methodologies.

Table 2: Comparative performance analysis of supervised and unsupervised approaches for DME and RVO classification

Task	Method	Accuracy	Precision	Recall	F1-Score
RVO	Supervised three grading	0.5241	0.5065	0.4840	0.4949
	Supervised binary grading	0.8056	0.8837	0.7037	0.7835
	Ours	0.8621	0.8234	0.8621	0.8423
DME	MobileNet	0.7903	0.8214	0.7419	0.7797
	ResNet18	0.7581	0.9000	0.5806	0.7058
	Ours	0.8600	0.8606	0.8600	0.8603

Experimental results demonstrate that our Label Transfer Hypothesis framework achieves remarkably competitive diagnostic performance across both experimental domains through disease-specific

model configurations tailored to distinct pathological characteristics. For DME classification, our framework employs GoogLeNet (Szegedy et al., 2015) as the feature extraction backbone combined with k-medoids clustering, achieving 86.00% accuracy, while RVO subtype differentiation utilizes VGG19 paired with k-modes clustering, yielding 86.21% accuracy. These results, presented in Table 2, not only rival but occasionally exceed those of traditional supervised approaches across multiple evaluation criteria. This performance parity is particularly noteworthy given the fundamental paradigmatic difference: our framework operates entirely without disease-level annotations, instead leveraging intrinsic pathological feature representations derived from lesion segmentation. The disease-specific model selection strategy reflects the heterogeneous nature of retinal pathologies and demonstrates the framework’s adaptability to diverse clinical scenarios while maintaining consistently high diagnostic performance.

The clinical significance of these findings extends beyond mere performance metrics. By achieving diagnostic accuracy comparable to supervised methods while circumventing the requirement for extensive disease-level annotation, our Label Transfer Hypothesis framework addresses a critical bottleneck in medical AI deployment—the dependency on labor-intensive expert labeling.

The comprehensive results of comparative analysis between various backbone networks and clustering algorithms for both DME and RVO classification are presented in Tables 5 and 6 in Appendix A.1.

4.3 ABLATION STUDIES

To systematically validate the contribution of each component within our proposed Label Transfer Hypothesis framework, we conducted comprehensive ablation experiments across three distinct configurations: (1) Classification-only baseline: Direct classification on raw fundus images without segmentation or feature extraction; (2) Feature extraction and classification without segmentation: Deep feature extraction using pre-trained networks followed by classification, bypassing the region segmentation module; and (3) Complete framework incorporating all components (Region Segmentation, Feature Extraction, Classification).

Table 3: Component-wise Evaluation of the Proposed Framework for DME and RVO

Task	Method			Metrics					
	A	B	C	Accuracy	Precision	Recall	F1-Score	CH index	DB index
DME	✓			0.5430	0.5430	0.5371	0.5225	2359.7505	0.3350
		✓	✓	0.8400	0.8400	0.8422	0.8397	29.7464	2.0853
	✓	✓	✓	0.8600	0.8606	0.8600	0.8603	36.3945	1.8753
RVO	✓			0.4947	0.4947	0.1658	0.2483	0.2310	1.9950
		✓	✓	0.5241	0.5065	0.4840	0.4949	327.3534	0.8380
	✓	✓	✓	0.8621	0.8234	0.8621	0.8423	6.2923	2.4017

Note: A = Region Segmentation; B = Feature Extraction; C = Classification; CH index = Calinski-Harabasz index; DB index = Davies-Bouldin index.

The experimental results, summarized in Table 3, demonstrate the critical importance of each framework component. For DME classification, the classification-only baseline achieved merely 54.30% accuracy, highlighting the inadequacy of raw image classification for medical diagnosis. The integration of deep feature extraction substantially improved performance to 84.00% accuracy, confirming the necessity of semantic feature representations. However, our complete framework incorporating lesion segmentation achieved optimal performance at 86.00% accuracy, validating our core hypothesis that explicit pathological localization enhances diagnostic capability.

Similar performance trends were observed in RVO classification, where the classification-only approach yielded suboptimal performance (49.47% accuracy), while feature extraction provided significant improvement (52.41% accuracy). The complete framework achieved 86.21% accuracy, representing a substantial performance gain of over 36 percentage points. Notably, the removal of attention mechanisms resulted in comparable accuracy but inferior feature discriminability metrics (Calinski-Harabasz and Davies-Bouldin indices), indicating that attention mechanisms enhance both feature quality and classification coherence.

These ablation results conclusively demonstrate that each component contributes synergistically to overall framework performance, with the lesion segmentation module serving as the most critical element for effective knowledge transfer from pixel-level annotations to image-level diagnosis.

5 CONCLUSION

This study presents and validates the Label Transfer Hypothesis, a framework that tackles the annotation shortage problem in medical imaging through knowledge transfer from lesion-level to disease-level classification. Evaluation on Diabetic Macular Edema (DME) and Retinal Vein Occlusion (RVO) demonstrates theoretical validity and clinical applicability.

The Label Transfer Hypothesis redefines supervision in medical imaging by utilizing pathological patterns from clinically-guided lesion segmentation instead of requiring disease-level annotations. This approach overcomes annotation limitations while integrating computational methods with clinical diagnostic procedures, improving model interpretability and clinical utility.

Results show classification accuracies of 86.00% for DME detection and 86.21% for RVO subtype classification without disease-level supervision. These findings confirm that lesion-level annotations provide sufficient discriminative information for accurate disease classification, demonstrating that structured pathological knowledge can replace explicit disease labels.

This work provides a foundation for clinically-guided supervision methods across medical imaging applications. By establishing knowledge transfer from pixel-level pathological annotations to image-level classification, we present a generalizable framework that combines clinical domain knowledge with machine learning, enabling interpretable and annotation-efficient medical AI systems compatible with clinical practice.

REFERENCES

- H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.
- T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.
- Q. Chen et al. Artificial intelligence can assist with diagnosing retinal vein occlusion. *International Journal of Ophthalmology*, 14(12):1895–1902, 2021.
- D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):224–227, 2009.
- Etienne Decencière et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pp. 231–234, 2014.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2016.
- Kelei He et al. Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78, 2023.
- Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2017.
- Z. Huang and M. K. Ng. A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, 7(4):446–452, 1999.
- S. Kumari and P. Singh. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Computers in Biology and Medicine*, 170:107912, 2024.
- Victor Lemos. Refuge2. <https://www.kaggle.com/datasets/victorlemosml/refuge2>, 2024.
- X. Li et al. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019.

- T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2017.
- S. Matta et al. Meta learning for anomaly detection in fundus photographs, 2022.
- Aarya Patel. Indian diabetic retinopathy image dataset. <https://www.kaggle.com/datasets/aaryapatel98/indian-diabetic-retinopathy-image-dataset>, 2020.
- Pablo Rocamora. Retinal lesions dataset. <https://www.kaggle.com/datasets/pablorocamora/retinal-lesions-dataset>, 2023.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- A. G. Roy, N. Navab, and C. Wachinger. Concurrent spatial and channel squeeze & excitation in fully convolutional networks, 2018.
- Noob Data Scientist. Retinal vessel segmentation combined. <https://www.kaggle.com/datasets/pradosh123/retinal-vessel-segmentation-combined/data>, 2022.
- SHIROUQSHAWKY16. final clean haemorrhage dataset. <https://www.kaggle.com/datasets/shirouqshawky16/final-clean-haemorrhage-dataset?select=labels>, 2024.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Christian Szegedy et al. Going deeper with convolutions, 2015.
- S. Yousefi et al. Recognizing patterns of visual field loss using unsupervised machine learning, 2014.

A APPENDIX

A.1 DETAILED SEGMENTATION PERFORMANCE

Table 4 presents the detailed segmentation performance metrics for different lesion types using various backbone networks. Through systematic evaluation, we identified optimal backbone architectures for each lesion type, achieving the best balance between segmentation accuracy and computational efficiency.

Table 4: Optimal backbone network selection and segmentation performance for different lesion types

Foci	Backbone	IoU	Accuracy	Precision
Hemorrhage	ResNet34	0.4942	0.9956	0.8008
Edema	ResNet50	0.9925	0.9867	0.9988
Neovascularization	VGG19	0.7415	0.9692	0.8921
Hard exudation	VGG16	0.8719	0.9968	0.9625
Cotton-Wool-Spots	VGG19	0.3636	0.9920	0.5671

A.2 BACKBONE NETWORK COMPARISON FOR DME CLASSIFICATION

Table 5: Performance comparison across different backbone networks and clustering algorithms for DME classification

Clustering	Backbone	Accuracy	Precision	Recall	F1-Score	Calinski-Harabasz	Davies-Bouldin
k-medoids	AlexNet	0.8333	0.8437	0.8333	0.8321	15.9312	2.2379
	GoogLeNet	0.8600	0.8606	0.8600	0.8603	36.3945	1.8753
	ResNet18	0.8333	0.8437	0.8333	0.8321	161.6702	0.7050
	VGG16	0.8133	0.8283	0.8133	0.8112	12.4736	3.1007
Kmeans	GoogLeNet	0.8000	0.8054	0.8000	0.7991	267.8952	0.5985
	ResNet34	0.8333	0.8437	0.8333	0.8321	19.4077	2.6955
	VGG16	0.7667	0.7749	0.7667	0.7649	12.0028	3.1943
	VGG19	0.8067	0.8134	0.8067	0.8056	14.0618	2.7996
Agg	AlexNet	0.8267	0.5630	0.5511	0.5539	14.7440	2.1688
	GoogLeNet	0.8400	0.5742	0.5600	0.5626	16.1686	2.1582
	ResNet18	0.7267	0.5265	0.4844	0.4982	71.7643	1.5988
	VGG16	0.8533	0.5780	0.5689	0.5728	11.0147	2.2108

A.3 BACKBONE NETWORK COMPARISON FOR RVO CLASSIFICATION

Table 6: Performance comparison across different backbone networks and clustering algorithms for RVO classification

Clustering	Backbone	Accuracy	Precision	Recall	F1 Score	Calinski-Harabasz	Davies-Bouldin
Agglomerative	AlexNet	0.8483	0.8805	0.6229	0.6280	10.5400	2.0712
	VGG16	0.8759	0.6844	0.6329	0.6374	9.6975	2.0309
Kmeans	VGG19	0.8759	0.6844	0.6329	0.6374	9.6975	2.3316
GMM	VGG19	0.8897	0.9085	0.6531	0.6571	9.0810	2.1721
Kmedoid	VGG19	0.8552	0.8149	0.8552	0.8312	7.0336	2.0823
Kmodes	VGG16	0.8552	0.8829	0.8552	0.8384	7.4719	2.0309
	VGG19	0.8621	0.8234	0.8621	0.8423	6.2923	2.4017

A.4 TRAINING PARAMETERS AND IMPLEMENTATION DETAILS

A.4.1 HARDWARE CONFIGURATION

All experiments were conducted on a distributed GPU cluster with the following configuration:

- 2× NVIDIA RTX 3080 (24GB VRAM)
- 1× NVIDIA RTX 2070 SUPER (8GB VRAM)
- 1× NVIDIA RTX 2080 Ti (11GB VRAM)
- 1× NVIDIA GTX 1080 (8GB VRAM)

A.4.2 SEGMENTATION MODEL TRAINING PARAMETERS

The U-Net segmentation models were trained with the following hyperparameters:

- **Optimizer:** Adam with initial learning rate of 1e-4
- **Learning rate scheduler:** ReduceLROnPlateau with patience=10, factor=0.5
- **Batch size:** 8 for RTX 3080, 4 for other GPUs
- **Epochs:** 100 with early stopping (patience=20)
- **Data augmentation:** Random horizontal/vertical flips, rotation ($\pm 30^\circ$), brightness/contrast adjustment

- **Loss function weights:** Dice Loss weight = 0.5, Focal Loss weight = 0.5
- **Focal Loss parameters:** $\alpha = 0.25$, $\gamma = 2.0$

A.4.3 FEATURE EXTRACTION PARAMETERS

The feature extraction stage employs the following configuration:

- **Pre-trained model:** VGG19 trained on ImageNet
- **Input image size:** 224×224 pixels
- **Feature map dimensions:** $512 \times 7 \times 7$
- **PCA components:** 50
- **Lesion importance weights** (optimized via grid search):
 - Hemorrhage: 0.25
 - Edema: 0.30
 - Neovascularization: 0.20
 - Hard exudates: 0.15
 - Cotton-wool spots: 0.10

A.4.4 CLUSTERING ALGORITHM PARAMETERS

The clustering algorithms were configured as follows:

- **K-modes/K-medoids:** k=2 for binary classification
- **Initialization:** k-means++ for numerical features
- **Maximum iterations:** 300
- **Convergence tolerance:** $1e-4$
- **Number of runs:** 10 with different random seeds

A.4.5 LOSS FUNCTION FORMULATION

The composite loss function for segmentation is defined as:

$$\mathcal{L}_{\text{combined}} = \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{focal}} \quad (1)$$

where:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2 \cdot |X \cap Y|}{|X| + |Y|} \quad (2)$$

$$\mathcal{L}_{\text{focal}} = -\alpha_t \cdot (1 - p_t)^\gamma \cdot \log(p_t) \quad (3)$$

with $\alpha_t = 0.25$ for lesion pixels and 0.75 for background pixels, and $\gamma = 2.0$ to focus learning on hard examples.