G-014

Metabolic Syndrome Diagnosis Using Fundus Images Without Clinical Data

シュシェン† 鎌田清一郎†
Sheng Xu Sei-ichiro Kamata

1. Abstract

Metabolic syndrome (MetS) is a growing global health concern, traditionally diagnosed through invasive and resource-intensive clinical measurements. To address the need for non-invasive diagnosis, we propose a two-stage deep learning framework for MetS classification using only fundus images. In the first stage, a RETFound-based model predicts key clinical risk factors directly from each fundus image. In the second stage, image features and these predicted clinical features are integrated using feature-wise linear modulation (FiLM) and an adaptive fusion mechanism to classify MetS. To improve robustness, we employ scheduled sampling during training, gradually shifting from ground-truth clinical inputs to the model ' s own predictions. Evaluated on an independent test set from the Japan Ocular Imaging Registry, our method achieved an accuracy of 81.4% and an area under the ROC curve (AUC) of 0.836, outperforming models that use only image or clinical data. These results demonstrate the feasibility of non-invasive MetS detection via retinal imaging and highlight the effectiveness of integrating learned clinical representations to improve systemic disease screening.

2. Introduction

Metabolic syndrome (MetS) is a cluster of interrelated risk factors－such as central obesity, hypertension, hyperglycemia, high triglycerides, and low HDL cholesterol－that affects approximately 20‐25% of adults worldwide, a prevalence that continues to rise[2]-[4]. Individuals with MetS face a markedly elevated risk of developing type2 diabetes (around five-fold higher) and cardiovascular disease (roughly two-fold higher) compared to healthy individuals[2][3]. Early diagnosis is critical for prevention; however, MetS is traditionally diagnosed through multiple clinical measurements (e.g., blood tests, blood pressure readings, waist circumference) that are invasive, time-consuming, and resource-intensive[2]. This clinical burden motivates the search for non-invasive screening approaches for MetS. One promising non-invasive approach is retinal fundus imaging, which has long been used in ophthalmology (for example, to screen for diabetic retinopathy) and provides a unique window into systemic vascular health[1][5]. The retina is the only internal tissue visible non-invasively, and its microvascular features reflect systemic conditions associated with MetS[1][5]. Individuals with MetS, regardless of age, sex, or ethnicity, have an increased risk of retinal lesions, arteriovenous nicking, focal arteriolar narrowing, decreased retinal arteriolar diameter, and increased venular diameter[1][5].

† Sheng Xu and Sei-ichiro Kamata are with the Image Media Laboratory, Graduate School of Information, Production and Systems, Waseda University, Kitakyushu 8080135, Japan.

In recent years, the growing availability of large retinal image datasets and advances in deep learning have expanded fundus imaging applications beyond ocular disease. For example, Poplin et al. demonstrated that a single fundus photograph can predict multiple cardiovascular risk factors (e.g., blood pressure, smoking status) as well as future cardiac events[6]. These findings underscore the potential of fundus images as non-invasive biomarkers of systemic disease.

Recent work by Lee et al. proposed a non-invasive approach for MetS diagnosis by combining fundus images with basic clinical information such as age, gender and BMI[7]. However, their method does not incorporate metabolic indicators like blood pressure or blood glucose, which are key components of international MetS diagnostic criteria[2]. This omission may limit the method ' s ability to fully capture the clinical complexity of MetS. Therefore, we propose a new framework that leverages clinically validated metabolic features－including blood pressure and blood glucose－as support during training, while requiring only fundus images as input during inference. By doing so, our approach aims to enable truly non-invasive MetS screening, grounded in established diagnostic standards, and to evaluate the feasibility of fundus-image‐based MetS diagnosis using deep learning.

3. Related Work

3.1 Fundus Image‐Based MetS Diagnosis

Research directly using retinal fundus images for MetS diagnosis is a relatively new field. Lee et al. recently demonstrated that fundus images can be used to detect MetS, with diagnostic performance further improved by incorporating basic clinical variables such as age, sex, and BMI (AUC ≈ 0.87)[7]. This finding suggests that retinal photographs encode systemic information relevant to metabolic risk. However, achieving high diagnostic accuracy using only fundus images－without any additional clinical inputs－remains a significant challenge and an open area for further research.

3.2 Fundus for MetS-Associated Diseases Diagnosis

MetS shares pathophysiological features with conditions such as type 2 diabetes, hypertension, and dyslipidemia, all of which have been extensively studied using retinal fundus imaging. Numerous studies have demonstrated that deep learning models can accurately identify diabetes from fundus photographs, even in the absence of overt retinal disease, achieving AUCs as high as 0.88 for hyperglycemia prediction[8][9]. Similarly, hypertension and dyslipidemia have been detected from fundus images with AUCs of 0.77 and 0.70, respectively[10][11]. These results confirm that the individual components of MetS−elevated blood glucose, blood pressure, and cholesterol−manifest as subtle, yet discernible, retinal features. Beyond individual risk factors, fundus-based algorithms have also been used to estimate

composite cardiovascular risk, such as predicting a patient’s 10-year risk of myocardial infarction or stroke at a level comparable to traditional risk scores[10]. Furthermore, deep learning has enabled high-accuracy detection of metabolic complications, including diabetic retinopathy and glaucoma, directly from retinal images[8][9]. Collectively, these advances underscore the potential of fundus imaging as a non-invasive tool for systemic disease assessment and provide a strong foundation for investigating MetS diagnosis through ocular biomarkers.

3.3 MetS Diagnosis Using Only Clinical Data

Recent studies have demonstrated the utility of clinical features in predicting systemic conditions such as hypertension and metabolic risk[12]. These approaches often incorporate demographic and anthropometric variables like age, sex, BMI, and blood pressure, leveraging statistical or deep learning models for disease risk stratification. Such findings confirm the clinical relevance of structured health data in non-invasive diagnostics.

In contrast to these works, our study involves a richer set of clinically validated indicators－such as fasting glucose, lipid profile, and waist circumference−aligned with international MetS diagnostic criteria. These features serve as targets during model training, enabling our system to learn medically meaningful surrogate markers from fundus images.

4. Proposed Method

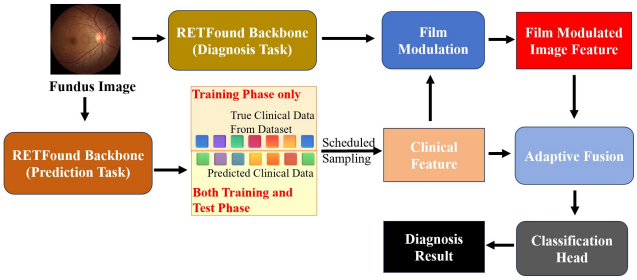

Figure1: Overview of the proposed two-stage pipeline for MetS diagnosis from fundus images. Stage 1 uses a RETFound backbone (Prediction Task), fine-tuned to regress clinical data from fundus images. Stage 2 employs a separately fine-tuned RETFound[13] model (Diagnosis Task) to extract image features, which are modulated via FiLM using clinical data (ground-truth during training, predicted during inference). The modulated features are adaptively fused and passed to a classification head for final MetS prediction. Scheduled sampling is applied during training to gradually replace ground-truth clinical inputs with predicted ones.

Our proposed approach is a two-stage deep learning pipeline that diagnoses MetS from fundus images without clinical data at test phase. As illustrated in Figure 1, the first stage predicts key clinical data from a fundus image, and the second stage classifies metabolic syndrome by fusing fundus features with the predicted clinical values. Both stages leverage a Vision Transformer backbone. Specifically, we adopt RETFound[13]－a recently developed ViT-based foundation model pretrained on ~1.6 million retinal images−to extract informative features. RETFound’s large-scale pretraining enables robust and generalizable retinal representations, which we further fine-tune for each stage: the prediction task uses a RETFound model specialized for clinical data regression, while the diagnosis task employs a separately fine-tuned RETFound model for MetS classification.By performing task-specific fine-tuning of RETFound in each stage, our architecture fully leverages the model’s powerful retinal feature extraction capabilities, enabling it to effectively address the distinct objectives of clinical biomarker prediction and MetS diagnosis.

4.1 Prediction Task: Clinical Feature Regression

In the first stage, the model predicts a set of clinical data from a single fundus image. The image is processed by a RETFound backbone, which is fine-tuned for this regression task, followed by a shallow MLP head that outputs predicted values for key metabolic syndrome‐related parameters. The network is trained end-to-end by minimizing mean squared error between the predicted and true clinical data, encouraging the model to encode fundus features into quantitative systemic health markers.

4.2 Diagnosis Task: Fused Feature Classification

In the second stage, a separate model classifies metabolic syndrome using both the fundus image and clinical features (real values during training; predicted values at inference). A RETFound backbone extracts image features, while clinical data are projected into an embedding of matching dimension. The two streams are integrated via feature-wise linear modulation (FiLM), allowing clinical information to modulate fundus features. An adaptive fusion mechanism then weights and combines the modulated image and clinical embeddings, producing a fused representation for final classification. The network is trained with cross-entropy loss to predict MetS status. This architecture enables flexible integration of both real and predicted clinical data, allowing training on ground truth while supporting fully image-based inference.

4.3 Joint Training with Scheduled Sampling

We train the two stages in a joint fashion, optimizing both the regression and classification objectives together. The overall training loss L is a sum of the prediction MSE loss and the diagnosis classification loss:

$$L = L_{MSE} + L_{CE}$$

This multi-task learning framework encourages the predictor to generate clinical features that are both accurate and informative for downstream classification, while the classifier learns to leverage both image and clinical cues. To align training and inference conditions, we adopt scheduled sampling: the classifier initially receives ground-truth clinical data (“teacher forcing”) to stabilize learning and provide an upper performance bound. As training progresses, real clinical inputs are gradually replaced with predicted values from the first stage, following a predefined schedule. By the end of training, the classifier operates exclusively on the predictor’s outputs, ensuring robustness to prediction noise and mitigating the gap between training and deployment. Both stages are jointly

optimized via backpropagation, with the predictor minimizing MSE and the classifier minimizing cross-entropy loss as the input distribution evolves.

4.4 Inference Pipeline

During inference or deployment, our pipeline requires only the fundus image as input. First, the trained predictor network processes the 224 × 224 fundus image to output the estimated clinical feature vector. These predicted clinical features are then fed, along with the same fundus image, into the diagnosis network ’ s RETFound backbone and fusion modules as described. The classifier consequently produces the final diagnosis of metabolic syndrome (positive or negative). All clinical data used by the classifier at test time come from the model ’ s own predictions, eliminating the need for any actual clinical measurements from the patient. In summary, the system can identify individuals with metabolic syndrome using fundus images only, by internally generating and leveraging surrogate clinical biomarkers. This two-stage design, combined with adaptive feature fusion and scheduled sampling during training, enables a clear and structured exploitation of fundus images for metabolic syndrome diagnosis without clinical data.

5. Experiments

5.1 Dataset & Experimental Setup

Dataset Description: We conducted experiments using the Japan Ocular Imaging Registry (JOIR) dataset[14], which includes 5,000 labeled retinal fundus images, each with associated clinical data for training and validation. All images were obtained from Japanese male health checkup participants. Each subject's metabolic syndrome status was annotated as a binary label (MetS-positive or MetS-negative), with the dataset balanced between positive and negative cases. An additional 500 fundus images (without access to labels or clinical data during training) were held out as an independent test set for final evaluation.

The available clinical data include seven routinely measured metabolic indicators: age at the time of examination, abdominal circumference (AC), systolic blood pressure (SBP), diastolic blood pressure (DBP), HDL cholesterol (HDLC), triglycerides (TG), and blood glucose (BS). These variables align with standard diagnostic criteria for metabolic syndrome and were used as supervision targets in the prediction stage of our model.

Preprocessing: Fundus images were automatically cropped to remove peripheral black borders and resized to 224×224 pixels. Clinical data were standardized using z-score normalization to ensure consistent scale across variables. These steps ensured better data quality for multimodal learning.

Evaluation Metrics: Model performance was primarily assessed using classification accuracy and the Area Under the ROC Curve (AUC). Accuracy measures the fraction of correct MetS classifications, while AUC evaluates the model’s ability to discriminate between MetS-positive and MetS-negative cases across all classification thresholds. We report both metrics on the validation folds during cross-validation and on the held-out test set, following standard practices in medical deep learning work. These metrics provide a balanced view of model performance, with AUC being particularly important for imbalanced-risk scenarios as it is threshold-independent.

5.2 Implementation Details

We employed 5-fold cross-validation on the 5,000 labeled samples. In each fold, model training proceeded in three stages: (1) a predictor network was trained to regress clinical features from fundus images; (2) a classifier was trained using image features and clinical inputs to diagnose MetS; (3) the entire pipeline was joint trained end-to-end. We used the Adam optimizer with an initial learning rate of 1e-4 and a batch size of 16. Training was performed on a system with an NVIDIA RTX 3090 GPU and an Intel Core i7-9300K CPU. Early stopping was applied based on validation Acc, and the best-performing model in each fold was used for evaluation.

5.3 Comparison

Features	Method	Accuracy	AUC
Fundus image, Age, Gender, BMI	Lee, et al.[7]	0.7837	**0.8725**
Age, Gender, BMI	Lee, et al[7]	0.7747	0.8640
Clinical Data(AC, SBP, DBP...)	Lee, et al[7]	0.7936	0.8154
Fundus Image only	RETFound[13]	0.706	0.748
Fundus Image + predicted clinical data(AC, SBP, DBP...)	Ours	**0.814**	0.836

Table1:Comparison of accuracy and AUC for different Feartures and models. The data in the first two rows are from the paper.

Table 1 contrasts MetS classification performance across modalities. The first two rows reproduce Lee et al.’s reported results: a RETFound-based ViT model using fundus images plus age/gender/BMI attained 78.37% accuracy (AUC 0.8725), while using only age/gender/BMI gave 77.47% accuracy (AUC 0.8640). The third row shows our reproduction of the clinical-data-only classifier (L1-regularized logistic regression classifier (C=0.03, solver=saga) on the extracted features) on our dataset’s clinical measurements, yielding 79.36% accuracy (AUC 0.8154). The final two rows report our fundus-only baseline and our proposed model. Our model (fundus image plus predicted clinical features) achieves 81.4% accuracy (AUC 0.836), the highest among all single-modality baselines and exceeding the clinical-only result. Although our model achieves the highest accuracy among all single-modality baselines, its AUC remains slightly lower than that of models combining fundus images with true clinical measurements (AUC ≈0.8725). Nonetheless, the proposed approach demonstrates clear improvements over the fundus-only baseline, indicating that incorporating predicted clinical cues enhances classification performance. These findings suggest that while our architecture effectively boosts accuracy by leveraging

surrogate clinical features, there is still room for improvement in terms of overall discriminative power as reflected by the AUC.

5.4 Ablation Study

Modal	Accuracy
DenseNet121+ResNet50	0.692
RETFound	0.758
RETFound+FILM	0.796
RETFound+FILM+Adaptive fusion	0.814

Table2:Ablation study results on JOIR dataset

Table2 presents an ablation analysis of key components in our model. Starting from a baseline dual-CNN architecture (DenseNet121 + ResNet50) achieving 0.692 accuracy, we observe stepwise improvements as each component is introduced. Replacing the baseline image backbone with the RETFound model (pretrained on retinal images) raises accuracy to 0.758. Incorporating FiLM modulation to inject the predicted clinical features further boosts accuracy to 0.796. Finally, adding the adaptive fusion mechanism to optimally combine image-based and feature-based cues yields the highest accuracy of 0.814.

This progression highlights the impact of each innovation: the retina-specialized RETFound backbone enhances feature learning, FiLM effectively integrates the predicted clinical information, and adaptive fusion provides an optimal multimodal combination. The final configuration (RETFound + FiLM + Adaptive Fusion) achieves the best performance at 81.4% accuracy (vs 69.2% for the baseline). This outcome demonstrates the cumulative contribution of all components to MetS classification.

6. Conclusion

In this study, we presented a two-stage deep learning framework for diagnosing metabolic syndrome (MetS) using only fundus images. The model first predicts key clinical data from fundus images using a RETFound-based network, and then integrates the predicted values with image features through FiLM modulation and adaptive fusion for final classification. Our method achieved 81.4% accuracy and 0.836 AUC on an independent test set, outperforming baseline approaches using either image or clinical data alone.

Despite promising results, our study is limited by the use of a male-only dataset, which may restrict generalizability. In future work, we plan to focus on optimizing the performance of the model and validating it on more diverse populations.

Acknowledgement

I would like to express my deepest gratitude to my parents for their unwavering financial support and enduring care, which have been the cornerstone of my academic journey. I am also sincerely thankful to my friends and classmates, whose companionship and encouragement have made my two years of studying abroad both joyful and fulfilling. Their presence has been a great comfort during my time away from home. Special thanks are reserved for Professor Sei-ichiro Kamata, whose insightful guidance and constructive advice have been invaluable to my research. His expertise and support have played a crucial role in shaping the direction and quality of this work.

References

[1] Lima-Fontes, Mário, et al. "Ocular findings in metabolic syndrome: a review." Porto biomedical journal 5.6 (2020): 104.
[2] Alberti, George, et al. "The IDF consensus worldwide definition of the metabolic syndrome." Brussels: International Diabetes Federation 23.5 (2006): 469-80.
[3] Grundy, Scott M. "Metabolic syndrome update." Trends in cardiovascular medicine 26.4 (2016): 364-373.
[4] Blüher, Matthias. "Obesity: global epidemiology and pathogenesis." Nature Reviews Endocrinology 15.5 (2019): 288-298.
[5] Liew, Gerald, et al. "Retinal vascular imaging: a new tool in microvascular disease research."Circulation: Cardiovascular Imaging 1.2 (2008): 156-161.
[6] Poplin, Ryan, et al. "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning." Nature biomedical engineering 2.3 (2018): 158-164.
[7] Lee, Tae Kwan, et al. "Vision transformer based interpretable metabolic syndrome classification using retinal Images." npj Digital Medicine 8.1 (2025): 205.
[8] Yang, Yehui, et al. "Robust collaborative learning of patch-level and image-level annotations for diabetic retinopathy grading from fundus image." IEEE Transactions on Cybernetics 52.11 (2021): 11407-11417.
[9] Rom, Yovel, et al. "Predicting the future development of diabetic retinopathy using a deep learning algorithm for the analysis of non-invasive retinal imaging." BMJ Open Ophthalmology 7.1 (2022): e001140.
[10] Lee, Yeong Chan, et al. "Multimodal deep learning of fundus abnormalities and traditional risk factors for cardiovascular risk prediction." npj Digital Medicine 6.1 (2023): 14.
[11] Baharoon, Mohammed, et al. "HyMNet: a Multimodal Deep Learning System for Hypertension Classification using Fundus Photographs and Cardiometabolic Risk Factors." arXiv preprint arXiv:2310.01099 (2023).
[12] Worachartcheewan, Apilak, et al. "Predicting metabolic syndrome using the random forest method."The Scientific World Journal 2015.1 (2015): 581501.
[13] Zhou, Yukun, et al. "A foundation model for generalizable disease detection from retinal images." Nature 622.7981 (2023): 156-163.
[14] Miyake, Masahiro, et al. "Japan ocular imaging registry: a national ophthalmology real-world database." Japanese Journal of Ophthalmology 66.6 (2022): 499-503.
[15] Mohseni-Takalloo, Sahar, et al. "Metabolic syndrome prediction using non-invasive and dietary parameters based on a support vector machine." Nutrition, Metabolism and Cardiovascular Diseases 34.1 (2024): 126-135.
[16] Choe, Eun Kyung, et al. "Metabolic syndrome prediction using machine learning models with genetic and clinical information from a nonobese healthy population." Genomics & informatics 16.4 (2018): e31.
[17] Wong, Tien Yin, et al. "Retinal microvascular abnormalities and their relationship with hypertension, cardiovascular disease, and mortality." Survey of ophthalmology 46.1 (2001): 59-80.