COMPRER: A MULTIMODAL MULTI-OBJECTIVE PRETRAINING FRAMEWORK FOR ENHANCED MEDICAL IMAGE REPRESENTATION

A PREPRINT

Guy Lutsker[1]
guy.lutsker@weizmann.ac.il

Hagai Rossman[2]
hagai@pheno.ai

Nastya Godiva[1]
nastya.godiva@weizmann.ac.il

Eran Segal[1,3]
eran.segal@weizmann.ac.il

March 17, 2024

ABSTRACT

Substantial advances in multi-modal Artificial Intelligence (AI) facilitate the combination of diverse medical modalities to achieve holistic health assessments. We present COMPRER , a novel multi-modal, multi-objective pretraining framework which enhances medical-image representation, diagnostic inferences, and prognosis of diseases. COMPRER employs a multi-objective training framework, where each objective introduces distinct knowledge to the model. This includes a multi-modal loss that consolidates information across different imaging modalities; A temporal loss that imparts the ability to discern patterns over time; Medical-measure prediction adds appropriate medical insights; Lastly, reconstruction loss ensures the integrity of image structure within the latent space. Despite the concern that multiple objectives could weaken task performance, our findings show that this combination actually boosts outcomes on certain tasks. Here, we apply this framework to both fundus images and carotid ultrasound, and validate our downstream tasks capabilities by predicting both current and future cardiovascular conditions. COMPRER achieved higher Area Under the Curve (AUC) scores in evaluating medical conditions compared to existing models on held-out data. On the Out-of-distribution (OOD) UK-Biobank dataset COMPRER maintains favorable performance over well-established models with more parameters, even though these models were trained on $75\times$ more data than COMPRER. In addition, to better assess our model's performance in contrastive learning, we introduce a novel evaluation metric, providing deeper understanding of the effectiveness of the latent space pairing.

1 Background

The evolution of AI within healthcare is promoting an era of precision medicine, marked by enhanced diagnostic accuracy, improved prognostic evaluations, and personalized treatment strategies (Bajwa et al., 2021; Esteva et al., 2019), where deep learning is increasingly central to medical imaging analysis (Huang et al., 2023). Technologies such as fundus imaging and carotid ultrasound are pivotal in cardiovascular health assessments, granting insights into micro and macrovascular structures and pathologies (Poplin et al., 2018; Spence, 2006). Fundus imaging, a non-invasive procedure, reveals the retinal microvasculature and is used to detect early manifestations of diseases like diabetes and hypertension (Dai et al., 2021; Yan et al., 2019). Such microvascular changes are significant indicators of systemic conditions, enabling broader health monitoring. Carotid ultrasound complements fundus imaging by providing a structural assessment of the carotid arteries, crucial for identifying risks of stroke and atherosclerosis through blood flow

[1]Department of Math and Computer Science, Weizmann Institute of Science, Rehovot, Israel
[2]Pheno.AI, Tel Aviv, Israel
[3]Department of Machine Learning, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

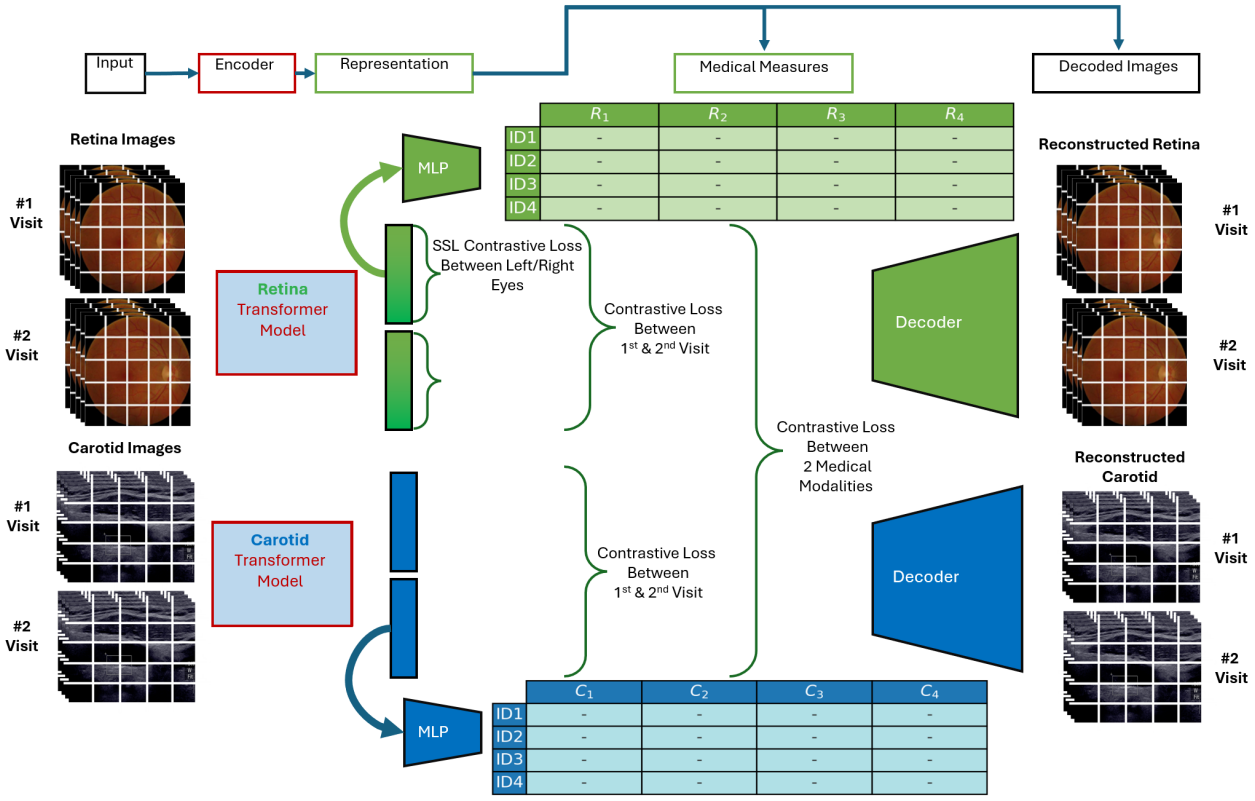

Figure 1: Schematic representation of COMPRER, our Contrastive Multi-objective Pretraining approach for Multi-modal Representation. We utilize ViT-Base encoders equipped with DINOV2 pre-trained weights for processing each imaging modality, accompanied by a linear projection head. Our method is defined by a combination of multiple loss objectives: (1) an ID-centric multi-modal contrastive loss that bridges features between fundus images and carotid ultrasound images; (2) a patient visit-based contrastive loss that discerns temporal discrepancies across repeat visits for each patient; (3) a contrastive scheme for the bilateral fundus images to ensure coupling of right and left eye data per patient; (4) a decoding objective to restore original images from condensed latent representations; and (5) a predictive mechanism to estimate general medical measures directly from modality-specific embeddings.

dynamics and plaque visualization (Yu et al., 2021; Siontis et al., 2021). The integration of both the fundus imaging, and carotid ultrasound modalities, offers a comprehensive representation of cardiovascular health, capitalizing on their individual strengths to assess conditions at various scales of the vascular system. However, the full potential of AI in medical imaging is challenged by the limited availability of large, annotated datasets necessary for traditional supervised learning (Zhou et al., 2023). This dataset scarcity is addressed by initiatives such as the Human Phenotype Project (HPP), which embraces a multi-modal deep-phenotyping approach, capturing a vast range of data modalities, from high-resolution images to comprehensive clinico-pathological records (Shilo et al., 2021). Such datasets are ideal for investigating and improving AI models that surpass existing boundaries in medical diagnostic capabilities. (Moor et al., 2023). To overcome these limitations self-supervised learning (SSL) has become a key tool in this field. (Huang et al., 2023). SSL circumvents the need for extensive labeled datasets by utilizing the data itself to derive informative features through the resolution of proxy tasks. It enables the extraction of significant patterns intrinsic to the data, fostering a model's ability to generalize robustly to unseen data (Grill et al., 2020; Chen et al., 2020). SSL with multi-modal data are particularly potent, with each modality enhancing the model's capabilities and utility across various health signals (Radhakrishnan et al., 2023).

2 Introduction

In this paper, we present COMPRER (COntrastive Multi-objective PREtraining for multi-modal Representation), a multi-modal, multi-objective pretraining framework that can be used in numerous downstream tasks and applications through the analysis of fundus imaging and carotid ultrasound. These include: diagnosing current patient diseases, predicting clinically significant medical features, and prognosing the probability of developing a medical condition in the future. As seen in Figure 1, COMPRER using a multi-modal approach, where it integrates distinct but complementary data sources— fundus imaging and carotid ultrasound imaging. Each modality offers a unique glimpse into the cardiovascular health of patients, capturing a diverse array of medical measures that, when combined, provide a comprehensive assessment framework. Our framework leverages ViTs (Dosovitskiy et al., 2020), specifically the DINOV2-Base pre-trained model (Oquab et al., 2023), as our architectural backbone, augmented by a multi-objective learning strategy that incorporates reconstruction via an image decoder, predictive heads, and contrastive learning losses.

2.1 COMPRER Training Objectives

In this section, we describe the training objectives utilized for our model. Our approach employs paired batches of fundus images and carotid ultrasound images to learn a joint embedding space, inspired by the CLIP training paradigm (Radford et al., 2021). This multimodal training maximizes the similarity of embeddings from matching image pairs and minimizes it for non-matching pairs within a batch.

The contrastive loss function specific to a set of two embeddings types u and v is defined as:

$$\mathcal{L}_{\text{contr}}(u, v) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp\left(\frac{\text{sim}(u_i, v_i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\text{sim}(u_i, v_j)}{\tau}\right)}\right)$$

where $\text{sim}(u, v) \equiv \frac{u^\top v}{\|u\|\|v\|}$ and N is the batch size. The contrastive loss is as in CLIP:
$\mathcal{L}_{\text{contr\_CLIP}}(u, v) = \frac{1}{2}\left(\mathcal{L}_{\text{contr}}(u, v) + \mathcal{L}_{\text{contr}}(v, u)\right)$
Several contrastive losses based on the relationship between the embeddings compared: For fundus (f) and carotid (c) image embeddings: $\mathcal{L}_{\text{contr\_fc}} = \mathcal{L}_{\text{contr\_CLIP}}(f, c)$.

For fundus images across different visits t and t': $\mathcal{L}_{\text{contr\_fv}} = \mathcal{L}_{\text{contr\_CLIP}}(f^t, f^{t'})$
For carotid images over time t and t': $\mathcal{L}_{\text{contr\_cv}} = \mathcal{L}_{\text{contr\_CLIP}}(c^t, c^{t'})$
For right (R) and left (L) eye fundus images: $\mathcal{L}_{\text{contr\_eye}} = \mathcal{L}_{\text{contr\_CLIP}}(f^R, f^L)$

Predictive accuracy for fundus and carotid embeddings is assessed with Mean Squared Error (MSE) losses:

$$\mathcal{L}_{\text{pred}}(m, \hat{m}) = \frac{1}{N}\sum_{i=1}^{N}(m_i - \hat{m}_i)^2$$

Yielding $\mathcal{L}_{\text{pred\_r}}$ for fundus with $N = N_r$, and $\mathcal{L}_{\text{pred\_c}}$ for carotid with $N = N_c$, with their respective medical measurements predictions. Finally, the total loss $\mathcal{L}$ combines individual contrastive and predictive components:

$$\mathcal{L} = \mathcal{L}_{\text{contr\_fc}} + \mathcal{L}_{\text{contr\_fv}} + \mathcal{L}_{\text{contr\_cv}} + \mathcal{L}_{\text{contr\_eye}} + \mathcal{L}_{\text{pred\_r}} + \mathcal{L}_{\text{pred\_c}},$$

2.2 Model Summary and Key Contributions

We evaluate COMPRER through validation across individual learning objectives and demonstrate its capabilities not only in extracting meaningful representations but also in projecting these representations into actionable clinical insights. Notably, we show that our multi-objective framework results in enhanced diagnostic and prognostic accuracy. Moreover, the model's ability to outperform not only a baseline pretrained DINOV2 but also dedicated models with substantial advantages in terms of parameters and data scale demonstrates the efficacy of our approach. In conclusion, our contributions are:

1. We introduce COMPRER, a novel deep learning framework that leverages multi-modal, multi-objective pretraining to forecast and predict the development of future diseases from medical imaging data.
2. We provide evidence for the efficacy of our modeling approach through an internal validation scheme, showing that our embeddings are capable of predicting medical measures with high R^2 scores.

3. We introduce a novel, understandable metric for assessing the performance of contrastive learning, offering an approach to measure the quality of embeddings in identifying correct image pairs across different modalities.
4. We substantiate the translational value of our model through its application in predicting cardiovascular health conditions, in both our cohort as well as in external cohorts.

3 Related Work

The combination of AI with healthcare represents a significant shift toward redefining clinical methodology and patient care. At the heart of this transformation, deep learning, particularly through convolutional neural networks (CNNs), has played a pivotal role in enhancing medical diagnostics. CNNs have demonstrated their efficacy in detection and classification tasks across a range of medical imaging modalities, including dermatology (Kwasigroch et al., 2020), radiology (Tiu et al., 2022; Jaiswal et al., 2019), and neuroradiology (Pereira et al., 2016). The profound pattern recognition capabilities of CNNs have thus become instrumental in medical image interpretation. Recently, the success of SSL methods, such as SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020), has redirected the focus from supervised learning reliant on extensive labeled datasets to SSL in extracting features from unlabeled data (Huang et al., 2023). SSL's resilience to dataset imbalances (Liu et al., 2021) particularly proves its adaptability in medical contexts, making it a cornerstone for foundation models, designed for broad application across multiple tasks (Bommasani et al., 2021; Moor et al., 2023). Zhou et al. (Zhou et al., 2023) exemplified this adaptability in a self-supervised masking strategy applied to a vast array of unlabeled fundus images, yielding a foundation model with an ability to perform disease detection across multiple scenarios. The breakthrough with OpenAI's Contrastive Language-Image Pretraining (CLIP) system has provided a novel perspective on utilizing versatile architectures, specifically transformers (Vaswani et al., 2017), for a wide range of modalities (Radford et al., 2021; Ramesh et al., 2022, 2021; Brown et al., 2020). CLIP's revolutionary approach to interpreting images through a natural language lens has revealed the potential of transformer architectures to tokenize and process multimodal data efficiently. Our COMPRER framework adopts a similar stance, leveraging the transformer architecture for both image encoders in a contrastive mechanism to align information across medical imaging types. The methodology rooted in CLIP's cross-modal learning inspired both the multi-visit and multi-modal contrastive losses of COMPRER, allowing it to not only decode spatial characteristics but also trace temporal patterns indicative of disease progression. In the era of multimodal data, significant strides have been made in cross-modal representations (Radhakrishnan et al., 2023) and feature extraction (Holmberg et al., 2020), revealing the intersecting pathways of SSL and multimodal methodologies. Complementing this trend are ViTs, which have remodeled the AI landscape with their extraordinary image processing capabilities (Dosovitskiy et al., 2020) and interpretability (Chefer et al., 2020). The advent of multi-task learning frameworks further amplifies these models' ability to assimilate diverse data and objectives, showcasing their robustness across a spectrum of tasks relevant to cardiovascular health analytics (Ruder, 2017; Crawshaw, 2020).

4 Methodology

In our study, we present a multimodal, multi-objective deep learning architecture designed to create a versatile pretrained model suitable for a wide range of health-related tasks. This architecture can generate adaptable embeddings for predicting a multitude of medical features or be fine-tuned for diverse medical applications. The model achieves this through the analysis of both fundus images and carotid ultrasounds. This part delves into the intricacies of our approach, which capitalizes on the robust capabilities of ViTs. Initially, we assembled a dataset encompassing approximately 11.5K participants' fundus and carotid ultrasound images, of which 1.5K have returned for a follow-up visit after two years. The dataset is divided into training (80% of data), validation (validation is 20% of the training set), and test sets (20% of the data), with the latter consisting solely of new participants arriving after the start of this research to ensure the integrity of our evaluation. Our preprocessing protocols ensure high-quality, artifact-free images using AutoMorph (Zhou et al., 2022) for fundus images and custom preprocessing to isolate relevant regions in carotid ultrasounds. Both image types are standardized to a resolution of 280x280 pixels, facilitating uniform processing where even grayscale ultrasound images are converted to three-channel format to align with the fundus images. The structural backbone of our model derives from the pretrained DINOV2-Base ViT, a vision transformer by Meta that has shown great performance in image representation tasks, as well as multiple vision downstream tasks (Caron et al., 2021; Oquab et al., 2023). As DINOV2 was trained on millions of images, we can capitalize on its extensive pretrained ability to represent image datasets and we can fine-tune it to our unique medical imaging context for enhanced efficiency. To complement the ViT, a linear projection head condenses high-dimensional embeddings to a more manageable state, serving a dual purpose: reducing computational demands and assisting with stability and shown to be essential by Balestriero et al. (2023). A transposed convolution neural network, comprising of transposed convolutional layers with gaussian error linear unit (GELU) activations, reconstructs the original images from latent embeddings, introducing a regularization effect that underpins the self-supervised learning within our framework. Lastly, a small 2-layer Multi-Layer Perceptron (MLP) is

added to predict medical measures from the latent space embeddings. To train the model we use multiple different optimization objectives simultaneously. These include the multimodal and multi-visit contrastive losses used as shown in CLIP by OpenAI (Radford et al., 2021), which fortify the model's time-awareness and cross-modality inference capabilities. A classic mean squared error (MSE) decoder loss ensures fidelity in image reconstruction, and another MSE loss is applied when predicting medical measures from the embeddings. We trained the model over four days, distributed across 8 NVIDIA A40 GPUs. We trained with an AdamW optimizer, with learning of 3×10^{-4} and weight decay of 0.5 with a StepLR scheduler, and a batch size of 9 per GPU. We found that the model does not overfit the train set in this time period, but due to limited resources we chose to run for only 4 days (even though the model might not have saturated the training set during this period). Addressing missing data, we introduce four parallel data loaders that guarantee optimal usage of the available data by including every sample where possible in the training, even when dealing with missing modalities or visit data. During training, we employ the validation set to discern the model's evolving accuracy. To evaluate performance on the validation set, we used different metrics for the different losses we employed. For the medical measures task, we relied on the R^2 to gauge the medical measure predictions from latent embeddings. For the decoding task, we relied on the decreasing MSE loss, as well as human evaluation of the resulting reconstructions. To evaluate the contrastive learning performance, we have crafted a novel metric, assessing the proximity of paired image embeddings and optionally adjusting for random chance, thus providing an intuitive measure of the model's learning. Essentially, by viewing the contrastive task as a classification task, we can view this metric as top-K accuracy.

Algorithm 1 Top-K Metrics for Contrastive Learning

procedure TOPK(sim_mat, k)
  $correct \leftarrow 0$
  for $i = 0$ **to** $|sim\_mat|$:
    if index i in top-k similar items **then**
      $correct \leftarrow correct + 1$
  return $correct$ / $|sim\_mat|$

$kVals \leftarrow [5, 25, 100, \ldots]$
$emb_i, emb_j \leftarrow$ embeddings
$norm_i \leftarrow$ l2_normalize(emb_i)
$norm_j \leftarrow$ l2_normalize(emb_j)
$cosSimMat \leftarrow norm_i \cdot norm_j^\top$
for $k \in kVals$:
  $rand\_base \leftarrow k$ / $|cosSimMat|$
  $metricScore \leftarrow$ TOPK($cosSimMat, k$)
  $angleTopK \leftarrow metricScore$
  $multAngleK \leftarrow metricScore/randBase$

In contrastive learning, the central goal is to learn representations such that similar or "paired" samples are brought closer together in the embedding space, while dissimilar samples are pushed apart. As we deal with batches of N samples, we inherently face an N-way classification problem during training. Achieving perfect performance is often challenging, and a binary assessment of model proficiency via top-1 prediction accuracy may not sufficiently capture the nuances in the embeddings the model has learned. In practice, it may appear that the model is underperforming when, in fact, it has developed a representation where correct matches are amongst the nearest neighbors, not necessarily the immediate first. By introducing a Top-K metric (Algorithm 1) specifically tailored for contrastive learning, we extend the single-label evaluation to a multi-neighbor perspective, which is analogous to considering a set of K nearest neighbors in the embedding space. Selecting different values of K enables us to explore the depth of the model's understanding of data relationships. Lower values of K can indicate fine-grained discriminatory power, while larger values suggest a broader comprehension of sample similarity. Furthermore, by adjusting for random chance in our metric — by dividing the raw Top-K score by the expected score under random matching — we gain insight into how much more effectively our model is at reconciling these pairs compared to a trivial random embedding model. Notably, this also allows us to use our hardware efficiently, as to avoid evaluating our model directly using the downstream tasks, we can evaluate the contrastive task performance at pretraining time.

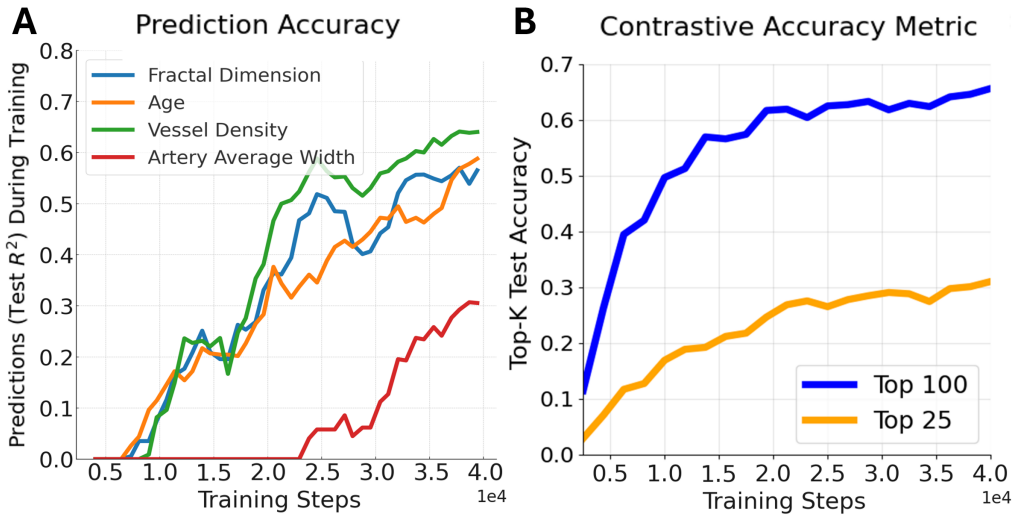

Figure 2: Interval Validation, **2.A** (left panel). Prediction accuracy of various medical measures over iterative steps. This figure illustrates the test R^2 values achieved during the training of predictive models for four different medical measures: fractal dimension, age, vessel density, and artery average width. The x-axis is the training steps iterations (in tens of thousands), while the y-axis indicates the R^2 value observed on held-out test data. **2.B** (right panel). Evolution of Top-K Test Accuracy in Contrastive Learning for Multimodal Image Matching. This figure depicts the test accuracy of a contrastive learning model for two different values of k: 25 and 100, as labeled Top-25 (orange line) and Top-100 (blue line), respectively. The x-axis is, again, training steps iterations, and the y-axis denotes the Top-K Test Accuracy. The Top-100 accuracy increases as training goes on, reaching top performance of 0.65. Similarly, the Top-25 accuracy ends around 0.35.

5 Results

5.1 Internal Validation

Our exploration of COMPRER's results began with internal validation metrics. Interval validation metrics are composed of metrics that validate the pretraining phase of our model with multiple objectives. For each one of these objectives, we report their score on a held-out test set.

5.1.1 Medical Measures Prediction

To evaluate the generalization capabilities of the medical measures prediction head of COMPRER, we engaged in predicting measures from the test set that have direct clinical applicability. Among the predicted medical measurements were age, fundus image fractal dimension, vessel density, and artery average width. Predicting age from fundus images is particularly intriguing, as it suggests a correlation between ocular characteristics and biological aging, which can have various medical implications (Ahadi et al., 2023). The Fundus Image Fractal Dimension is a measure of the complexity and branching patterns of the retinal vasculature, indicative of overall vascular health(Dinesen et al., 2021; Macgillivray et al., 2007) . Vessel Density refers to the proportion of the retina occupied by blood vessels, a crucial factor in assessing retinal and systemic circulatory health. Artery Average Width provides insights into vascular caliber, important for understanding cardiovascular risks. The ability to predict these measures from fundus images is noteworthy, indicating that our model retains spatial understanding of the images despite the multiple objectives enforced on this image representation. As detailed in Figure 2, the generalization performance, quantified by R^2 score — showed differing degrees of success across the medical measures. An R^2 score of approximately 0.6 was observed for most medical measures, indicating a meaningful predictive relationship between the learned representations and the clinical measures. However, the prediction for Artery Average Width presented a lower R^2 score of around 0.3, signifying a less robust prediction capability, or a harder prediction task.

5.1.2 Evaluation of Contrastive Learning

A critical component of the COMPRER framework is the multimodal contrastive loss, which plays a pivotal role in aligning features across distinct imaging modalities—namely, fundus and carotid ultrasound images. In fact, in the

whole multiple objective framework, this loss is the only connection between the two distinct image encoders. To measure the effectiveness of contrastive learning, we devised a scaleless, interpretable metric that provides a concrete understanding of model performance. While the loss term itself provides an indication of model learning during training, it lacks direct translatability to practical outcomes. In Figure 3.a, we show the Top-K test accuracy of $k \in [5, 25, 100]$, and in Figure 3.b we show the same plot for the multiplicative metric, which highlights the model getting better than random performance. The plots show the metric as calculated from algorithm 1. The model exhibited non-trivial, non-random results in the multimodal matching task, which is notable considering the inherent challenge of this problem - one that even skilled clinicians do not typically address. As part of our experimental setup, we developed an ablation model, the Multi-Modal Contrastive Learning (MMCL) model, which was trained on the same data as COMPRER. Unlike COMPRER, MMCL was trained using a single objective with pretraining - focusing exclusively on multi-modal contrastive learning. This approach allowed us to evaluate the impact of the other objectives on our main model's performance. Interestingly, the COMPRER model outperforms the Multi-Modal Contrastive Learning (MMCL) model, which has trained on the same data, with only the multi-modal contrastive loss, in multimodal matching accuracy, emphasizing the advantage of a multiple objective training strategy. In figure 3.b we see that all start off with random

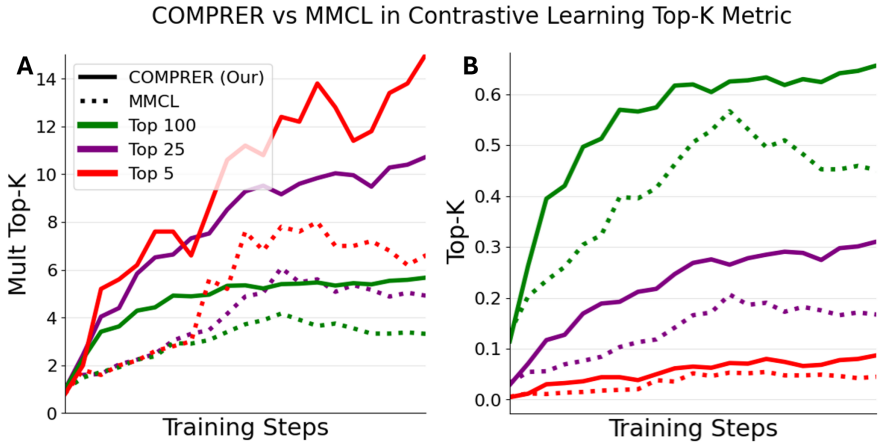

Figure 3: Comparative Analysis of Top-K Contrastive Metric on COMPRER and MMCL. We see in green, purple, and red the Top-100, Top-25, Top-5 contrastive metric respectively. We see in a dotted line the MMCL model, which trained solely on the multi-modal contrastive loss (trained using only $\mathcal{L}_{\text{contr\_rc}}$), and in the straight line COMPRER which trained on multiple objectives, including the multi-modal contrastive loss. In **3.A** We see the multiplicative tok-K metric, and in **3.B** we see the top-K metric. We see that COMPRER consistently outperforms MMCL.

performance (mult top-K score of 1 for all K) , and rise as the optimization starts. In COMPRER, the observed Top-100 accuracy reached 0.65, which is higher than the baseline set by MMCL of 0.56. For the more stringent Top-25 accuracy, COMPRER obtained an accuracy of 0.35, while the MMCL achieved a lower score of 0.2. **This difference not only underpins our model's superior matching capability but also implies the potential benefits of multi-objective training in augmenting the feature space for more nuanced discriminatory powers.**

5.1.3 Image Reconstruction Capability

Both fundus image and carotid image reconstructions, while losing some fine details, remain structurally similar to the original general structure, an encouraging sign for the model's comprehension of microvascular features. While select losses in high-frequency details were observed—likely attributable to the inherent information compression within the network—the structural integrity was maintained.

5.2 Predictive Performance on Cardiovascular Conditions in the HPP Cohort

While internal validation schemes during the pretraining phase provide essential insights into the immediate learning dynamics of our COMPRER model, the true test of its effectiveness lies in its clinical application. Thus, our goal is to demonstrate that our model pretraining not only captures intricate data patterns but also translates into significant improvements in real-world clinical diagnostics. To this end, we focused our attention on fine-tuning COMPRER to predict cardiovascular health conditions. Figure 4 shows the model's capacities in both a diagnostic and prognostic context, providing valuable insights into cardiovascular health. The performance metrics presented in Figure 4 were constructed using the models, which was fitted with a 1-layer MLP regression head. To find appropriate hyperparameters,

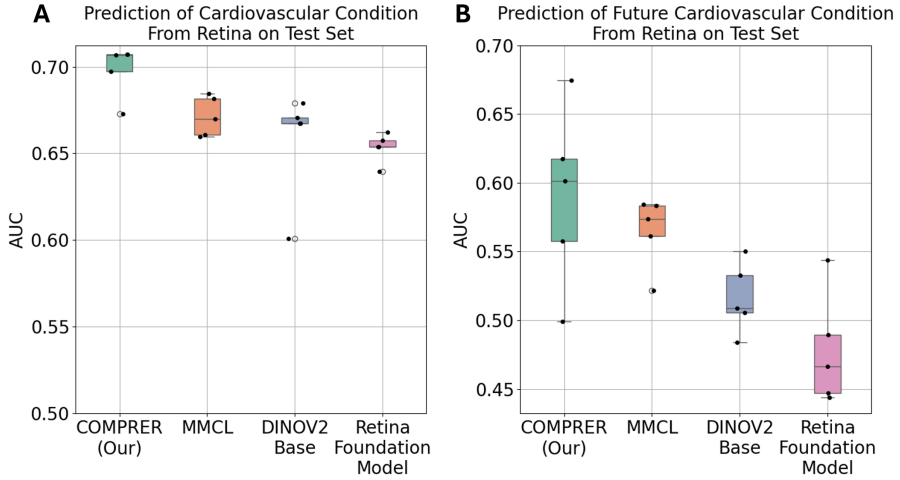

Figure 4: Comparative Analysis of Model Performance in Predicting Cardiovascular Conditions from Fundus Images. **4.A** illustrates the AUC results for various models fully finetuned to predicting current cardiovascular condition based on fundus images taken at the participants' first visit. The models compared include, our proposed COMPRER, MMCL, DINOV2-Base, and a Retina Foundation Model. The boxplots represent the distribution of AUC scores achieved on the test set after selection of top 5 hyperpameters for each model based on validation performance. **4.B** depicts the AUC results for the same set of models, but focuses on predicting the future onset of cardiovascular conditions. This analysis only includes participants who were healthy at baseline, with the goal of determining if the models can predict the development of cardiovascular conditions at a follow-up visit. While the AUC scores on their own are not high, COMPRER consistently outperforms the competition. It's also worth observing that MMCL has also some gains with respect to DINOV2 and the retina foundation model, which goes to show that even just applying multi-modal contrastive learning on its own can increase downstream performance.

we employed a systematic hyperparameter search on the validation set, from which the top 5 models of each type were identified. These leading models were then assayed on an independent test set, yielding a distribution of results, which denotes the robustness and consistency of performance across model instances. The MMCL model is an ablation model, representing a version of our architecture and data trained with only multimodal contrastive loss, to provided a baseline to quantify the value added by the multi-objective learning. The Retina Foundation Model embodies a high-parameter (300M parameters, which is $3.5\times$ larger than all other competitor models) alternative, leveraging a considerably larger latent space (1024, which ≈ 1.3 times larger than all other competitors) and trained on an extensive dataset of 1.6M fundus images (which is $75\times$ larger than our fundus dataset). Despite these advantages of the retina foundation model, COMPRER shows superior performance. We also observed that on the prognosis task (figure 4.b) almost all model runs except COMPRER's are random. This is interesting because COMPRER is the only model that had in its pretraining any signal of future events - based on the temporal contrastive learning objective.

5.3 Performance on an OOD External Cohort

Validating the predictive power of a model on an out of distribution (OOD) dataset is often considered the gold standard for demonstrating the real-world applicability and robustness of a predictive framework. In this section, we evaluate COMPRER's performance on the external dataset - the UK Biobank (UKBB), an extensive, well-characterized external cohort that has been at the forefront of large-scale biomedical research (Sudlow et al., 2015). After data filtering and cleaning, it comprises of 44K participants with fundus images. The Retina Foundation model has previously showcased its traction on this dataset, establishing a performance benchmark for the field. Figure 5 illustrates the comparison between COMPRER and the Retina Foundation model in predicting various cardiovascular and related diseases from fundus image representations. Our approach demonstrates competitive, if not superior, predictive performance across numerous conditions. In the realm of ischaemic stroke prediction, COMPRER conspicuously outperforms the Retina Foundation model, denoting a higher AUC value. These results affirm the appropriateness of COMPRER's multi-modal, multi-objective pretraining paradigm, reinforcing its utility in extracting salient features pertinent to disease states from medical imagery. Moreover, they underscore our method's efficiency; by achieving these competitive performance metrics, COMPRER evidences that well-conceived model architectures coupled with sophisticated pretraining strategies

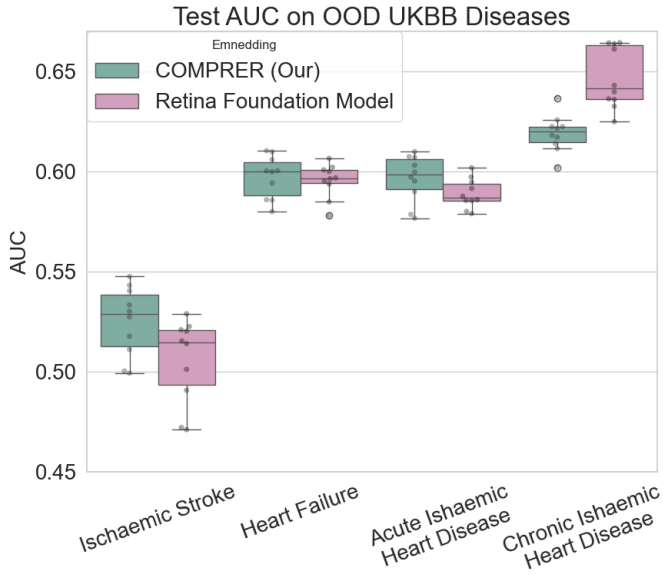

Figure 5: Comparative Analysis of COMPRER and Retina Foundation Model Performance in Predicting Diseases from Fundus Image Representations on an External Dataset (UKBB). The prediction is done from frozen image embeddings using logistic regression. We see that COMPRER outperforms the Retina Foundation model in the prediction of ischaemic stroke, evidenced by a higher AUC value. Conversely, the Retina Foundation model exhibits superior performance in predicting chronic ischaemic heart disease, although COMPRER also shows a test AUC above 0.6. For the conditions of stroke, heart failure, and acute ischaemic heart disease, the COMPRER method slightly outperforms the Retina Foundation model, though the margins are not significant. Its also notable that the the retina foundation model was trained using $75\times$ more fundus images, and with a $3.5\times$ larger model.

can level the playing field against models with ostensibly more advantageous training conditions (using $75\times$ the data, and $3.5\times$ the parameters).

6 Discussion

We presented COMPRER, a novel pretraining method targeted at extracting medical features, specifically aimed at the enhancement of downstream tasks such as disease diagnosis and prognosis. Our multi-modal, multi-objective pretraining approach bridges the gap between advancing machine learning architectures and their practical applications in clinical settings, exhibiting superior predictive abilities compared to specialized models. In the realm of image representation, COMPRER has demonstrated the ability to distill meaningful insights from a large-scale, longitudinally derived dataset, diminishing the reliance on arduously curated labels. The empirical advancements evidenced by robust internal validation and insightful applications in disease prediction advocate for COMPRER's adoption. Notably, the model provides competitive performance, transcending the need for vast data volumes and extensive computational resources, highlighting a quality-centric approach in medical data analytics. We have shown that the integration of multiple objectives stretches the training duration but enriches the model's feature representations. This comprehensive approach, though intensive, equips the model with a deepened understanding and enhanced performance that may surpass models optimized for single objectives alone. The superior performance in multimodal contrastive matching demonstrates this phenomenon, suggesting that a more holistic, integrated training regimen can indeed yield models with robust generalization capacities across various tasks.

In addition, we have shown that COMPRER is capable of achieving higher test AUC scores than models that have been trained with an order of magnitude more data, and with larger models. We have shown this both in our HPP cohort, as well as, in an external cohort - the UKBB, which shows our model can generalize out of distribution, across diverse populations from different continents. Interestingly, within the OOD UKBB dataset, a notable divergence in the test AUC was observed for ischaemic stroke prediction between COMPRER, and the Retina Foundation model. This finding is particularly intriguing given that ischaemic stroke is a condition diagnosable through carotid ultrasound (Zhang et al., 2014). It is worth noting that COMPRER was trained using both fundus images and carotid ultrasound, suggesting that this multi-modality approach may have played a role in the increase of COMPRER's predictive accuracy compared to

the Retina Foundation model. It is also noteworthy that in the internal cohort validation section (5.2), where we showed future prognosis performance, the majority of model runs, with the exception of COMPRER, exhibit random test AUC scores. This phenomenon might be interesting, as it underscores the distinctive attribute of COMPRER's pretraining methodology. In contrast to the other models, COMPRER benefits from the inclusion of a temporal information during its pretraining, notably - the temporal, visit-based, contrastive learning objective. This objective provided the model with insights into future events within the longitudinal data. We hypothesize that this unique aspect of COMPRER's training regimen is contributing to its superior performance, even surpassing models like MMCL that have encountered the same data but lack exposure to the temporal, visit-based contrastive loss. This finding underscores the potential advantages of incorporating temporal information in pretraining, shedding light on the nuances of disease prognosis prediction and highlighting the efficacy of our approach.

However, our work has several caveats, calling attention to the simultaneous challenges and potential trajectories for improvement. Dataset scope and representativeness remain pivotal for model generalization. While the HPP dataset underpins our current findings, incorporating datasets with broader demographic diversity is essential for enhancing model robustness and ensuring its translational relevance across patient populations. A more diverse dataset would help mitigate bias and uphold the model's diagnostic integrity, especially when encountered with dataset shifts in real-world scenarios. The generalizability of COMPRER across different diseases and imaging modalities is another frontier to be explored. Expanding the disease spectrum and experimenting with a variety of modalities are key to consolidating the framework's applicability in diverse medical contexts. Moreover, the limited computational resources constrained our model's training to 40k steps, hinting at the potential for further refinement. Investments in computational infrastructure and collaborative efforts could uncover latent performance enhancements and insights into the optimization dynamics of our model. Interpretability and explainability are indispensable for clinician and patient acceptance. Despite the strides made with our transformer-based model, elucidating the AI's decision-making processes remains crucial. Validation in clinical environments could unravel the efficacy and adaptability of our model and solidify its role within clinical workflows. It is important to note that we conducted our research on only two modalities, however, we do see how this method could easily extend beyond only two modalities. In the HPP dataset, we have access to a rich variety of over 20 distinct data modalities, ranging from visual information and time series data to textual records and tabular measurements.

We recognize the untapped potential of leveraging multiple modalities within our pretraining scheme. Future iterations of the methods we described here can harness this diverse data landscape by incorporating additional losses for multi-modal contrastive learning, introducing multi-visit losses that span across these various modalities, and exploring other innovative techniques. This multi-modal approach holds promise in further enriching the model's understanding of complex medical data, potentially leading to even more robust generalization across a wide array of clinical tasks.

7 Availability of Code and Model Weights

In the interest of transparency and facilitating future research, we plan to release the code and model weights associated with our study upon publication.

8 Bibliographical References

References

Sara Ahadi, Kenneth A Wilson, Boris Babenko, Cory Y McLean, Drew Bryant, Orion Pritchard, Ajay Kumar, Enrique M Carrera, Ricardo Lamy, Jay M Stewart, Avinash Varadarajan, Marc Berndl, Pankaj Kapahi, and Ali Bashir. 2023. Longitudinal fundus imaging and its genome-wide association analysis provide evidence for a human retinal aging clock. *eLife*, 12:e82364.

F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova, and M. Prunotto. 2019. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *npj Digital Medicine*, 2:92.

J. Bajwa, U. Munir, A. Nori, and B. Williams. 2021. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthcare Journal*, 8:e188–e194.

R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, et al. 2023. A cookbook of self-supervised learning.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, et al. 2021. On the opportunities and risks of foundation models.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al. 2020. Language models are few-shot learners.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, et al. 2021. Emerging properties in self-supervised vision transformers.

H. Chefer, S. Gur, and L. Wolf. 2020. Transformer interpretability beyond attention visualization.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. 2020. A simple framework for contrastive learning of visual representations.

M. Crawshaw. 2020. Multi-task learning with deep neural networks: A survey.

L. Dai, L. Wu, H. Li, C. Cai, Q. Wu, H. Kong, et al. 2021. A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nature Communications*, 12:3242.

Sebastian Dinesen, Pia S Jensen, Maria Bloksgaard, Søren Leer Blindbæk, Jo De Mey, Lars M Rasmussen, Jes S Lindholt, and Jakob Grauslund. 2021. Retinal vascular fractal dimensions and their association with macrovascular cardiac disease. *Ophthalmic Research*, 64(4):561–566.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29.

J-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning.

O. G. Holmberg, N. D. Köhler, T. Martins, J. Siedlecki, T. Herold, L. Keidel, et al. 2020. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. *Nature Machine Intelligence*, 2:719–726.

S-C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari. 2023. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6:74.

A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. P. C. Rodrigues. 2019. Identifying pneumonia in chest x-rays: A deep learning approach. *Measurement*, 145:511–518.

A. Kwasigroch, M. Grochowski, and A. Mikołajczyk. 2020. Self-supervised learning to increase the performance of skin lesion classification. *Electronics*, 9:1930.

H. Liu, J. Z. HaoChen, A. Gaidon, and T. Ma. 2021. Self-supervised learning is more robust to dataset imbalance.

T J Macgillivray, N Patton, F N Doubal, C Graham, and J M Wardlaw. 2007. Fractal analysis of the retinal vascular network in fundus images. *Annu Int Conf IEEE Eng Med Biol Soc*, 2007:6456–6459.

M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, et al. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, et al. 2023. Dinov2: Learning robust visual features without supervision.

S. Pereira, A. Pinto, V. Alves, and C. A. Silva. 2016. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Transactions on Medical Imaging*, 35:1240–1251.

R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, et al. 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2:158–164.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al. 2021. Learning transferable visual models from natural language supervision.

A. Radhakrishnan, S. F. Friedman, S. Khurshid, K. Ng, P. Batra, S. A. Lubitz, et al. 2023. Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nature Communications*, 14:2436.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. 2022. Hierarchical text-conditional image generation with clip latents.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, et al. 2021. Zero-shot text-to-image generation.

S. Ruder. 2017. An overview of multi-task learning in deep neural networks.

S. Shilo, N. Bar, A. Keshet, Y. Talmor-Barkan, H. Rossman, A. Godneva, et al. 2021. 10 k: a large-scale prospective longitudinal study in israel. *European Journal of Epidemiology*, 36:1187–1194.

K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman. 2021. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18:465–478.

J. D. Spence. 2006. Technology insight: ultrasound measurement of carotid plaque–patient management, genetic research, and therapy evaluation. *Nature Clinical Practice Neurology*, 2:611–619.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. 2015. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779.

E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6:1399–1406.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al. 2017. Attention is all you need.

Q. Yan, D. E. Weeks, H. Xin, H. Huang, A. Swaroop, E. Y. Chew, et al. 2019. Deep-learning-based prediction of late age-related macular degeneration progression. *medRxiv*.

J. Yu, Y. Zhou, Q. Yang, et al. 2021. Machine learning models for screening carotid atherosclerosis in asymptomatic adults. *Scientific Reports*, 11:22236.

Y. Zhang, E. Guallar, Y. Qiao, and B.A. Wasserman. 2014. Is carotid intima-media thickness as predictive as other noninvasive techniques for the detection of coronary artery disease? *Arteriosclerosis, Thrombosis, and Vascular Biology*, 34(7):1341–1345. Epub 2014 Apr 24.

Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, et al. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622:156–163.

Yukun Zhou, Siegfried K Wagner, Mark A Chia, An Zhao, Moucheng Xu, Robbert Struyven, Daniel C Alexander, Pearse A Keane, et al. 2022. Automorph: Automated retinal vascular morphology quantification via a deep learning pipeline. *Translational vision science & technology*, 11(7):12–12.