CVAE-based Causal Representation Learning from Retinal Fundus Images for Age Related Macular Degeneration(AMD) Prediction

Dae young, Kim

Abstract— **Regarding relatively poor prognosis and acute vision impairment, analyzing Age-Related Macular Degeneration, or AMD has been one of the most important tasks in retinal disease analysis. Especially, constructing methods to analyze and predict Wet AMD, which is characterized by rapid RPE damage due to neovascularization, has been a demanding task for many ophthalmologists for decades. Recently, with advancements in ML/DL frameworks and computer vision AI, these previous efforts are now leading to drastic enhancements in AMD prediction and mechanism analysis. Specifically, use of attention mechanism based CNNs or XAI methods are leading to higher performance in predicting AMD status and reliable explanations. Under current success in the usage of cutting-edge techniques, this research implemented a novel latent causal representation learning framework to further enhance AI-based models to comprehend complex causal AMD mechanisms with only access to retinal fundus images, while constructing a more reliable type of AMD prediction model. Results show that valid convolutional VAE and GAE based explicit latent causal modeling can lead to successful causal disentanglements of underlying AMD mechanisms, while returning essential causal factors that can be utilized to reliably distinguish normal fundus and AMD fundus images in downstream tasks such as diagnosis prediction.**

Key Words **— Age-related macular degeneration, Causal representation learning, Computer vision, Retinal disease, Neural Causal Model, VAE**

I. INTRODUCTION

Having relatively poor prognosis, Age-Related Macular Degeneration(AMD or ARMD) has been a significant threat to the human vision along with a variety of retinal diseases such as Glaucoma or Diabetic Retinopathy [1,2]. Especially among retinal damage, Wet AMD has been one of the most major causes of acute vision impairment, which rapidly leads to complete or permanent loss of vision. Aside from having severe prognosis, AMD is also getting substantial attention for its increase in scope. According to the CDC, while being the primary cause of permanent vision impairment in people who are aged 65 and older, it was found that approximately 18.34 million in the US who are aged 40 and older suffered from early-stage AMD in 2019 [3,4]. In global aspects, increase rate of AMD patients is expected to be 47% in 2040 (from 196 million in 2020 to 288 million in 2040) [5]. Due to its current severity and linkage to mental health [6], numerous researchers probed for methods to enhance the possibility of earlier AMD detection or analyzed mechanisms of AMD-based vision loss to find effective treatments. In traditional works, manual OCT/FA image interpretations to track optical components (such as RPE or photoreceptors) and biomarker-based risk evaluations were mainly implemented for AMD analysis [7-10]. However, due to the intrinsic complexity in AMD pathogenesis and a broad overlap between other dystrophic macular disorder features, accurate or timely diagnosis of AMD still remained as a demanding task [11].

In recent works, with advancements in machine learning and neural network analysis, the possibility of overcoming these limitations is being increased in various aspects. Under extreme advancements in artificial intelligence, ML/DL based modelling approaches show high performance in fulfilling previous objectives. Specifically, methodologies such as attention mechanism based Convolutional Neural Networks (CNN) for retinal image classification or XAI-based analysis for acute image classifications are repeatedly being implemented for DL-based AMD analysis, leading to substantial advancements in retinal disease recognition or disease prediction performance [12-14]. Furthermore, fast developments in deep generative AI have enabled researchers to generate synthetic retinal images, which led to precise recognition or comprehension regarding optical components.

Based on the success in numerous applications of DL/ML methods in retinal analysis, this study combined cutting edge techniques in deep learning with AMD analysis to probe for novel mechanisms that can lead to accurate AMD diagnosis and valid interventional analysis in the aspect of treatments under better reliability. To be specific, this study attempted to extract latent causal representations related to AMD diagnosis from patient's retinal fundus images using Convolutional VAE and

*Affiliation Information: The Department of Applied Statistics, Yonsei University, Seoul, South Korea.
*First Author, Corresponding Author. (Email: lgtlsafg@yonsei.ac.kr)
*Data available at: Kaggle Dataset Repository: "Retinal Disease Classification". (https://www.kaggle.com/datasets/andrewmvd/retinal-disease-classification)
*Relevant codes for analysis can be accessed through the following URL: GitHub Repository: (https://github.com/Daeyoung25-Kim/AMD_LatentCausalRepresentationLearning_BioRxiv)
*Date of Creation: (version II) May, 12, 2025.

Neural Causal DAG extraction. Unlike previous works, which only focused on correlation or classification performance, this research also focused on making computer vision models comprehend the causal mechanism of AMD and return valid inference regarding interventions (treatments), while achieving high prediction performance regarding AMD development.

II. Preliminaries

A. AMD mechanism as domain knowledge

To check success in causal disentanglements under latent causal representations extracted from retinal fundus images, a basic empirical causal graph, which can be deduced from domain knowledge of AMD development, was set as a benchmark. AMD diagnosis, or vision impairment due to AMD is a result of either two distinct processes: Dry AMD and Wet AMD [15,16]. Characterized by atrophy in RPE, Dry AMD is resulted from cumulation of drusen, an undisposed metabolic waste. With increased oxidative stress, proteins and lipids are abnormally deposited under the retina as specks. These specks, or drusen, disturb photoreceptors from receiving nourishment as it hinders interaction between blood vessels and the retina [16,17]. This not only leads to deterioration of RPE cells but also vision impairment as the size of drusen gets larger.

On the other hand, in Wet AMD, which is characterized by hemorrhage or intraretinal fluids within the retina, neovascularization is considered as a major cause. Due to structural changes in choroidal vasculature from abnormal vascular endothelial growth factor(VEGF) levels, angiogenesis or neovascularization occurs in Wet AMD, which leads to damage in RPE, bruch's membrane and photoreceptors. This leads to generation of intraretinal fluids and hemorrhage which may result in focal loss and vision impairment by damaging the macula itself [15]. Thus, AMD mechanisms can be summarized into a simple causal diagram as in Fig 1. (Though there were some experimental results where possibility of interactions between drusen and neovascularization was found [17,18], explicit causal links between two features are not yet identified). In this work, as exact pathologies of AMD are not established, causal graph in Fig 1 was considered as an alternative (domain-based) benchmark causal network when validating or evaluating extracted latent causal variables from retinal fundus images in experiments.

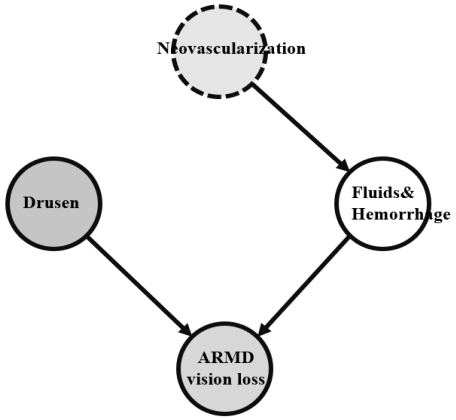


Fig 1. Visualization of Simplified AMD Causal Mechanism.

B. Related Works

With recent advancements in ML/DL architectures, various applications of AI in AMD analysis have been reported to be successful in diverse environments. Especially for OCT image analysis, CNN based modelling was found successful in most research. For example, using modified VGG16 CNN for 61-line raster macula scan OCT image classification, 92.77% AUROC and 87.63% binary AMD classification accuracy were achieved [12]. In [12], it was also found that pathologic regions can be retrieved from optical coherence tomography image based VGG16, which implies that key features in detecting AMD can be extracted through CNN architecture based deep learning approaches. Furthermore, implementing two CNN based deep residual neural networks: ResNet50 and attention mechanism based ResNet, AMD classification modelling was driven to achieve higher performance in detecting macular degeneration from OCT images, which led to higher classification accuracy when compared to HOG-SVM and VGG [13].

To make classification performance more robust to image preprocessing steps such as RoI extraction and retinal flattening, multi-level dual-attention based CNN approaches were also implemented for advancements in related works [19]. Under Duke and NEH datasets, multi-level DAM achieved higher performance, compared to existing methodologies such as multi-scale CNN ensemble or LACNNs. Apart from focusing on enhancing classification performance, increasing the interpretability of deep learning based AMD classification models was also studied by researchers. For example, under a finely trained VGG16 AMD classification model, gradient-weighted class activation mapping (GradCAM) was implemented as an XAI method to visualize major features that affected the decision of the model [14]. Successfully achieving interpretability, it was found that reliable classification results can be returned under appropriate CNN architectures.

Meanwhile, not only application of Convolutional Neural Networks but also applications of transfer learning in filters or clustering in low dimensional OCT representations were also reported to have significant progress in analyzing AMD [20,21]. For example, Transfer learning based on ImageNet data was implemented for fine-tuning of GoogLeNet in [20], leading to extreme advancements in AMD classification accuracy compared to results from random initialization of filters in normal GoogLeNet based CNNs. Linking expert decisions of ophthalmologists with DNN analysis was also implemented in certain works. In [21], based on a ResNet50 CNN encoder, interpretation of K-means clustering results from low dimensional representations of AMD diagnosed OCT images by retinal specialists was implemented to find novel biomarkers for AMD risk evaluation.

C. Neural causal DAG extraction

To increase analytic flexibility in conventional causal analysis and to overcome limitations in ML/DL approaches which only guarantee correlation between explanatory variables and target variable, neural causal modeling, which links NNs with Pearl, J's structural causal model(SCM), was introduced [22-24]. Under the theorem of $\mathcal{G}$ -Consistency and L_2- $\mathcal{G}$ representation dealt by K, Xia et al. 2022, it was found that DNNs with causal diagram $\mathcal{G}$ as a constraint can fully achieve expressivity, while enabling canonical causal inference [22]. That is, when valid causal graph is set as constraint $\mathcal{G}$, non-linear causal effects or relations, which were hard to extract in conventional approaches, can be relatively easily extracted from training neural networks under gradient descent algorithms. Thus, finding valid $\mathcal{G}$ became a crucial task to perform successful NCM analysis, which was linked to vast exploitations of algorithms that can extract adjacency matrices or DAGs from observational data. Among recent works, methodologies such as NOTEARS, DAG-GNN and Graph Autoencoder based causal structure extraction were introduced and found to be valid in finding ground truth adjacency matrices from observational data in research [25-27].

Among various methods, considering low-complexity and high performance, this study implemented Graph Autoencoder based DAG extraction methods in [27] to find causal interactions between latent representations from retinal fundus images. Graph AutoEncoders (GAE) are AE based models that extract causal relationships between variables by setting an additional adjacency matrix layer between the encoder and decoder under SCM assumptions. (Though GAE introduced in [27] does not explicitly cover permutation invariance or equivariance of GNNs, the term "graph" is used based on its intrinsic ability to compute message passing and the ability to process causal graphs). In generalized SCM, ground-truth causality for feature set X can be defined as structural equation in (1). Here, any existing DAG relationships can be statistically modeled when g_1, g_2 are assumed as high dimensional functions that can cover both linearity and non-linearity.

As universal approximation theorem holds for deep neural network structures, use of multi-layer perceptrons (MLP) can approximate true g_1, g_2 functions with validity, which implies that under significant convergence in MLPs of Fig 2, true adjacency matrix A in causal graph can be successfully returned by setting A's components as trainable weight parameters in neural layers. In GAE, this weight training procedure is computed based on the objective function of (2).

$$X = g_2\left(A^T g_1(X)\right),\ A\text{: Adj mat of directed graph} \quad \textbf{(1)}$$

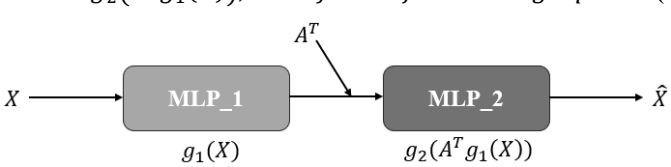


Fig 2. Visualization of Graph AutoEncoder process [27]. Weights are updated under the discrepancy between X and $\hat{X}$.

In the optimization objective of (2), parameters θ_1,θ_2 denote network weights from MLP_1 and MLP_2 , $\lambda\|A\|_1$ denotes L_1 regularization for adjacency matrix A, $\sum_{i=1}^{n}\left\|X^{(i)} - \hat{X}^{(i)}\right\|^2/n$ denotes the reconstruction error of a vanilla autoencoder, and $tr(e^{A\odot A}) = d$ works as an equality constraint that constrains weights in A to follow the acyclicity assumption of DAG. With Lagrange multiplier set as α, optimization problem of (2) can be approximated by augmented Lagrange multiplication of (3). Based optimization of (3), trainable weights in A, θ_1 and θ_2 are updated using gradient descent methods. On the other hand, parameters α, ρ are updated under iterative algorithms in (4).

$$\min_{A,\theta_1,\theta_2} \frac{1}{n}\sum_{i=1}^{n}\left\|X^{(i)} - \hat{X}^{(i)}\right\|^2 + \lambda\|A\|_1$$
$$\text{subject to } \ tr(e^{A\odot A}) = d \quad \textbf{(2)}$$

$$\min_{A,\theta_1,\theta_2} \mathcal{L}(A,\theta_1,\theta_2) = \min_{A,\theta_1,\theta_2} \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left\|X^{(i)} - \hat{X}^{(i)}\right\|^2 + \lambda\|A\|_1$$
$$+\alpha h(A) + \tfrac{\rho}{2}|h(A)|^2, \text{where } h(A) = tr(e^{A\odot A}) - d \quad \textbf{(3)}$$

$$\alpha^{(k+1)} = \alpha^{(k)} + \rho^{(k)} h(A^{(k+1)})$$
$$\rho^{(k+1)} = \begin{cases} \beta\rho^{(k)}, (if\ \left|h\left(A^{(k+1)}\right)\right| \geq \gamma\left|h(A^{(k)})\right|) \\ \rho^{(k)}, (o.w.), (\beta > 1, \gamma < 1) \end{cases} \quad \textbf{(4)}$$

III. Methodology

A. Data

For reproducibility, Retinal Fundus Multi-Disease Image public Dataset(RFMiD) was selected for analysis [28]. RFMiD dataset is an open access fundus image dataset acquired from the Eye Clinic of Sushrusha Hospital(India). Consisting of 1920 training images, 640 validation images and 640 test images, each retinal fundus image was taken in examinations from 2009 to 2020 using TOPCON TRC-NW300, TOPCON 3D OCT-2000 and Kowa VX-10. Each set of data was consisted of retinal fundus images and corresponding diagnosis or labels provided by ophthalmology experts. Specifically, each image in RFMiD was first classified into abnormal/normal retina, and then further classified into 45 types of retinal abnormalities or diseases such as presence of diabetic retinopathy, AMD, myopia, etc. As a multiclass retinal disease dataset with diverse quality, RFMiD dataset can highly represent real medical environments where patients can have multiple diseases at the same time and where medical devices can be aged or outdated. Based on its advantages and high reproducibility, many researchers implemented RFMiD in creating CNN models for diverse retinal risk forecasting [29-31].

Focusing on AMD, this research first utilized AMD labels and corresponding images in train data for representation learning (*each image was reshaped to 176×176×3). In RFMiD, AMD was labeled to images if at least one of the following conditions was met: features of neovascular AMD, drusen around the macula, and geographic atrophy related to fovea. Here, to avoid

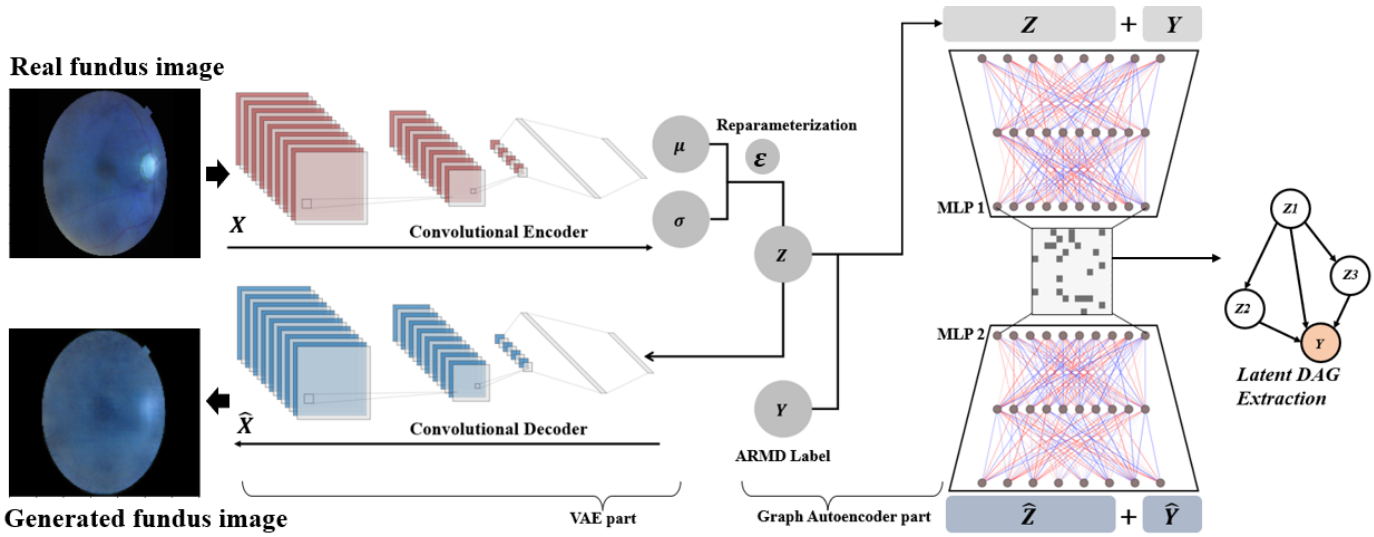


Fig 3. Visualization of CVAE based Causal Representation Learning from RFMiD images.

issues from extreme class imbalance (Non AMD: AMD = 1820:100) in training, random under sampling in the majority class was implemented to match the proportion of 200:100. After causal representation learning being completed, AMD patient data and normal retinal fundus data were used to construct high performing AMD classification ML/DL models. With training being complete, AMD and normal retinal fundus image data from RFMiD test dataset was implemented to check prediction performance of fitted ML/DL models.

B. Latent causal representation learning from fundus image

To find underlying causal interactions from only retinal fundus images with less complexity, explicit latent causal representation learning with graph autoencoder set as the latent DAG search neural algorithm was implemented. Unlike noise encoding approaches in [32,33], explicit modeling encodes causal variable itself for representation. To be specific, latent causal representation learning [32-34] is based on the assumption that when observational space X is defined, there exists an unobservable underlying latent mechanism: $Z \rightarrow_g X$, where latent variables in set Z are connected under Pearl, J's structural causal modeling(SCM) depicted in (5)-(6) [23,24].

$$\boldsymbol{X} = g(\boldsymbol{Z}) \tag{5}$$

$$\boldsymbol{Z} = A^T q(\boldsymbol{Z}) + \boldsymbol{\varepsilon} \;\; s.t. \;\; \mathcal{G} = \langle A, Z \rangle, \boldsymbol{\varepsilon}\text{: noise} \tag{6}$$

$$\boldsymbol{Z} \approx q_1(A^T q_2(\boldsymbol{Z})) \;\; s.t. \;\; tr(e^{A \odot A}) = d, \boldsymbol{Z} \in \mathbb{R}^d \tag{7}$$

Here, causal graph and the corresponding adjacency matrix are each denoted as $\mathcal{G}$ and A. For causal graph $\mathcal{G}$ to meet the conditions of DAG, when variables $\boldsymbol{Z}$ are sorted from parent node to children node, reformulated adjacency matrix A should be a triangular matrix with diagonal elements all being 0 to avoid self-loop based edges. For simplicity, these conditions can be replaced by the term in (2), which leads to an alternative definition of (7) to denote a generalized version of latent causal model under DAG $\mathcal{G}$. In this research, by setting retinal fundus image data as observational variables $\boldsymbol{X}$ and assuming AMD mechanism components in Fig 1 as $\boldsymbol{Z}$, causal representation learning was computed as in Fig 3-"Graph Autoencoder part".

By first constructing a Convolutional VAE structure to reconstruct retinal fundus images and concatenating AMD label Y with extracted latent representation vector Z, W = $\langle Z, Y \rangle$ was considered as input for Graph Autoencoder based inference. Based on the idea of simultaneous continuous gradient descent for latent causal DAG extraction and VAE fitting in [32], Optimization objective of CVAE weights were defined as in (8). For $\mathcal{L}_{VAE}$, negative Evidence Lower Bound(ELBO) was used to approximate VAE loss under proof of (9), whereas for $\mathcal{L}_{GAE}$, optimization problem formation dealt in II.C. was implemented to achieve latent causal DAG $\mathcal{G}$. In (9), as KL-divergence is a non-negative value, ELBO can be defined as the lower bound for log-likelihood. Therefore, under an MLE approach, maximization of ELBO loss leads to the maximization of log-likelihood, which makes it possible to set negative ELBO loss values as optimization objectives in VAE [35]. Under $p(\cdot) \sim Normal$, 1-step (L=1) Monte-Carlo simulation in reparameterization and an identity covariance matrix assumption, ELBO can be further reduced to a mean squared error (MSE) form in (10) [35]. Although assuming a Bernoulli distribution for $p(\cdot)$ would be ideal for image generation as it

$$\begin{aligned} \mathcal{L}_{total} &= \mathcal{L}_{VAE} + \mathcal{L}_{GAE} \\ &\approx -ELBO(\phi) + \mathcal{L}_{GAE}, \;\; \phi\text{: parameter of encoder} \end{aligned} \tag{8}$$

$$\begin{aligned} &\log(p(\boldsymbol{x})) = \int \log(p(\boldsymbol{x}))\, q_\phi(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= \int \log\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}\right) q_\phi(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= \int \log\left(\frac{p(\boldsymbol{x},\boldsymbol{z}) q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x}) q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right) q_\phi(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= \int \log\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right) q_\phi(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} + \int \log\left(\frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x})}\right) q_\phi(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= ELBO(\phi) + KL(q_\phi(\boldsymbol{z}|x) || p(\boldsymbol{z}|\boldsymbol{x})) \geq ELBO(\phi), \because KL \geq 0 \\ &\therefore \text{argmax ELBO}(\phi) \implies \text{argmax} \log(P(\boldsymbol{X})), q_\phi\text{: VAE encoder} \end{aligned} \tag{9}$$

leads to cross-entropy loss, due to large discrepancies generated between $\mathcal{L}_{VAE}$ and $\mathcal{L}_{GAE}$ under the use of cross-entropy, normal distribution assumptions were chosen to achieve balanced convergence in both objectives. Therefore, $\mathcal{L}_{total}$ for CVAE optimization was defined as (11). For updating weights in VAE, loss function of $\mathcal{L}_{total}$ was implemented. For updating weights

$$
\begin{aligned}
ELBO(\phi) &= \int \log\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right) q_\phi(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} \\
&= \int \log\left(\frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right) q_\phi(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} \\
&= \int \log(p(\boldsymbol{x}|\boldsymbol{z}))\, q_\phi(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} - \int \log\left(\frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z})}\right) q_\phi(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} \\
&= E_{q_\phi}(\log(p(\boldsymbol{x}|\boldsymbol{z}))) - KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \\
&\simeq \frac{1}{L}\sum_l \log\left(p(\boldsymbol{x}|\boldsymbol{z}^l)\right) - KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \\
&\Rightarrow \log \prod_{j=1}^{D} p(x^{(j)}|\boldsymbol{z}^l) - KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})), if\ L = 1, \boldsymbol{X} \in \mathbb{R}^{N\times D} \\
&= \sum_{j=1}^{D} \log\left(p(x^{(j)}|\boldsymbol{z}^l)\right) - KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \\
&\propto \sum_{j=1}^{D} -\left(x^{(j)} - \mu^{(j)}\right)^2, (if\ p(\cdot) \sim Normal \text{ and } Cov \sim I) \\
&\therefore\ \mathcal{L}_{VAE} = \textstyle\sum_{i=1}^{N} \left\|X_i - \hat{X}_i\right\|_F^2 / N = MSE(X_i, \hat{X}_i)
\end{aligned} \tag{10}
$$

in GAE structures, loss function of $\mathcal{L}_{GAE}$ was implemented to put more emphasis on valid latent adjacency matrix extraction. Thus, latent causal representations $\boldsymbol{Z}$ that correspond with underlying features of retinal fundus status were extracted with validity and low complexity.

$$
\begin{aligned}
&\min_{\Theta,A} \mathcal{L}_{total} = \min_{\Theta,A}(\mathcal{L}_{VAE} + \mathcal{L}_{GAE}) \\
&= \min_{\Theta,A} MSE(X_i, \hat{X}_i) + MSE(W_i, \widehat{W}_i) + \lambda\|A\|_1 + \alpha h(A) + \frac{\rho}{2}|h(A)|^2 \\
&\alpha^{(k+1)} = \alpha^{(k)} + \rho^{(k)} h(A^{(k+1)}) \\
&\rho^{(k+1)} = \begin{cases} \beta\rho^{(k)}, (if\ |h(A^{(k+1)})| \geq \gamma|h(A^{(k)})|) \\ \rho^{(k)}, (o.w.), (\beta > 1, \gamma < 1) \end{cases}
\end{aligned} \tag{11}
$$

C. Model Structure Settings-VAE, GAE

Specific structures for explicit causal representation modeling were constructed as follows. For Convolutional VAE's encoder structures, three convolutional layers and two dense layers after flattening were implemented for latent space retrieval. First convolutional layer was set as having 64 filters with kernel size set as 4×4 and stride size set as 2×2. The second convolutional layer was set as having 64 filters with kernel size set as 6×6 and stride size set as 3×3. The third convolutional layer was set as having 16 filters with kernel size set as 4×4 and stride size set as 2×2. After flattening data from the third layer and adding an additional dense layer with 256 hidden nodes, fully connected dense layer with $2|\mathcal{Z}|$ hidden nodes was set for $\boldsymbol{\mu}$ and $\log(\boldsymbol{\sigma}^2)$ in VAE. After generating latent variable $\boldsymbol{Z}$ using standard gaussian noise $\boldsymbol{\varepsilon}$ based reparameterization: $\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon}$, a convolutional decoder which has a symmetrical structure with the former encoder was constructed for retinal image reconstruction. Visualization of VAE structures was computed as in Fig 4. To avoid over-fitting issues, dropout layers with a dropout rate of 10% were implemented between Conv2D or Conv2DTranspose layers. For activation functions, ReLU functions were implemented for conv layers, while ELU functions were implemented for dense layers. Furthermore, for efficient weight initialization, He-Normal initialization [36] was implemented in all hidden layers. All structures for CVAE were constructed based on Python Tensorflow Keras package under tensorflow version 2.18.0.

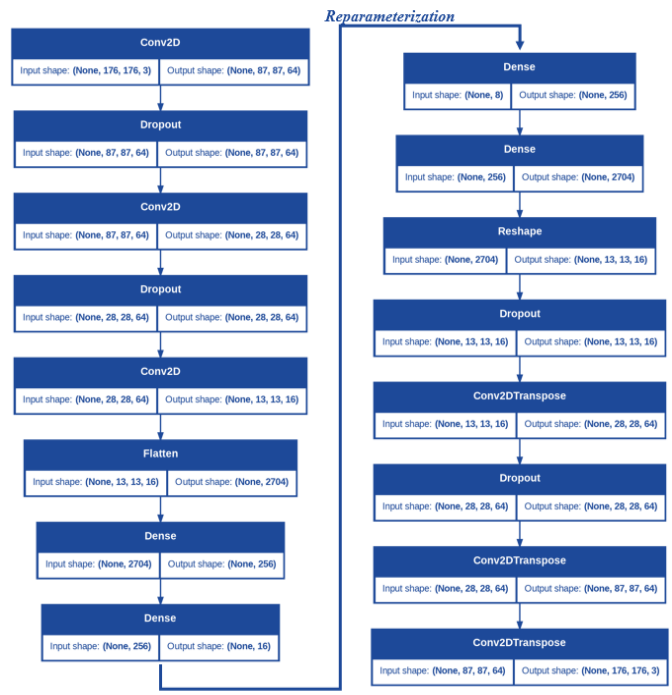


Fig 4. Visualization of CVAE structure framework (number of latent variables = 8)

For GAE structures, generalized additive modeling approach within the GAE encoder and decoder was applied to constrain MLPs to follow the form of (7). To perform additive modeling in layers, weight masking based on tensorflow.keras. constraints.Constraint was implemented [37]. Working as a base class for custom constraint generation, tensorflow.keras. constraints.Constraint can be embedded as a kernel constraint in most keras layers with high efficiency. Using weight masking, each input variable in W was assigned a separate MLP with two hidden layers, each having 5 hidden nodes, and an output layer, which has only one node. Therefore, a neural network that has 2 non-FCN hidden layers with each having $5(|\mathcal{Z}| + 1)$ hidden nodes and an output layer with $(|\mathcal{Z}| + 1)$ nodes was constructed for both the encoder and the decoder: Fig 5. Apart from basic AE structures, an adjacency weight matrix layer with trainable weights was implemented between two NNs for GAE to extract true edges in latent causal graph $\mathcal{G}$. In further steps, weighted adjacency matrix after training was transformed to a binary adjacency matrix using importance based sequential selection.

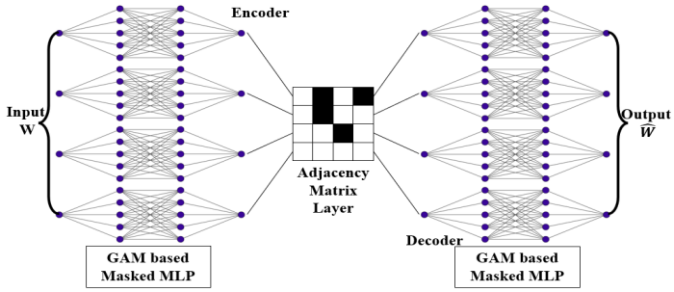


Fig 5. Example of GAE framework with 4 inputs. Each feature was assigned individual MLPs using weight masking.

D. Experimental Design

Process of training weights in CVAE-GAE and checking performance of latent causal variable extraction were computed as follows. For updating weights, different learning rates were implemented due to non-identical loss function usage. For CVAE, loss function of $\mathcal{L}_{total}$ and an Adam optimizer with learning rate=0.0004 were implemented for optimization. For GAE, loss function of $\mathcal{L}_{GAE}$ and an Adam optimizer with learning rate=0.003 were implemented. For parameter settings regarding GAE's DAG constraint, initial values were set as α=0.6, ρ=0.1, γ=0.9, β=1.01 and λ=1.0. Additionally, to enforce fast convergence in DAG constraint satisfaction, diagonal elements in adjacency matrix and outgoing edges of Y were blacklisted by multiplying a corresponding binary mask matrix. Using simultaneous updating for CVAE and GAE weights, 800 epochs under full-batch gradient descent were computed for optimization.

After training being complete, for convergence check, MSE of CVAE and loss value of CVAE were used as metrics. Under convergence, optimal weighted adjacency matrix was binarized by converting absolute weight values above the 100(1-q)th percentile as 1 and absolute weights below the threshold as 0. To extract only highly influential weights, hyperparameter q was set as q=0.15. Based on these results, true causal graph recovery was checked by computing the structural hamming distance(SHD) between causal graph in Fig 1 and the subgraph from extracted DAG. Specifically, to focus only on AMD mechanisms, only edges that are within 2-hop neighbors from Y and edges toward Y were considered. Furthermore, under the assumption that extracted causal features are all elements of true causal set, when SHD values can vary under diverse interpretations, minimum SHD was selected using best possible match. Increasing the number of latent causal variables by +1, optimal latent space dimension was selected from the candidate set: {5,6,7,8} under the objective of minimizing SHD. Among candidates with identical SHD, largest candidate was set as the optimal value to minimize information loss in encoding steps.

To check AMD mechanism identification in latent space, success in achieving causal disentanglement was checked by comparing reconstructed AMD images and original images under modification of individual z_i, with other latent variables being fixed. Lastly, after checking one-to-one identification or correspondence between causal variables in set Z and AMD factors in Fig 1. (neovascularization, drusen, hemorrhage and fluids), intervention analysis based on fitted GAE was discussed as an example to illustrate clinical advantages of using causal latent space. (*All experiments were computed under Google Colab's basic T4 NVIDIA GPU environments).

E. AMD prediction model construction

Under successful causal disentanglement in latent representations, ML/DL based AMD prediction models were trained to check prediction performance under latent causal variable usage. Using normal patient data and AMD patient data images in RFMiD training dataset, latent representations Z were extracted from trained CVAE-encoders, which were then used as input for ML/DL models with label Y. Specifically, from ML models, Random Forest, Extra Trees and Gradient Boosting Machines in sklearn.ensemble were implemented, whereas from DL models, a simple DNN with two hidden layers(128, 64 nodes) was implemented(epochs=300, learning rate=4e-4). Each model was then evaluated using test dataset with five distinct metrics: classification accuracy, macro-F1 score, macro-precision score, macro-recall score and ROC AUC score. Based on these results, not only causal mechanism comprehension in vision models but also latent causal variable usage for enhanced AMRD prediction was checked.

IV. Experimental Results

A. CVAE and GAE fitting results

Among candidate values for optimal latent causal variables, 8 was deduced as the optimal value to minimize SHD in TABLE I. Therefore, based on $|\mathcal{Z}| = 8$, CVAE and GAE were trained for optimal latent causal representation learning. Fitting results of CVAE+GAE were deduced as Fig 7. In both neural network structures, convergence to 0 in $\mathcal{L}_{VAE}$ and $\mathcal{L}_{GAE}$ was checked after 800 epochs of training, which implies that both latent causal representation learning and image reconstruction objectives were successfully achieved. Specifically, MSE value of CVAE and $\mathcal{L}_{GAE}$ were computed as 0.0019 and 0.7659.

TABLE I
Structural Hamming Distance(SHD) from candidate set

True Graph	**Extracted Sub-Graph from GAE**		
Structure	**Latent dim**	**Structure**	**Min SHD**
	5		1.0
	6		1.0
	7		2.0
	8*		**1.0***

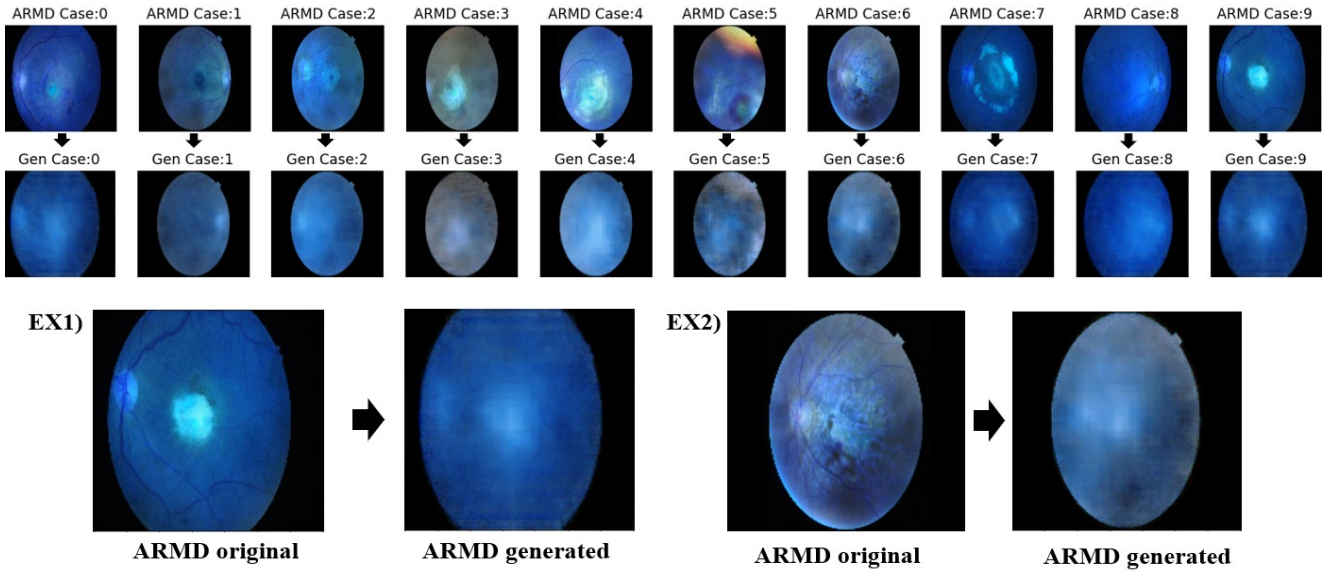


Fig 6. Visualization of CVAE+GAE fitting results. Original AMD patient retinal fundus images and corresponding generated images

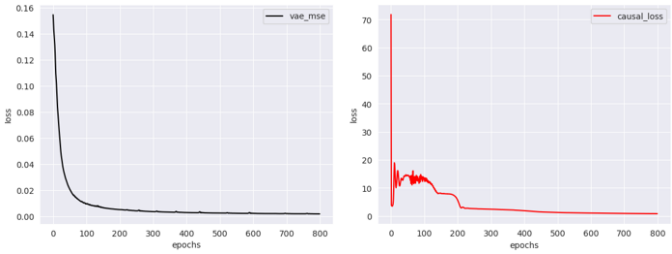


Fig 7. Visualization of latent causal representation learning procedure. (Left) MSE of VAE per epoch (Right) Total loss

Visual comparisons between original AMD retinal fundus images and generated images were also processed to check validity of training. In Fig 6, 10 images among successful AMD retinal fundus image reconstructions were visualized as examples of CVAE results. Images in the first row denote original retinal fundus images of AMD diagnosed patients, whereas images in the second row denote generated images from CVAE. As it can be seen in Fig 6., abnormally dark light reflections (hypo-reflections), which can be presumed due to retinal fluids or hemorrhage [38,39], and bright clusters, which can be presumed due to AMD related drusen or exudates [38,40], were well captured and reconstructed using latent causal modeling. For example, regarding seventh and tenth AMD images in Fig 6., CVAE successfully retrieved retinal fundus contaminations(hypo-reflective) and highly reflective legions, which are each presumed to be results of subretinal fluid leakage or hemorrhage (second example in third row) and results of drusen or drusen-oriented geographic atrophy around the macular region (first example in third row).

Based on trained CVAE+GAE structures, weighted adjacency matrix and binarized adjacency matrix from training latent DAGs were also computed as in Fig 8. Among latent variables, z_0 and z_4 were found to directly affect AMD status with high importance. Not having bidirectional edges or cycles in all nodes, conditions of DAG constraint (especially acyclicity for latent representations) were found to be successfully achieved. Furthermore, with SHD from the extracted subgraph from latent DAG being only 1, high possibility of valid AMD mechanism identification was found in training results. Specifically, among latent causal representations, variables z_0, z_4 and z_5 were found to be possible candidates that can match with components in Fig 1.

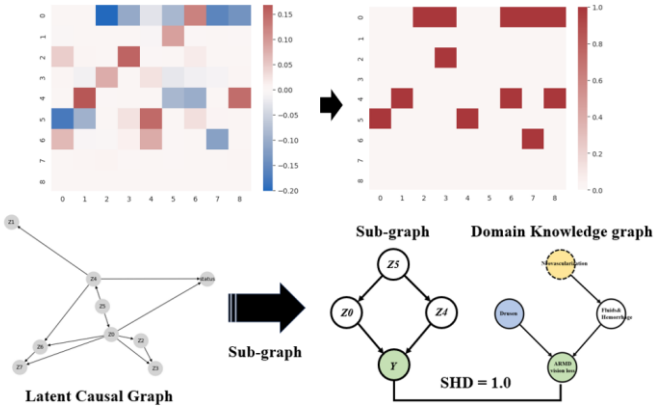


Fig 8. Visualization of trained adjacency matrix from latent causal variables and *Y*: AMD status. (Upper right) Binarized *A*. (Lower row) Latent causal DAG extraction.

B. Causal disentanglement check

To validate whether latent causal variables can be match with domain knowledge-based AMD mechanism factors, causal disentanglement was checked under multiple images by modifying each z within (z_0, z_4, z_5), with other variables being fixed. Here, set (z_0, z_4, z_5) is the encoded latent causal representation of the original retinal fundus image. As retinal status or disease development is not a discrete process, and as no specific segmentations of hemorrhage or drusen were given, qualitative evaluations based on visual characteristics were mainly computed instead of implementing intense quantitative approaches such as entropy-based disentanglement score

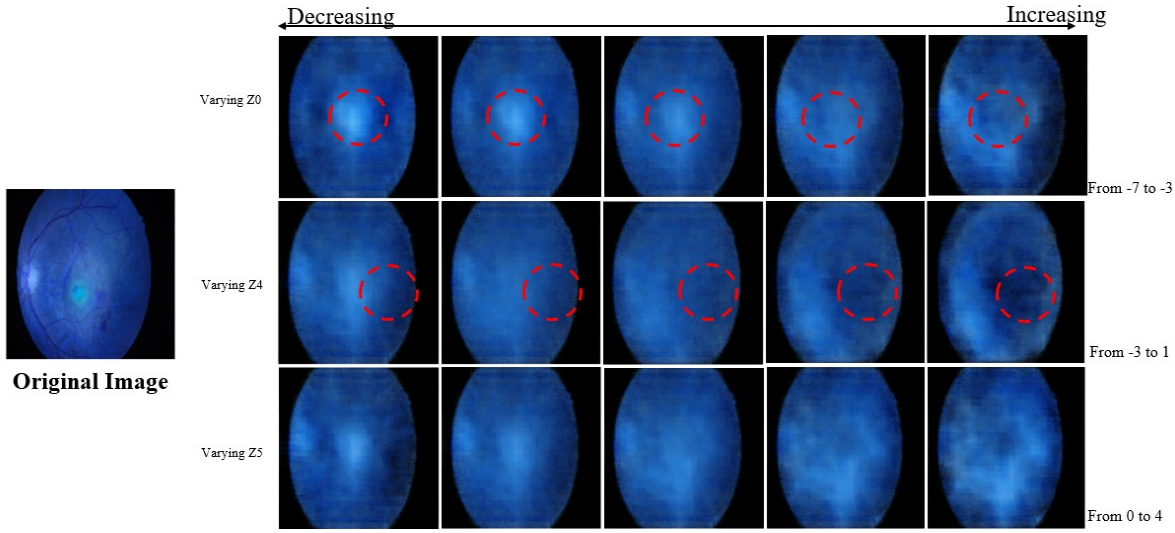


Fig 9. Visualization of Causal disentanglement check on random image. (Right) Original retinal image, (Left) Varying individual latent causal factors in CVAE reconstruction. Z0 and Z4 were both found to be valid causal representations of AMD mechanisms.

metrics [41]. Under clinical features or degree of reflectiveness, it was found that z_4 can correspond with hemorrhage and intraretinal fluids, whereas z_0 can correspond with drusen which is centralized around the fovea. For example, in Fig 9., by varying z_0 from -7 to -3 under equal intervals, only the area around the fovea was found to be altered. In this case, as z_0 values decreased, a highly reflective cluster around the fovea was found to be generated, which is a representative feature of drusen within retinal fundus images. Furthermore, for quantitative evaluation, one-sided Wilcoxon rank sum test for z_0 values from drusen diagnosed and non-diagnosed groups was also computed. Having a test statistic of 1.6840 and p-value of 0.0461(< 0.05), latent variable z_0 was also found to be highly related with drusen under statistical significance (significance level=0.05). Meanwhile, in Fig 9., by varying z_4 from -3 to 1, only the area which corresponds to low reflectional components or dark matters in the original retina was found to be altered. In this case, as z_4 increased, both size and darkness of specific regions where existence of hemorrhage is presumed in the original image increased dramatically. (More representative examples of disentanglement checks or variable modification results regarding z_0 and z_4 were illustrated in Appendix).

On the other hand, modifying z_5 in latent representations led to non-local changes across the retina. For example in Fig 9., as z_5 increased from 0 to 4, hyper reflective regions, which were found to be concentrated around the fovea, spread towards the outer edge of the retina. Considering characteristics of neovascularization, which is hard to identify under fundus images and is found to be hypo-reflective, it was deduced that z_5 cannot be classified as a proper causal representation of angiogenesis. However, considering characteristics in visualized results for disentanglement checks and causal linkage between z_4 as a parent node in adjacency matrix of Fig 8., it was found plausible to say that z_5 has the possibility of representing certain states of parental variables regarding neovascularization, such as aging or inflammations in retina.

Thus, not only having significant success in extracting valid latent causal representation learning in the aspect of MSE loss under unsupervised environments, but also having substantial success in validly identifying major components in the AMD mechanism using low-dimensional spaces, it was implied that trained CVAE-GAE models can return accurate retinal fundus images under significant comprehension on latent structures of retinal status and can simulate/visualize interventional effects or results of treatments on AMD patients by altering target values among variable set Z with reliability. For example, when sufficient regression or ATE analysis between novel treatments and z_4 is driven, clinicians can simulate customized visual results of treatments, such as the progress of intraretinal fluid elimination within the retina, in AMD risk patient's retina beforehand by using interventional GAE outputs: $\hat{Z}$, which can be acquired from modifying current z_4 to estimated $\hat{z_4}$, in generating individual retinal fundus images.

C. AMD prediction modeling using latent causal variables

Based on the success in latent causal AMD-related factor extractions, this research attempted to construct a causal feature based AMD diagnosis prediction model using pre-trained CVAE encodings. Explanatory variables for training were computed via encoding Normal and AMD diagnosed patient's retinal fundus images among original train data into 8-dimensional latent causal features, whereas target variables for training were set as corresponding labels(0: normal, 1: AMD). After training, prediction models were evaluated using test data under identical approach where encodings of images are considered as explanatory variables. Regarding each ML model settings, a Random Forest Classifier(RF) with 100 trees(max depth=10), Gradient Boosting Machine with 100 trees, and an Extra Trees Classifier with 100 trees(max depth=10) were implemented. For DL model settings, a deep neural network with two fully-connected dense layers was implemented. Specifically, for analysis, simple DNN which has 128 and 64 hidden nodes for each hidden layer, ELU activation functions for hidden layers, batch normalization between layers and a sigmoid activation function for binary outputs was constructed

TABLE II
TRAINING AND TEST RESULTS FOR AMD PREDICTION MODELS

Model	Train data		Test data				
	Accuracy	F1-score	Accuracy	Macro Precision	Macro Recall	Macro F1-score	ROC AUC
DNN	0.9641	0.9118	**0.9212**	**0.9020**	**0.8275**	**0.8582**	**0.9201**
RF	**1.0000**	**1.0000**	0.8424	0.7890	0.6178	0.6453	0.9035
GBM	0.9840	0.9583	0.8545	0.8707	0.6253	0.6586	0.8763
ExtraTree(ET)	0.9920	0.9804	0.8424	0.7507	0.6674	0.6936	0.8912

***F1-score, precision, recall scores were based on Macro-averaging**

for experiments. To consider class imbalance in ML model fittings, hyperparameter "class_weight" in sklearn.ensemble models were set as "balanced" in RF and ET. In DL model settings, TensorFlow's Binary Focal Crosse Entropy with class balancing=True, alpha=0.6 and label_smoothing=0.3 was implemented as a loss function to consider class imbalance. Based on above settings, experimental results were deduced as TABLE II. When fitting models to training data, all four models achieved high accuracy and F1-scores in classifying train data. Among models, RF achieved highest accuracy and F1-score of 100%. In all models, accuracies and F1-scores exceeded 90%. Thus, it was found that using latent causal representations instead of image itself can also lead to successful fitting in AMD classification. Meanwhile, in evaluation based on test data, though achievement of high classification accuracy was found in all models, significant variance in other prediction performance metrics was identified among different types of models. Among four models, DNN showed highest prediction performance in all metrics (accuracy: 92.1%, macro precision: 90.2%, macro recall: 82.8%, macro f1-score: 85.8%), whereas RF showed lowest performance in all metrics(accuracy: 84.24%, macro precision: 78.90%, macro recall: 61.78%, macro f1-score: 64.53%) based on test data. Unlike ML models, which showed over-fitting result or limitations in extracting true positives (AMD cases) in terms of macro-recall score and f1-scores, DNN showed consistent performance in all metrics, leading to a ROC AUC score of 92.01%. Specific results of DNN prediction performance were visualized in Fig 10. According to the confusion matrix in Fig 10., though high class imbalance was present in test data, it was found that DNN had substantial success in detecting AMD retinas among mixed data,

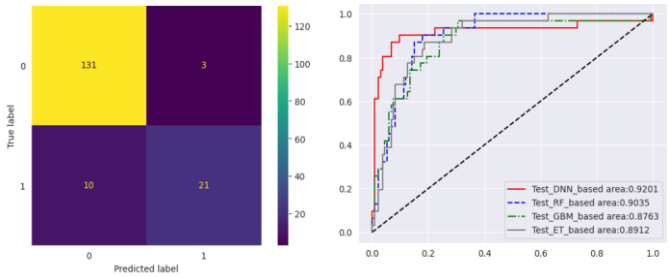


Fig 10. Visualization prediction performance of trained DNN based on latent causal variables extracted from test data. (Right) Confusion Matrix: 0 denotes Normal retinal fundus, 1 denotes AMD retinal fundus. (Left) ROC curves of models

while minimizing false positive predictions. Having a weighted F1-score of 91.7%, macro F1-score of 85.8% and accuracy of 92.1%, use of latent causal factors extracted from retinal fundus images itself in DL architectures was found to be effective in achieving reliable(robust) and valid AMD predictions. This implies that latent causal modeling based on previous CVAE + GAE structures not only enables valid causal intervention analysis or intervention simulations for specific AMD treatments, but also enables successful downstream tasks such as state-of-art AMD diagnosis by connecting deep learning based neural network models with encodings.

V. CONCLUSION

In this study, for ML/DL models to comprehend or consider the causal mechanism of AMD development in various downstream tasks, Convolutional VAE + Graph autoencoder based explicit latent causal representation learning was implemented to identify causal factors and relations from retinal fundus images directly. Based on successful convergence in both losses: $\mathcal{L}_{VAE}, \mathcal{L}_{GAE}$, it was deduced that the achieved latent causal graph contains valid underlying causal factors linked to AMD. Furthermore, by first comparing topological structures of extracted graph and pre-defined domain knowledge-based graph under SHD, and then modifying individual latent causal variables in extracted sub-graph, it was found that causal disentanglement can be fulfilled regarding drusen and NV-based hemorrhage or intraretinal fluids in latent causal representations. Thus, with validity found in CVAE + GAE based latent causal encoders, downstream tasks of classifying and predicting AMD diagnosis were computed based on Normal/AMD diagnosed patient's retinal fundus images. Results show that though high class imbalance was present, latent causal encoding based DNN achieved extreme prediction performance in all metrics: Accuracy = 92.1%, Macro F1-score = 85.8%, ROC AUC score = 92.0%. This implies that through suggested CVAE+GAE modeling for AMD, not only valid causal intervention analysis can be computed with extracted latent causal AMD DAG, but also reliable AMD predictions can be returned while using only low quality retinal fundus images.

Though accurate predictions or treatment (intervention) simulations regarding AMD can be expected using results from this research, there were also clear limitations which should be thoroughly discussed in further works. For example, though

hyper/hypo reflective components or contaminations in retinal fundus were well reconstructed in CVAE, reconstruction of specific blood vessels around the macula failed in most image generation cases, which were presumed due to low quality or low resolution in sample images. Limitations in implementing quantitative metrics were also found to be present in causal disentanglement checks dealt in IV.B. Though disentanglement within causal factors were visually apparent under numerous simulations, due to the intrinsic continuity (non-discreteness) in AMD related factors, conventional disentanglement scoring methods (except SHD) were not able to be applied in evaluation. Thus, in further developments, to achieve a more reliable and complex latent causal analysis regarding mechanisms of AMD, obtaining high resolution fundus images using diffusion models and converting continuous latent variables into binary spaces should be implemented in various research.

VI. Reference

[1] R. E. K. Man et al., Impact of incident age-related macular degeneration and associated vision loss on vision-related quality of life, *British Journal of Ophthalmology*, vol. 106, pp. 1063-1068, 2022.

[2] D. B. Rein et al., Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study, *Lancet Glob Health*, vol. 9, no. 2, pp. 144-160, 2022.

[3] CDC, About Common Eye Disorders and Diseases, (https://www.cdc.gov/vision-health/about-eye-disorders/index.html.)

[4] CDC, VEHSS Modeled Estimates for Age-Related Macular Degeneration, (https://www.cdc.gov/vision-health-data/prevalence-estimates/amd-prevalence.html)

[5] T. D. L. Keenan, C. A. Cukras, and Emily Y, Age-Related Macular Degeneration: Epidemiology and Clinical Aspects, *Adv Exp Med Biol*, vol. 1256, pp. 1-31, 2021.

[6] S. Hwang et al., Impact of Age-Related Macular Degeneration and Related Visual Disability on the Risk of Depression, *Ophthalmology*, vol. 130, no. 6, pp. 615-623, 2023.

[7] S. Schmitz-Valckenberg et al., Combined Confocal Scanning Laser Ophthalmoscopy and Spectral-Domain Optical Coherence Tomography Imaging of Reticular Drusen Associated with Age-Related Macular Degeneration, *Ophthalmology*, vol. 117, no. 6, pp. 1169-1176, 2010.

[8] A. Tomi and I. Marin, Angiofluorographic aspects in age-related macular degeneration, *Journal of Medicine and Life*, vol. 7, no. 4, pp. 4-17, 2014.

[9] N. G. Lambert et al., Risk Factors and Biomarkers of Age-Related Macular Degeneration, *Prog Retin Eye Res*, vol. 54, pp. 64-102, 2016.

[10] R. J. Ross et al., Genetic markers and biomarkers for age-related macular degeneration, *Expert Rev Ophthalmol*, vol. 2, no. 3, pp. 443-457, 2007.

[11] V. Hrishikesh and S. Pranaykumar, Age-Related Macular Degeneration: Epidemiology, Pathophysiology, Diagnosis, and Treatment, *Cureus*, vol. 14, no. 9, e29583, 2022.

[12] C. S. Lee, D. M. Baughman, and A.Y. Lee, Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images, *Ophthalmology Retina*, vol. 1, no. 4, pp. 322-327, 2017.

[13] S. Li et al., Attention-Aware Convolutional Neural Network for Age-Related Macular Degeneration Classification, *2020 12th ICCSN*, pp. 264-269, 2020.

[14] M. Wang et al., Applications of Explainable Artificial Intelligence Algorithms to Age-Related Macular Degeneration Diagnosis: A Case Study Based on CNN, Attention, and CAM Mechanism, *2023 IEEE InC4*, pp. 1-5, 2023

[15] M. Van Lookeren Campagne et al., Mechanisms of age-related macular degeneration and therapeutic opportunities, *The Journal of Pathology*, vol. 232, no. 2, pp. 151-164, 2013.

[16] N. Marchesi et al., Different Therapeutic Approaches for Dry and Wet AMD, *Int J Mol Sci*. vol. 25, no. 23, pp. 13053, 2024.

[17] P. V. Algvere, A. Kvanta, and S. Seregard, Drusen maculopathy: a risk factor for visual deterioration, *Acta Ophtalmol*, vol. 94, no. 5, pp. 427-433, 2016.

[18] S. M. Waldstein, W. D. Vogl, and H. Bogunovic et al., Characterization of Drusen and Hyperreflective Foci as Biomarkers for Disease Progression in Age-Related Macular Degeneration Using Artificial Intelligence in Optical Coherence Tomography, *JAMA ophthalmology*, vol. 138, no. 7, pp. 740-747, 2020.

[19] S.S. Mishra, B. Mandal, and N. B. Puhan, Multi-Level Dual-Attention Based CNN for Macular Optical Coherence Tomography Image Classification, *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1793-1797, 2019.

[20] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration, *Biomedical Optics Express*, vol. 8, no. 2, pp. 579-592, 2017.

[21] R. Holland et al., Deep Learning–Based Clustering of OCT Images for Biomarker Discovery in Age-Related Macular Degeneration (PINNACLE Study Report 4), *Ophthalmology Science*, vol. 4, no. 6, pp. 100543, 2024.

[22] K. Xia, K.Z. Lee, Y. Bengio, and E. Bareinboim, The Causal-Neural Connection: Expressiveness, Learnability, and Inference, *NeurIPS 2021*, pp. 1-14, 2022.

[23] J. Pearl, Causal inference in statistics: An overview, *Statistics Surveys*, vol. 3, pp. 96-146, 2009.

[24] J. Pearl, Causal diagrams for empirical research, *Biometrika*, vol. 82, pp. 669–710, 1995.

[25] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, DAGs with no tears: Continuous optimization for structure learning, *NeurIPS 2018*, pp. 9472-9483, 2018

[26] Y. Yu, J. Chen, T. Gao, and M. Yu, DAG-GNN: DAG structure learning with graph neural networks, *International Conference on Machine Learning*, vol. 97, pp. 7154-7163, 2019.

[27] I. Ng, S. Zhu, Z. Chen, and Z. Fang, A graph autoencoder approach to causal structure learning, *NeurIPS 2019 Workshop*, pp. 1-8, 2019.

[28] S. Pachade et al., Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for Multi-Disease Detection Research, *Data*, vol 6, no. 2, 2021. (DOI: https://doi.org/10.3390/data6020014)

[29] Y. Hazem et al., Retinal Fundus Diseases Detection and Identification Using CNN, *2023 ICICIS*, pp. 375-380, 2023.

[30] L. Lyu et al., Multi-Expert Deep Networks for Multi-Disease Detection in Retinal Fundus Images, *2022 EMBC*, pp. 1818-1822, 2022.

[31] Q. Abbas et al., Deep-Ocular: Improved Transfer Learning Architecture Using Self-Attention and Dense Layers for Recognition of Ocular Diseases, *Diagnostics*, vol. 13, no. 20. 3165, 2023.

[32] J. Brehmer et al., Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems(NeurIPS)*, vol. 35, pp. 38319-38331, 2022.

[33] M. Yang et al., CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models, *2021 CVPR*, pp. 9588-9597, 2021.

[34] I. Ng et al., Score-Based Causal Discovery of Latent Variable Causal Models, *PMLR*, vol. 235, pp. 37531-37552, 2024.

[35] P. Kingma Diederik and M. Welling, Auto-Encoding Variational Bayes, *arXiv:1312.6114*, pp. 1-14, 2013.

[36] K. He et al., Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, *ICCV*, pp. 1026-1034, 2015.

[37]tf.keras.constraints.Constraint,(https://www.tensorflow.org/api_docs/python/tf/keras/constraints/Constraint).

[38] G. A. Saleh et al., The Role of Medical Image Modalities and AI in the Early Detection, Diagnosis and Grading of Retinal Diseases: A Survey, *Bioengineering*, vol. 9, no. 8, pp. 366, 2022.

[39] S. Dinu et al., Submacular hemorrhage in neovascular age-related macular degeneration: A synthesis of the literature, *Survey of Ophthalmology*, vol. 61, no. 1, pp. 18-32, 2016.

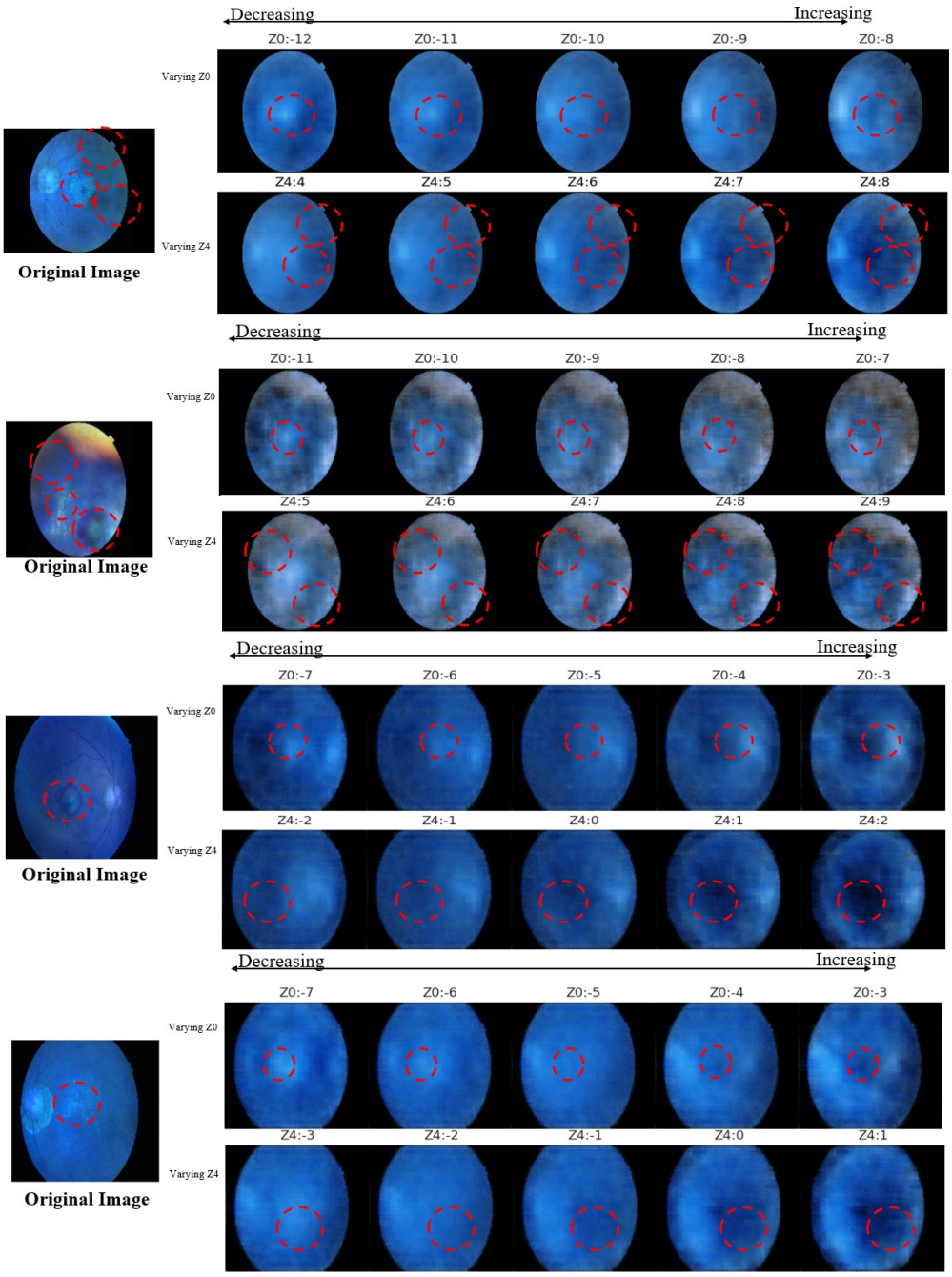
[40] A. Ly, Infrared reflectance imaging in age-related macular degeneration, *Ophthalmic Physiol Opt*, vol. 36, no. 3, pp. 303-306, 2016.

[41] E. Cian and K. Christopher, A Framework for the Quantitative Evaluation of Disentangled Representations, *ICLR 2018*, pp. 1-15, 2018.

VII. Appendix

A. Causal Disentanglement Check

Additional examples of causal disentanglement checks were visualized below (Selected original images among AMD patients for intense comparison). First row denotes varying z0, whereas second row denotes varying z4. Disentanglement success found in Fig 9. were also found in all additional cases. For example, in the first case, hyper reflective regions found in the original image around the macula, which can be presumed to be drusen based geographic atrophy, were well identified as Z0 decreased from -8 to -12 in CVAE+GAE based reconstructions. For the second example, auto-fluorescent features and locations of existing drusen clusters(identified in original image) were also identified when decreasing Z0. Hemorrhage or intra retinal fluid in the retina was also successfully identified in most experimental cases. For example, locations and hypo-reflective features of vivid hemorrhage next to the cluster of drusen in the second example and the hemorrhage around the macula in the third example were both clearly identified as Z4 was increased in simulation. Having consistent results in identifying causal AMD factors, CVAE+GAE based retinal fundus causal representation learning was found to not only succeed in returning valid underlying causal connections between retinal components, but also succeed in disentangling key factors of AMD, which can lead to reliable interventional analysis or simulations. (Github: https://github.com/Daeyoung25-Kim/AMD_LatentCausalRepresentationLearning_BioRxiv)

*For Z0, quantitative evaluation(Wilcoxon rank sum test) was also computed based on pre-defined labels for drusen in full RFMiD training data.

*Wilcoxon rank sum test under 0.05 significance level was computed based on ranksums(alternative="greater") function in Python scipy.stats package.