Supervised Temporal Graph Networks and Self-Supervised Vision Models for Clinical Risk Prediction

Gavriel David Hannuna [1]

Supervisors:
Prof. Niko Beerenwinkel, ETH Zurich, Department of BioSystems Science and Engineering (D-BSSE).
Prof. Eran Segal, Weizmann Institute of Science, Department of Applied Mathematics and Computer Science.

Abstract

Can we unify diverse healthcare data modalities to predict chronic disease onset before clinical symptoms emerge? This thesis addresses this fundamental challenge in preventive medicine by exploring two methodological pathways toward unified early prediction systems: (1) temporal graph-based integration of structured patient data, and (2) self-supervised learning architectures with inherent multi-modal unification capabilities.

The first approach constructs 3-day heterogeneous graphs representing patient trajectories by integrating glucose-derived physiological signatures, sleep analysis embeddings, general medical information, and diagnostic events to predict disease onset across seven conditions including diabetes, hypertension, and cardiovascular disease. Graph neural networks with attention mechanisms and Cox survival modeling were employed for short- and long-term onset prediction. Despite systematic engineering of temporal dynamics and node relationships, this framework achieved limited predictive accuracy, exposing fundamental challenges in graph-based modeling of sparse, noisy clinical data. The second approach leverages the Image Joint Embedding Predictive Architecture (I-JEPA), selected specifically for its ability to learn unified representations across diverse data modalities through joint embedding spaces. While this proof-of-concept focused on retinal fundus images for cardiovascular risk prediction, I-JEPA's architecture enables seamless integration of multiple data types within a single predictive framework. Through Low-Rank Adaptation (LoRA) fine-tuning, I evaluated model performance on external datasets including Messidor and IDRiD. Although results did not surpass established diabetic retinopathy detection models, the framework successfully captured clinically relevant retinal features, validating I-JEPA's potential as a foundation for unified healthcare prediction systems.

The findings suggest that while temporal graph approaches face substantial obstacles with real-world clinical data, self-supervised architectures like I-JEPA offer a promising pathway toward truly multi-modal clinical prediction systems.

1. Introduction

Chronic disease prediction remains a significant challenge in preventive healthcare (Khera et al., 2024; Armoundas et al., 2024). Early detection of conditions like cardiovascular disease or diabetes can markedly improve patient outcomes (American Diabetes Association Professional Practice Committee, 2025), yet conventional clinical indicators often emerge too late, while patients frequently delay response to concerning symptoms.

AI-powered models combined with continuous health monitoring offer substantial promise (Shajari et al., 2023; Daniore et al., 2024) through two distinct paradigms. The first leverages temporal modeling of structured patient data using graph-based approaches to capture relational and longitudinal patterns in electronic health records (Boll et al., 2024a). Graph neural networks (GNNs) trained on patient similarity graphs from datasets like MIMIC-III (Johnson et al., 2016) have shown encouraging performance in outcome forecasting, such as heart failure prediction (Boll et al., 2024b), with interpretability benefits through attention mechanisms (Choi et al., 2017). However, recent surveys highlight persistent challenges that include irregular sampling, heterogeneous feature spaces, and limited generalization across diverse populations.

The second paradigm exploits deep learning on retinal fundus imaging as a non-invasive diagnostic approach. Large-scale investigations, notably from Google and UK Biobank collaborations, reveal that vascular and anatomical features in fundus photographs can predict cardiovascular risk factors, including biological age, blood pressure, smoking status, and major adverse cardiac events, with accuracy rivaling

traditional clinical assessments (Poplin et al., 2018; Qian et al., 2024; Rim et al., 2021).

Despite this progress, significant gaps persist in both paradigms. Temporal patient modeling approaches face evaluation limitations, as existing graph-based methods are typically tested on well-curated datasets with dense, uniform sampling, conditions rarely met in real-world clinical settings where measurements are sparse, noisy, and irregular (Tipirneni et al., 2022; Boll et al., 2024a). Additionally, clinical trajectories exhibit dynamic temporal changes that prove difficult to capture with static or semi-static graph structures (Longa et al., 2023). For retinal imaging, state-of-the-art models depend heavily on supervised learning with extensive labeled datasets (Poplin et al., 2018; Rim et al., 2021), and high-resolution medical images impose computational constraints that often force either aggressive downsampling or costly hardware.

Self-supervised learning presents novel solutions to these challenges. In imaging domains, architectures like the Image Joint Embedding Predictive Architecture (I-JEPA) facilitate context-based representation learning by predicting latent representations of missing image regions from visible areas (Assran et al., 2023). This approach aligns naturally with medical imaging characteristics, where high spatial redundancy and localized pathology patterns provide rich self-supervised signals (Huang et al., 2023). Combined with efficient fine-tuning strategies such as Low-Rank Adaptation (LoRA), these models can be adapted to specialized clinical domains despite hardware and dataset limitations (Hu et al., 2021).

This thesis investigates both directions:

- Temporal heterogeneous graph modeling to represent and predict patient trajectories from longitudinal glucose monitoring data, sleep monitoring data, general health data (age, sex, BMI, and existing medical conditions), and diagnosis histories.
- I-JEPA fine-tuning for learning predictive retinal representations for cardiovascular risk assessment under realistic computational constraints.

The graph-based approach reveals methodological and practical challenges in applying GNNs to sparse, heterogeneous longitudinal health data, identifying critical limitations for future research (Tipirneni et al., 2022; Boll et al., 2024a). The imaging approach demonstrates successful adaptation of self-supervised vision models for high-resolution medical imaging within standard GPU constraints (Assran et al., 2023; Hu et al., 2021). Together, these complementary investigations advance the development of robust, data-efficient, and clinically viable predictive models for preventive healthcare.

This study utilizes data from the Human Phenotype Project (HPP), a large-scale prospective cohort study in Israel mapping health-to-disease transitions. HPP participants undergo comprehensive baseline phenotyping, incorporating medical history, anthropometric measurements, and lifestyle assessments; extensive physiological tests (two-week continuous glucose monitoring, three-night sleep studies, ECG, ankle-brachial index, DEXA, liver ultrasound); multi-omic profiling (genomics, transcriptomics, proteomics, metabolomics, microbiome); and high-resolution retinal fundus imaging - with long-term biospecimen storage and linkage to national health registries allowing decades-long follow-up (Shilo et al., 2021; Reicher et al., 2025). This unique combination of longitudinal follow-up, diverse data modalities, and high temporal resolution creates an exceptional platform for investigating complementary disease risk modeling strategies (Reicher et al., 2025). It is worth noting that part of the information in this project is **self-reported** through a mobile application, specifically, a diagnosis and its date, and dietary intake. This will inevitably affect the precision of our models.

From HPP's comprehensive dataset, I focused on three key modalities: CGM-derived physiological signatures combined with sleep analysis and medical history data for temporal graph construction, and high-resolution retinal photographs for I-JEPA adaptation (Shilo et al., 2021; Reicher et al., 2025). The graph-based approach examines the feasibility of modeling real-world clinical data complexity using temporal GNNs, while the imaging approach explores self-supervised representation learning for clinical feature extraction from large medical images (Assran et al., 2023).

Together, these complementary directions reflect the broader vision of building a multimodal framework capable of unifying structured and unstructured health data over time. Such a framework could capture both the dynamic progression of physiological states and the latent risk factors visible through imaging, offering a more holistic approach to chronic disease prediction. While the temporal graph approach achieved limited predictive success, it generated valuable insights into the challenges of applying sophisticated graph architectures to sparse clinical datasets (Tipirneni et al., 2022; Boll et al., 2024a). The I-JEPA adaptation, conversely, demonstrated promising performance under practical constraints (Assran et al., 2023; Hu et al., 2021), highlighting the potential of context-predictive vision transformers within a multimodal preventive healthcare pipeline.

1.1. Related Work - GNN

Temporal graph modeling has emerged as a promising approach for healthcare AI, particularly for capturing evolving patient health trajectories. Unlike static representations,

temporal graphs model dynamic changes in node and edge relationships over time, a critical capability for clinical data analysis where patient states continuously evolve. These networks consist of time-activated edges that enable analysis of temporal paths, reachability, and causality within dynamic systems (Holme & Saramäki, 2012; Skarding et al., 2021; Rossi et al., 2020).

In biomedical domains, temporal GNNs have shown potential across various prognostic tasks. Notable applications include Alzheimer's disease progression prediction using interpretable GNNs built on longitudinal neuroimaging data, successfully integrating graph structure with temporal dynamics (Kim et al., 2021). Boll et al. provide a comprehensive review of GNN applications in clinical risk prediction using electronic health records, demonstrating how graph models capture complex relational structures while highlighting persistent challenges with data sparsity and irregular temporal sampling in medical records (Boll et al., 2024a).

Recent advances have expanded temporal modeling capabilities. Researchers have formulated EHR data as temporal heterogeneous graphs, representing visits and medical events as distinct node types within dynamic structures, enabling more sophisticated clinical trajectory modeling (Chen et al., 2024). STEM-GNN exemplifies cross-disciplinary innovation by capturing spatial and temporal dependencies for multivariate time series forecasting, offering insights applicable to clinical temporal modeling (Cao et al., 2020). In a healthcare-specific implementation, temporal graph convolutional networks predicted hip replacement procedures one year in advance using primary care event codes, achieving 0.724 AUROC and demonstrating the feasibility of long-term medical procedure forecasting (Hancox et al., 2024).

Self-supervised strategies have shown particular promise in addressing the limitations of clinical data. Lu et al. developed models leveraging hyperbolic embeddings and multi-level attention mechanisms for pre-training medical code representations, improving disease complication prediction through hierarchical healthcare data structures (Lu et al., 2023). Hybrid architectures combining LSTM networks with relational GNNs, such as LIGHTED, enhance both interpretability and temporal modeling capabilities in clinical settings (Dong et al., 2022).

Beyond individual patient modeling, temporal graphs demonstrate utility in population-scale epidemiological forecasting. HeatGNN embeds epidemiological mechanisms into heterogeneous, time-varying graphs for epidemic trend prediction (Zheng et al., 2024), while dual-topology GNNs integrate geographical and functional networks for influenza-like illness forecasting, achieving competitive performance at fine spatio-temporal resolutions (Luo et al., 2025).

Despite these advances, clinical adoption remains constrained by the inherent complexity, irregularity, and sparsity of real-world health trajectories. Existing approaches typically evaluate on well-curated datasets with regular sampling patterns that poorly reflect clinical reality. This work addresses these limitations through a tailored temporal heterogeneous graph approach, incorporating glucose-derived physiological signatures, sleep-derived embeddings, longitudinal patient embeddings, and diagnostic events to assess both the potential and practical constraints of temporal graph modeling for disease risk prediction.

1.2. Related Work - IJEPA

Self-supervised learning has transformed visual representation learning by moving beyond traditional augmentation-based strategies toward more sophisticated predictive frameworks (Huang et al., 2023). The JEPA represents a significant advancement in this direction, building world models by training representations to predict latent embeddings of missing image regions from visible context (LeCun, 2022). This predictive learning mechanism aligns more closely with theoretical models of human cognition while enabling semantically meaningful embedding acquisition without pixel-level reconstruction or handcrafted transformations (LeCun, 2022).

The Image-based JEPA (I-JEPA), formally introduced by Assran et al., provides the first concrete implementation of JEPA principles for computer vision. The architecture strategically divides image patches into context and target blocks: a context encoder processes visible portions, a predictor infers embeddings of masked target patches, and a target encoder (updated via exponential moving average) generates reference embeddings for supervision. I-JEPA demonstrates impressive scalability and representation quality: for example, a ViT-H / 14 model pre-trained on ImageNet in less than 72 hours achieved strong performance across diverse downstream tasks, including classification, object counting, and depth estimation (Assran et al., 2023). The approach excels in both efficiency and semantic generalization, representing a paradigm shift from traditional self-supervised learning pipelines.

Several JEPA variants have emerged, extending the framework across modalities and addressing technical limitations. MC-JEPA integrates motion and content learning for video processing through unified encoders capturing both static and dynamic information (Bardes et al., 2023). V-JEPA applies predictive objectives across spatiotemporal frames for large-scale video representation learning without labels or contrastive pairs (Drozdov et al., 2024). In neuroscience applications, Brain-JEPA leverages JEPA principles for modeling functional brain dynamics from fMRI data, achieving strong prognosis and demographic prediction via gradient

positioning and spatiotemporal masking (Dong et al., 2024).

Recent refinements address limitations in the original I-JEPA framework. C-JEPA (Contrastive-JEPA) tackles instability and collapse risks in EMA-target setups by integrating VICReg-inspired variance–invariance–covariance regularization, improving convergence speed and embedding quality on ImageNet benchmarks (Mo & Tong, 2024; Bardes et al., 2021). CNN-JEPA adapts JEPA mechanisms to convolutional architectures using sparse encoders and convolutional predictors, reporting competitive results to ViT-based I-JEPA with reduced compute and minimal augmentations (Kalapos & Gyires-Tóth, 2024).

JEPA variants show particular promise where image acquisition is expensive or labeling is limited, common in medical imaging. By predicting in latent space, these models learn semantic and structural representations from raw inputs while remaining compute-efficient and modality-flexible (Huang et al., 2023). Despite this potential, JEPA architectures remain underexplored in healthcare, representing a significant opportunity for clinically oriented foundation models.

2. Models and Methods

2.1. Temporal Graph Construction and Input Features

We model patient health trajectories as discrete-time sequences of heterogeneous graphs $\{G_t\}_{t=1}^{T}$, where each G_t represents a three-day sliding window. The index t denotes the window's starting day, so G_t encompasses days $t, t+1, t+2$ inclusively. This design balances temporal resolution with graph density, enabling each snapshot to capture short-term physiological trends while mitigating single-day measurement noise.

Each graph consists of three node families: patient nodes, representing individuals active within the window; data group nodes, which can in principle encode any quantifiable medical information about patients but in this work are restricted to glucose-signature nodes derived from Non-negative Matrix Factorization (NMF) of three-day CGM segments; and fixed medical-condition nodes, representing target diagnoses. Edges capture patient–signature relationships, temporal continuity across consecutive windows, and condition-onset events.

CGM Processing and Glucose-Signature Nodes. To derive canonical glucose signatures, we aggregated all available continuous glucose monitoring (CGM) signals from the cohort into a single large matrix

$$X \in \mathbb{R}^{(N\times 96)}, \tag{1}$$

where N denotes the total number of patient-day segments (on the order of 10,000) and 96 represents the 15-minute intervals in a day. We applied Non-negative Matrix Factorization (NMF) with $K = 12$ components (Lee & Seung, 1999):

$$X \approx WH, \quad W \in \mathbb{R}_{\geq 0}^{96\times K}, \quad H \in \mathbb{R}_{\geq 0}^{K\times N}. \tag{2}$$

Here, the columns of W correspond to global one-day glucose *signatures*, while the columns of H contain patient- and day-specific activations. The basis W was chosen by optimizing reconstruction error and subsequently frozen for downstream modeling. In the temporal graphs, signature nodes $\{s^{(k)}\}_{k=1}^{K}$ are instantiated per window with features given by the corresponding signature $w^{(k)}$. Patient-to-signature edges are weighted by normalized activations: $\tilde{h}^{(k)} = \frac{h^{(k)}}{\sum_j h^{(j)}}$, ensuring that edge weights reflect the relative contribution of each signature to the observed glucose pattern for that patient-day segment.

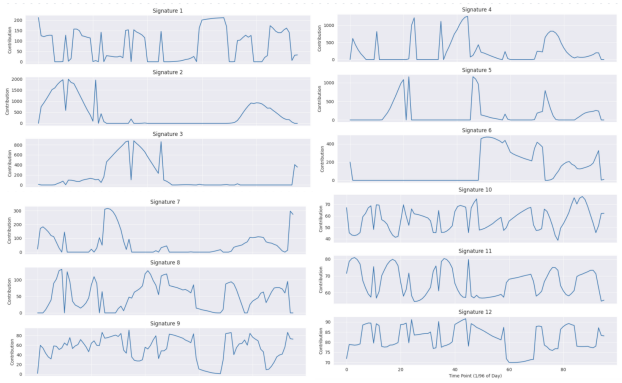

Figure 1. Non-negative Matrix Factorization (NMF) glucose signatures. Each curve represents one of the $K = 12$ basis patterns obtained from the factorization of one-day CGM segments. These signatures capture recurring intra-day glucose dynamics across the cohort and serve as feature vectors for signature nodes in the temporal heterogeneous graphs. The K was found by optimizing based on reconstruction error, obtaining a reconstruction error of 10%.

Sleep Embeddings. For each three-day window t, sleep features derive from multi-night physiological recordings available for a patient subset. Most participants possess high-resolution sleep data for only 2–3 nights during the entire observation period. For each recorded night τ, we compute an embedding $u_\tau \in \mathbb{R}^{d_s}$ using a pretrained sleep representation model capturing macro-architecture, stage composition, arousal patterns, and fragmentation metrics (Perslev et al., 2021; Thapa et al., 2024). For unrecorded nights, we impute embeddings by averaging all available nightly embeddings for that patient. When no sleep recordings exist, we substitute a learned "missing-sleep" token with accompanying binary mask features. All embeddings undergo standardization using training-set statistics exclusively. The resulting sleep embedding concatenates directly

with the health embedding z_t to form a combined patient node context vector.

Health Embeddings. To capture slowly-evolving patient health status at window t onset, we construct health embeddings $z_t \in \mathbb{R}^{d_h}$ through two pipelines. When sufficient free-text medical history exists, we encode all past medical notes (timestamped $\leq t$) using BioBERT, pooling contextualized token representations into a single vector (Lee et al., 2019). For insufficient textual data, we construct compact demographic vectors (BMI, sex, age) and project them into $\mathbb{R}^{d_h}$ via a learned linear transformation. We also tried replacing the LLM embeddings fully with those obtained from BMI, sex, and age, but the change in performance was minor.

The health embedding z_t concatenates with the corresponding sleep embedding to form: $z_t^{\text{ctx}} = [z_t \,\|\, u_t]$ where $\|$ denotes concatenation. This combined context vector captures both static/slowly-evolving health information and short-term sleep patterns, joining CGM-derived features, coverage metrics, calendrical encodings, and the latest precomputed patient embedding $\hat{e}_t$ in the final patient node representation.

Patient-Day Nodes and Temporal Continuity. Patient-day nodes p_t concatenate the latest available precomputed patient embedding $\hat{e}_t$ (selected from baseline/follow-up data with timestamp $\leq t$), sleep embedding u_t, health embedding z_t, CGM coverage scalar c_t, and calendrical encodings using sine/cosine transformations (Kazemi et al., 2019; Vaswani et al., 2017). Temporal edges $(p_t \to p_{t+1})$ capture within-patient dynamics. Once a diagnosis is made for a given condition, the patient is removed from subsequent graphs for the prediction task for that condition (only in the training set), following dynamic prediction principles that avoid post-outcome information (van Houwelingen & Putter, 2011). For inductive use of evolving node features over time, we follow the spirit of feature-driven inductive graph embedding (Hamilton et al., 2017).

Condition Nodes and Diagnosis Edges. We define fixed condition nodes $\{c^{(1)}, \ldots, c^{(M)}\}$. When patient diagnosis m occurs on day $t^\star$, we add event edge $(p_{t^\star} \to c^{(m)})$ with timestamp $t^\star$. No condition edges appear before onset. During validation and testing, these edges remain hidden from model inputs to prevent leakage while defining labels (Hu et al., 2020a; Kapoor & Narayanan, 2022).

Target Conditions. We modeled seven target diseases: *Essential hypertension*, *Other specified spine conditions (intervertebral disc displacement)*, *Osteoporosis*, *Diabetes mellitus (type unspecified)*, *Non-alcoholic fatty liver disease*, *Coronary atherosclerosis*, and *Malignant neoplasms of breast*. These conditions were selected for their cohort prevalence and clinical relevance for longitudinal risk prediction.

Disease	Events
Essential hypertension	229
Other spine conditions (disc displacement)	85
Osteoporosis	77
Diabetes mellitus, type unspecified	69
Non-alcoholic fatty liver disease	62
Coronary atherosclerosis	55
Malignant neoplasms of breast	35

Table 1. **Target conditions and cohort statistics.** Counts of diagnosis events used as condition nodes in the temporal heterogeneous graphs.

Data Splits and Graph Sparsity. We divided the longitudinal cohort into 70/15/15 train/validation/test splits, stratified by calendar time to prevent temporal leakage. Despite three-day aggregation, graphs remained extremely sparse: mean patient→signature edges per patient per graph were 0.0042 (training), 0.0042 (validation), and 0.0041 (test). Patient→condition edges averaged only 0.000087 per patient per graph in training. This sparsity, combined with the asymmetric treatment of diagnostic edges (present in training, hidden in evaluation), forced predictive models to infer diagnoses from indirect signals rather than direct edge presence, preserving evaluation integrity (Hu et al., 2020a; Kapoor & Narayanan, 2022) and reflecting the irregular and sparse nature of clinical data noted in previous work (Tipirneni et al., 2022).

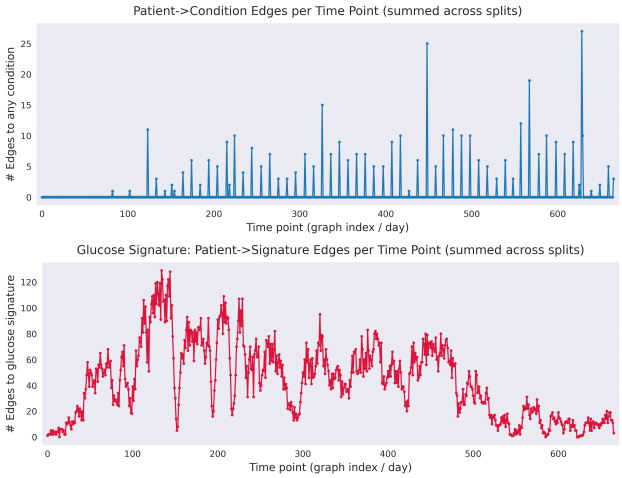

Figure 2. Density of edges over time in the temporal graphs. Up: distribution of patient→condition edges across time. Down: distribution of patient→glucose-signature edges across time. Both plots aggregate across the training, validation, and test splits.

2.2. Graph Neural Networks and Graph Attention for Temporal, Heterogeneous Patient Graphs

We operate on the three-day time-indexed heterogeneous graphs $\{G_t\}_{t=1}^{T}$ defined in the previous section, where $G_t = (\mathcal{V}_t, \mathcal{E}_t)$ contains patient nodes $p_t \in \mathcal{V}_t^{\text{pat}}$, glucose–signature nodes $s_t^{(k)} \in \mathcal{V}_t^{\text{sig}}$ (from NMF with $K = 12$ components), and condition nodes $c^{(m)} \in \mathcal{V}^{\text{cond}}$. Edges include (i) patient→signature links with weights given by normalized NMF activations $\tilde{h}_t^{(k)}$, and (ii) temporal patient→patient "follows" links $p_t \to p_{t+1}$. Diagnosis edges $p_t \to c^{(m)}$ are added only at onset times and are hidden from the model inputs during validation/testing to avoid leakage. Each node v carries an initial feature vector $\mathbf{h}_v^{(0)}$ comprising the components described earlier (e.g., CGM coverage, the combined health-sleep context z_t^{ctx}, calendrical encodings, and the latest available pre-window patient embedding $\hat{e}_t$).

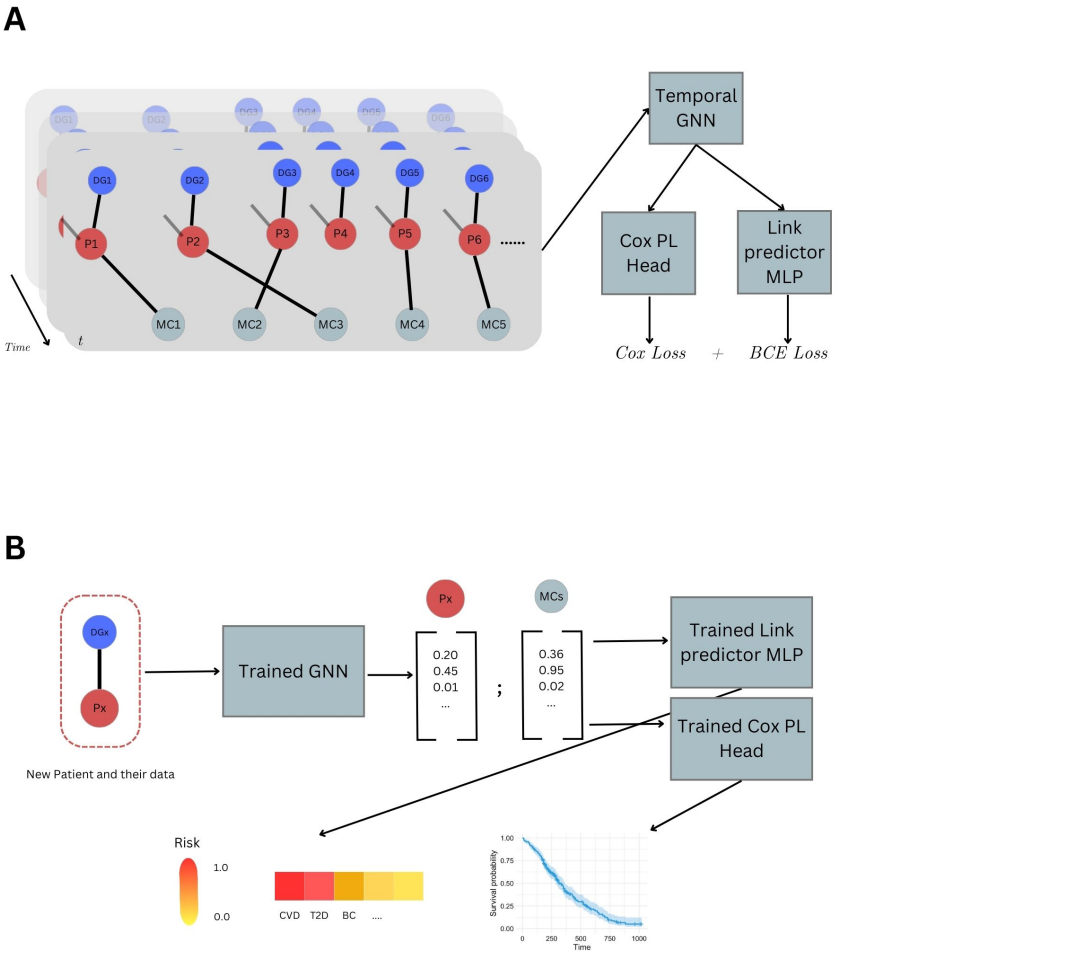

Figure 3. **Temporal GNN architecture and usage.** *(A) Modeling and training.* Three-day temporal graphs connect patient nodes (red), dynamically expressed data group nodes (blue, in this case, they are only glucose-signature nodes), and medical condition nodes (gray). Temporal progression is encoded by patient-follows-patient edges across windows; information also flows via shared signature nodes. A heterogeneous GAT encoder produces node embeddings that feed two predictive heads: a Cox partial likelihood survival head for modeling time-to-event and a link prediction MLP for near-term onset, optimized jointly with Cox loss and BCE loss. *(B) Inference.* For a new patient, the trained GNN encodes the patient–signature subgraph and outputs embeddings for patient and condition nodes. The link-prediction MLP yields short-horizon risk scores per condition, while the Cox PL head provides a survival function (or risk ranking) over time, enabling visualization as risk heatmaps and survival curves.

Message Passing on Typed Graphs. We adopt a message-passing neural network (MPNN) formalism (Gilmer et al., 2017) over a heterogeneous schema. Let $\mathcal{R}$ denote the set of edge types (relations). At layer ℓ, for an edge $(u \to v) \in$

$\mathcal{E}^r$ with type $r \in \mathcal{R}$ and (optional) edge attributes $\mathbf{e}_{u\to v}$ (e.g., $\tilde{h}_t^{(k)}$ on patient→signature links), we compute a typed message

$$\mathbf{m}_{u\to v}^{(\ell)} = \phi_{\mathrm{msg},r}^{(\ell)}\Big(W_{r,\mathrm{src}}^{(\ell)}\mathbf{h}_u^{(\ell)},\ W_{r,\mathrm{dst}}^{(\ell)}\mathbf{h}_v^{(\ell)},\ U_r^{(\ell)}\mathbf{e}_{u\to v}\Big) \tag{3}$$

where relation-specific projections follow the spirit of relational/heterogeneous GNNs (Schlichtkrull et al., 2018; Hu et al., 2020b). Node updates aggregate incoming messages with a permutation-invariant operator and update function:

$$\mathbf{h}_v^{(\ell+1)} = \phi_{\mathrm{upd}}^{(\ell)}\left(\mathbf{h}_v^{(\ell)},\ \bigoplus_{u\in\mathcal{N}(v)} \mathbf{m}_{u\to v}^{(\ell)}\right). \tag{4}$$

Attention-Weighted Message Passing (GAT). To allow the model to weight neighbors adaptively, we instantiate $\phi_{\mathrm{msg},r}$ with attention. For a neighbor $u \in \mathcal{N}(v)$ along relation r, we project node and edge features and compute a compatibility score with a learned vector $\mathbf{a}_r^{(\ell)}$:

$$e_{uv,r}^{(\ell)} = \mathrm{GeLU}\Big(\mathbf{a}_r^{(\ell)\top}\big[\tilde{\mathbf{h}}_{u,r}^{(\ell)} \,\|\, \tilde{\mathbf{h}}_{v,r}^{(\ell)} \,\|\, \tilde{\mathbf{e}}_{u\to v,r}^{(\ell)}\big]\Big) \tag{5}$$

$$\alpha_{uv,r}^{(\ell)} = \frac{\exp(e_{uv,r}^{(\ell)})}{\sum_{w\in\mathcal{N}_r(v)}\exp(e_{wv,r}^{(\ell)})}. \tag{6}$$

Attention coefficients $\alpha_{uv,r}^{(\ell)}$ modulate each neighbor–relation pair, yielding the attention-weighted aggregation

$$\mathbf{h}_v^{(\ell+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{u\in\mathcal{N}_r(v)} \alpha_{uv,r}^{(\ell)}\,\tilde{\mathbf{h}}_{u,r}^{(\ell)}\right), \tag{7}$$

with multi-head attention as in Vaswani et al. (2017), and heterogeneous/type-aware attention following Veličković et al. (2018); Hu et al. (2020b); Wang et al. (2019).

Prediction Heads and Loss Functions. On top of the learned node representations, we employ two types of supervision depending on the task: (i) *link prediction* between patient nodes and disease nodes, and (ii) *survival analysis* for time-to-event modeling. For link prediction we use a lightweight MLP edge predictor with a binary cross-entropy (BCE) loss. Given an edge score $\hat{y}_{uv} \in (0,1)$ for a patient–condition pair (u,v) and the ground-truth label $y_{uv} \in \{0,1\}$, the BCE loss is

$$\mathcal{L}_{\mathrm{BCE}} = -\Big(y_{uv}\log\hat{y}_{uv} + (1-y_{uv})\log(1-\hat{y}_{uv})\Big). \tag{8}$$

This encourages the model to assign high probability to observed (positive) edges and low probability to unobserved (negative) ones.

For survival modeling we adopt the partial likelihood of the Cox proportional hazards model (Cheng & Hu, 2025), which estimates relative risk while handling censored data. Each patient i with risk score r_i and event time T_i contributes

$$\mathcal{L}_{\mathrm{Cox}} = -\sum_{i\in\mathcal{E}}\Big(r_i - \log\sum_{j:T_j\geq T_i} e^{r_j}\Big), \tag{9}$$

where $\mathcal{E}$ denotes the set of uncensored events. This objective maximizes the likelihood that patients who experienced an event earlier receive higher predicted risk scores than those who remained event-free at that time.

In practice, we optimize a joint loss of the form

$$\mathcal{L} = \lambda_{\mathrm{BCE}}\,\mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{Cox}}\,\mathcal{L}_{\mathrm{Cox}}, \tag{10}$$

with task-specific weights λ balancing the contributions of binary link prediction and survival analysis. This design allows the model to capture both cross-sectional associations and longitudinal disease onset risk.

Evaluation Metrics. Given the extreme class imbalance in disease onset prediction (diagnosis events represent $\leq$ 0.01% of all patient-window pairs), we employ metrics specifically designed for imbalanced classification and survival analysis. For the link prediction task, we use the Precision-Recall Area Under Curve (PR-AUC) as the primary metric, which focuses on the model's ability to identify true positives among predicted positives, which is critical when positive cases are rare. We report the Receiver Operating Characteristic Area Under Curve (ROC-AUC) as a secondary metric, though it can be misleadingly optimistic in highly imbalanced settings due to the large number of true negatives. For survival analysis, we employ Harrell's concordance index (C-index), which measures the proportion of comparable patient pairs where the model correctly orders their risk scores relative to their observed event times. The C-index is particularly appropriate for censored data and provides an interpretable measure of discriminative ability, where 0.5 indicates random performance and 1.0 indicates perfect ranking. These metrics collectively assess both the model's ability to identify imminent disease onset (PR-AUC, ROC-AUC) and its capacity to rank patients by long-term risk (C-index).

Temporal Connectivity and Depth. Temporality is encoded with a minimal yet effective design: for each patient we add a single follows edge $p_t \to p_{t+1}$ linking successive three-day windows. Longer-range temporal dependencies are captured via multi-hop propagation over these edges. We stack three GAT layers, which offers two desirable propagation radii simultaneously: (i) within a window, patients influence each other indirectly via their shared glucose–signature nodes (patient→signature←patient), and (ii) across windows, information can travel forward up to three successive

windows. Because each window spans three days, this depth lets the representation at $t+1$ incorporate signals up to roughly nine days prior without introducing long-range temporal edges that often exacerbate over-smoothing and optimization instability (Oono & Suzuki, 2020; Li et al., 2018). Temporal position and calendar signals included in $\mathbf{h}_v^{(0)}$ further help the model to disambiguate recency and seasonality (Vaswani et al., 2017).

Why Attention Suits This Setting. Our graphs are heterogeneous and multi-scale: patient nodes carry slow-varying clinical context and sleep embeddings, signature nodes capture structured CGM patterns over the current window, and temporal edges express short-range dynamics. The informativeness of neighbors is highly context dependent. Attention allows the model to emphasize strongly expressed glucose signatures for a specific patient–window, downweight signatures when CGM coverage is low, and adapt the reliance on temporal edges when recent measurements are volatile or stable. The locally normalized coefficients are robust to degree variation across patients, and their values offer a readily interpretable attribution of which neighbors and relations dominated each update, useful when auditing physiological drivers of the model's predictions (Veličković et al., 2018; Wang et al., 2019; Choi et al., 2017).

Computational Considerations. For hidden width d and H heads, one GAT layer scales as $O(H\,|\mathcal{E}|\,d)$ flops per forward pass with memory proportional to the edges in the minibatch (Veličković et al., 2018). We employ typed sparse adjacency, neighbor sampling on high-degree relations, mixed-precision training with loss scaling, residual connections, and feature-wise LayerNorm to improve stability (Hamilton et al., 2017; Zeng et al., 2020; Ying et al., 2018; Micikevicius et al., 2018; Ba et al., 2016). To avoid over-parameterization on relations with few edges, we share projections across low-support types and use compact edge encoders (e.g., one- or two-layer MLPs) for $\mathbf{e}_{u\to v}$ (Schlichtkrull et al., 2018). In practice this yields stable training and inference times while preserving the capacity to model relation- and time-dependent effects.

Summary of the Instantiated Encoder. Concretely, we use a three-layer heterogeneous GAT with multihead attention, relation-specific projections on patient↔signature and temporal relations, and edge-aware attention on patient→ signature links via normalized NMF activations. The encoder produces patient-node embeddings for each window; these feed the downstream predictive heads described in the next subsection (a short-horizon link-prediction MLP targeting condition onsets and a Cox partial-likelihood head for time-to-event modeling), ensuring a single shared representation supports both near-term and survival objectives (Veličković et al., 2018; Hu et al., 2020b; Kipf & Welling, 2016; Cox, 1975; Katzman et al., 2018).

Model Development Attempts and Negative Findings. Given the extreme sparsity of the temporal graphs and the rarity of onsets, we undertook a series of interventions aimed at improving the short-horizon link-prediction head and the survival head. We summarize the most important attempts here and explain why, despite careful controls, none produced a reliable performance gain. Throughout, we evaluated with patient-level splits, early stopping on validation *PR-AUC* for the link task (preferred under heavy imbalance) and *ROC-AUC* as a secondary metric; the survival head was monitored with concordance index and time-dependent AUC. All decisions (hyperparameters, checkpoints) were fixed on validation and reported on the held-out test set (Saito & Rehmsmeier, 2015; Heagerty et al., 2000; Harrell, 2015; Hu et al., 2020a).

Synthetic Minority Over-sampling (SMOTE). We experimented with embedding-space SMOTE applied to minority positive examples within each minibatch. The synthetic samples were created by linear interpolation in the patient-node feature space and attached to the same local topology as their source nodes. Although this increased the apparent positive rate during training, it did not change the topological signal: synthetic nodes either duplicated neighborhoods (risk of overfitting) or, if detached, broke the correspondence between features and edges. In practice, validation PR-AUC and ROC-AUC remained at the level of the unbalanced baseline, suggesting that class imbalance was not the sole limiter; edge sparsity and weak relational signal dominated (Chawla et al., 2002).

Pseudo-labeling of Uncertain Positives. Motivated by recent work on pseudo-labeling for graphs, we iteratively identified high-confidence predictions in the training split and added them as provisional positives with temperature-scaled soft targets. We guarded against confirmation bias with conservative thresholds, temporal holdout of the target window, and consistency checks across checkpoints. Even so, the noise introduced by incorrect pseudo-labels was not offset by the additional supervision: the validation PR-AUC oscillated around the baseline and sometimes degraded, indicating that label noise in this regime was more harmful than the extra mass of positives was helpful (Alchihabi et al., 2024).

Risk-edge Augmentation. To inject clinically plausible priors without leaking labels, we added, for each patient node, a typed edge carrying a scalar *risk attribute* derived from a parametric function of BMI, age, and sex. For patients who never developed the condition in training, the risk rose smoothly up to a ceiling of 0.9; for those who did, the risk started from the same baseline and then increased

step-wise each window to approach 1.0 by onset. This created a dense auxiliary signal available to attention. Two failure modes emerged: either the network learned to shortcut via the risk attribute (inflating validation but collapsing on test once diagnosis edges were removed), or, when regularized to prevent shortcutting (dropout on edge attributes, relation-specific weight decay), the added signal acted as noise because it was only weakly aligned with the sparse onset edges. Net effect: no robust improvement over chance (Geirhos et al., 2020).

Reducing the Negative Pool. We tried training on subsets that down-weighted easy negatives by restricting to patients and windows with higher CGM coverage or with more active signature edges. While this raised the average degree and simplified the classification problem seen during training, it also narrowed the coverage of the operating regime. Generalization to the full validation/test graphs dropped slightly and PR-AUC did not improve, indicating that the model benefited from seeing the true negative distribution, even if heavily skewed.

Increasing Per-graph Temporal Span. To combat sparsity we "crammed" more days per snapshot (expanding the window beyond three days). This densified patient→signature connections and created more temporal edges, but at the cost of blurring event timing and increasing heterophily across amalgamated windows. Deeper GAT stacks partially compensated, yet over-smoothing and training instability appeared earlier, and the net effect on PR-AUC/ROC-AUC was neutral to negative (Oono & Suzuki, 2020; Li et al., 2018).

Outcome and Rationale. Across all interventions, validation PR-AUC remained close to the prevalence baseline and ROC-AUC hovered near 0.5; the held-out test mirrored validation. The common thread is that signal-to-noise in the available relations was insufficient: diagnosis edges were removed from validation/test to prevent leakage (by design), patient→signature edges were extremely sparse (mean ≈ 0.004 per patient per graph across splits), and the temporal chain alone did not carry enough discriminative information. Augmentations that add feature-level mass without strengthening topology either encouraged shortcuts or injected noise (Tipirneni et al., 2022).

Reporting Choice. Because none of these lines of work yielded conclusive gains, and several risked selection bias if reported in isolation, we do not include the GNN predictive metrics in the *Results* section. The main narrative in the *Results* section focuses on the I-JEPA imaging pathway, which produced stable and interpretable results.

2.3. Self-Supervised Predictive Learning with JEPA and I-JEPA

The Joint-Embedding Predictive Architecture (JEPA) is a self-supervised framework that learns semantic representations by predicting latent targets of masked regions from visible context (LeCun, 2022). Given an image x, JEPA partitions it into a context subset $\mathcal{C}$ and multiple disjoint target subsets $\{\mathcal{T}_i\}_{i=1}^{M}$. The architecture employs three components: a context encoder f_θ processing visible regions $x_\mathcal{C}$, a target encoder f_ξ processing masked regions $x_{\mathcal{T}_i}$, and a predictor g_ϕ that maps context representations to target feature space (Assran et al., 2023).

The training objective minimizes prediction error in latent space:

$$\begin{aligned} \mathbf{z}_\mathcal{C} &= f_\theta(x_\mathcal{C}), \qquad \mathbf{y}_i \;=\; \mathrm{sg}\big(f_\xi(x_{\mathcal{T}_i})\big), \\ \widehat{\mathbf{y}}_i &= g_\phi\big(\mathbf{z}_\mathcal{C},\, \mathbf{m}_i\big), \\ \mathcal{L}_{\mathrm{JEPA}}(\theta,\phi;\xi) &= \frac{1}{M}\sum_{i=1}^{M} \|\widehat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2 , \end{aligned} \tag{11}$$

where $\mathbf{m}_i$ encodes target geometry metadata and the target encoder is updated via exponential moving average: $\xi \leftarrow \tau\,\xi + (1-\tau)\,\theta$ with decay $\tau \in (0,1)$ (Assran et al., 2023). This design stabilizes training while encouraging semantically meaningful representations without pixel-level reconstruction.

I-JEPA Architecture and Medical Imaging Advantages I-JEPA implements JEPA principles using vision transformers operating on patch tokens (Assran et al., 2023). Input images are partitioned into non-overlapping patches grouped into rectangular blocks, with random subsets designated as context (visible) or targets (masked). The context encoder processes only visible blocks, while the predictor generates target predictions conditioned on pooled context representations and geometric metadata (Assran et al., 2023).

This design offers three key advantages for medical imaging: (1) **Semantic focus**: latent prediction emphasizes spatial relations and structural patterns crucial for clinical interpretation of vessels, lesions, and anatomical boundaries; (2) **Memory efficiency**: masking large blocks enables training at native high resolutions without the computational burden of pixel-level reconstruction; (3) **Preservation of clinical signals**: the absence of aggressive augmentations prevents distortion of subtle pathological features that may be critical for diagnosis (Huang et al., 2023).

Retinal Fundus Adaptation with LoRA For retinal fundus analysis, we adapt I-JEPA through Low-Rank Adaptation (LoRA), which introduces trainable low-rank matrices into transformer attention projections (Hu et al., 2021).

Given a base weight $W \in \mathbb{R}^{d\times d}$, LoRA computes the adapted weight as $W' = W + BA$ with $A \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{d\times r}$, and rank $r \ll d$. This parameter-efficient approach enables fine-tuning at high resolution while preserving pretrained knowledge.

Training Protocol and Objectives During adaptation, we retain the I-JEPA self-supervised objective while introducing LoRA parameters. The context encoder and predictor are updated jointly with LoRA weights, while the target encoder follows exponential moving average updates. This ensures predictions remain anchored to consistent latent targets while allowing efficient domain adaptation (Assran et al., 2023; Hu et al., 2021).

For supervised fine-tuning on external datasets, we attach lightweight classification heads to the representations $\hat{\mathbf{y}}$. The combined training objective balances self-supervised prediction with task-specific supervision, enabling the model to leverage both pretrained semantic knowledge and domain-specific signals (Huang et al., 2023).

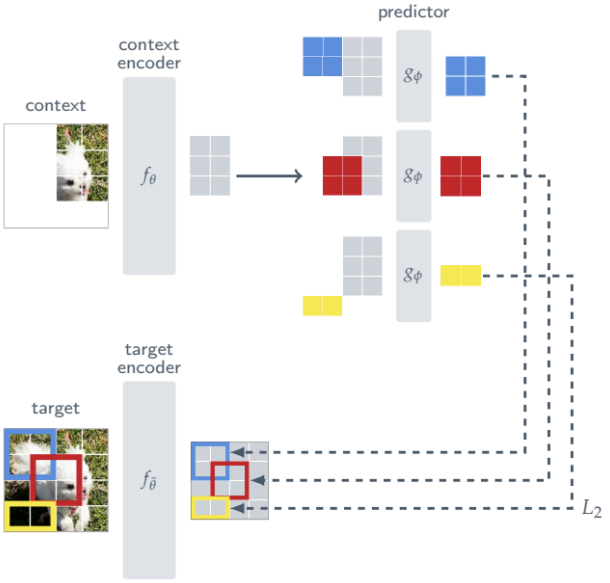

Figure 4. Schematic overview of the Image-JEPA (I-JEPA) framework. A context encoder generates latent embeddings from visible image patches. A predictor network maps these embeddings to predict the target representations of masked regions, which are computed by a frozen target encoder. This training objective drives the model to capture high-level semantic relationships, rather than reconstructing pixel-level details (Assran et al., 2023).

Preprocessing and Implementation Retinal images undergo standardized preprocessing, including center-cropping to isolate the retinal field of view, resizing to match the backbone's input resolution, and per-image normalization (Zhou et al., 2022). The masking of the large target region forces the model to integrate information across critical retinal structures, such as vessels, the optic disc, and the macula, anatomical characteristics central to ophthalmological evaluation (Poplin et al., 2018).

The combination of JEPA's representation learning strengths and LoRA's parameter efficiency creates a framework well-suited for medical imaging: latent space prediction emphasizes semantically meaningful features while avoiding augmentation-induced artifacts, EMA stabilization prevents representational collapse, and efficient adaptation enables deployment under realistic computational constraints (Assran et al., 2023; Hu et al., 2021; Grill et al., 2020).

External Validation Protocol To assess cross-dataset generalization, we evaluate the adapted model on three public retinal datasets: PAPILA (glaucoma assessment), IDRiD (diabetic retinopathy and macular edema), and Messidor (diabetic retinopathy screening) (Kovalyk et al., 2022; Porwal et al., 2018; Decencière et al., 2014). All datasets undergo identical and minimal preprocessing to ensure consistency across patient populations and imaging devices.

Two evaluation protocols test different aspects of transferability: (1) *linear probing* uses frozen backbone features with a linear classifier, assessing representation quality; (2) *lightweight fine-tuning* optimizes small MLP heads with optional LoRA adapter unfreezing, evaluating adaptation capacity. Subject-level data splits prevent leakage, and standard metrics (accuracy, F1, AUC) quantify performance across clinically diverse endpoints (Kapoor & Narayanan, 2022).

Figure 5. **Example retina fundus images.** Right eye (OD, *oculus dexter*, left) and left eye (OS, *oculus sinister*, right). These are very high-resolution images, and of much better quality than fundus data sets like the one offered by the UK Biobank, making them suitable for pretraining and fine-tuning. Because patient diagnoses in the cohort are self-reported and not particularly prevalent, we chose not to use them as predictive labels.

Dataset	Resolution (px)	Total images	Good (n, %)	Bad (n, %)
UKBB	1300×1300	133,791	68,574 (51.27%)	65,217 (48.73%)
10K (HPP)	3000×3000	22,737	21,134 (92.96%)	1,603 (7.04%)

Table 2. **Retina image datasets: resolution and quality split.** Quality classification for both UKBB and 10K (HPP) follows the *Automorph* pipeline. **Good images** are clear, centered, well-exposed, and diagnostically usable (sharp vessels, visible optic disc and macula). **Bad images** are blurred, artifact-heavy, off-centered, or insufficient for diagnostic use. Good images allow vessel extraction and structural analysis; bad images are excluded from predictive modeling. Resolutions listed are the native acquisition sizes used in preprocessing.

We systematically compare four adaptation strategies to disentangle the contributions of different training phases: **(1) HPP pretraining**: I-JEPA Imagenet backbone furhter pretrained on the HPP retina dataset; **(2) HPP fine-tuning**: Imagenet backbone fine-tuned (LoRA) to predict retinal features (vessel density, optic disc morphology); **(3) ImageNet baseline**; **(4) k-NN baseline**: non-parametric classification using frozen pretrained features from the Imagenet baseline (Assran et al., 2023; Hu et al., 2021). This framework enables a comprehensive assessment of in-domain pretraining benefits, auxiliary task contributions, and representation transferability. The external datasets span distinct clinical contexts, providing diverse evaluation scenarios.

3. Results and Discussion

3.1. I -JEPA pretraining on HPP retinal images

We initialized our experiments from the publicly released I -JEPA weights trained on ImageNet-22K (Assran et al., 2023; Deng et al., 2009). Specifically, we used the **ViT-H/14 backbone** (patch size 14×14, input resolution 224×224, 66 transformer layers) pretrained on ImageNet-22K (Dosovitskiy et al., 2020) obtained from the offcial GitHub repository of IJEPA. This large-scale initialization provided a strong general-purpose representation prior, which we then adapted to high-resolution retinal fundus photographs from the HPP.

We first examine how pretraining reshapes the geometry of the learned representation space (Huang et al., 2023). Figure 6 visualizes the first two principal components of image embeddings *before* and *after* pretraining, colored (or labeled) by age, BMI, and sex. Prior to pretraining, the embedding cloud traces a broad, curved manifold: PC1 explains roughly half of the variance and PC2 a further $\sim$20%. A pronounced color gradient appears along age and a weaker one along BMI, while the sex classes largely overlap (Poplin et al., 2018). After pretraining, variance concentrates strongly on PC1 ($\sim$80–90%), the cloud narrows into a band, and both age and BMI color gradients become visibly weaker; sex remains overlapping. A few outliers in the post-pretraining space stretch the axes, but they do not alter the overall pattern.

To quantify whether demographic attributes remain linearly decodable from the embeddings, we track probe AUC across corrected pretraining epochs (Figure 7). The trends mirror

the PCA: age AUC decreases slightly from an initial ≈ 0.87 to ≈ 0.82–0.85; sex falls from ≈ 0.64 to ≈ 0.58–0.60; BMI hovers near chance at ≈ 0.50–0.55 with large uncertainty. In short, this pre-training does not increase demographic decodability; if anything, it attenuates it, consistent with a representation that emphasizes the retinal structure useful for context prediction while becoming less entangled with subject demographics (Assran et al., 2023; LeCun, 2022; Grill et al., 2020). Whether that attenuation is desirable depends on the downstream goal: for unbiased risk modeling it can be beneficial; if demographic awareness is required, one could introduce supervised auxiliaries, adjust masking geometry, or tune EMA dynamics (Huang et al., 2023; LeCun, 2022). A key limitation that most probably prevented the success of this experiment was the low resolution of the images we were forced to use due to GPU constraints (616×616). Other solutions to this problem, like gradient accumulation (Zhang et al., 2023), were not effective.

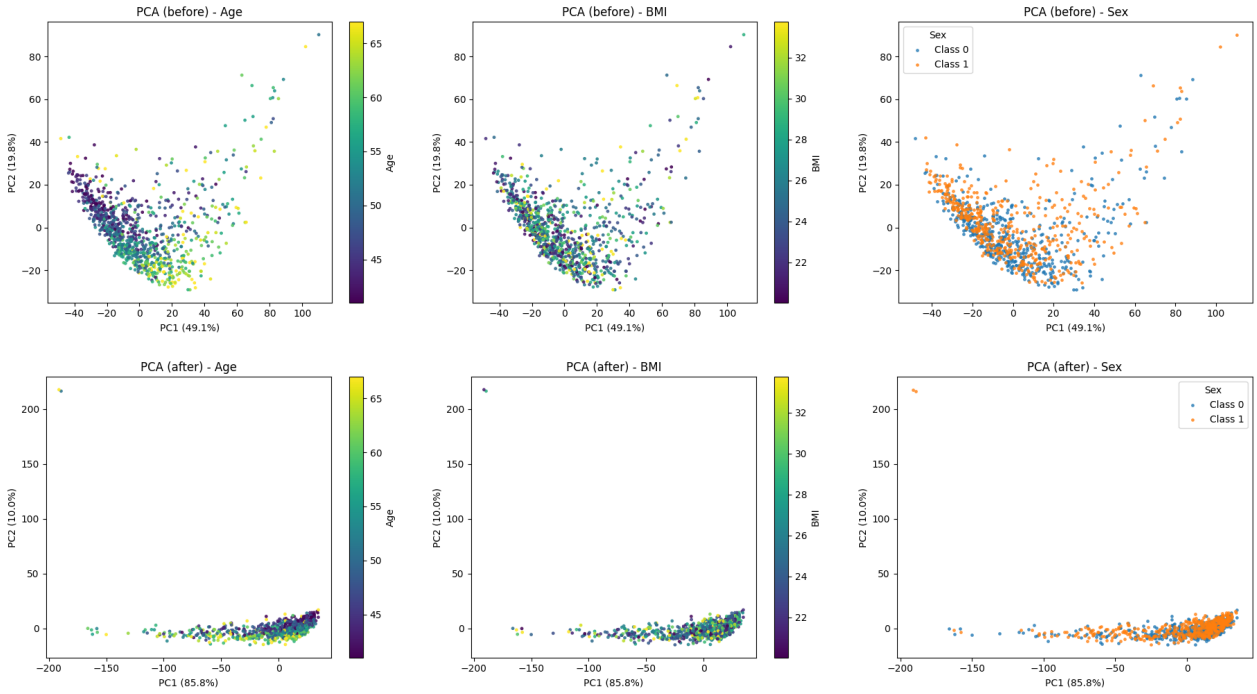

Figure 6. **I-JEPA pretraining reshapes representation geometry.** Top row: PCA before pretraining (colored by age/BMI; sex classes overlaid). Bottom row: PCA after pretraining. Before, embeddings span a broad, curved manifold with clear age and weak BMI gradients; sex largely overlaps. After pretraining, variance concentrates onto PC1 (band-like cloud), age/BMI gradients weaken, and sex remains overlapping. Isolated outliers stretch the axes post-pretraining but do not affect the overall pattern.

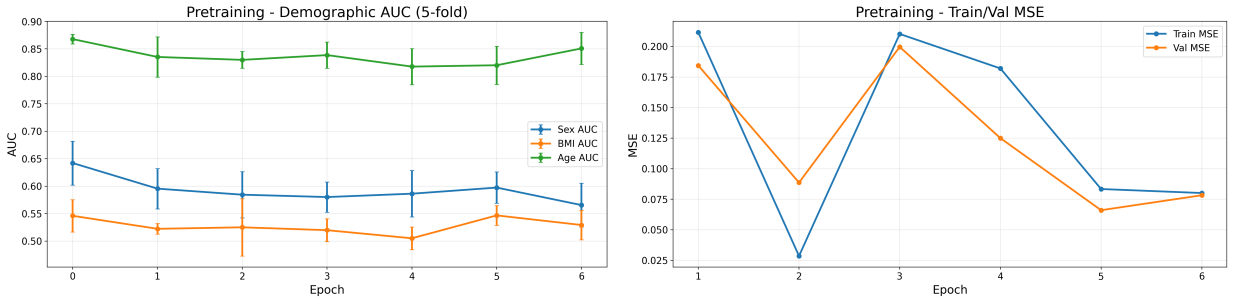

Figure 7. **Pretraining dynamics.** *Left:* Five-fold probe AUC trajectories for age, sex, and BMI across pretraining epochs. Age starts high ($\sim$0.87) and declines slightly to $\sim$0.82–0.85; sex drops from $\sim$0.64 to $\sim$0.58–0.60; BMI remains near chance at $\sim$0.50–0.55 with wide fold-to-fold uncertainty. *Right:* Training and validation mean squared error (MSE) across epochs. Despite fluctuations in early epochs, both curves steadily decrease and converge, confirming that the I-JEPA objective is optimized effectively. Together, these results show that demographic decodability declines while the pretext loss improves, producing embeddings less entangled with age, sex, or BMI.

Fine-tuning on Automorph retinal features. Building on the pretrained backbone, we fine-tuned I-JEPA with LoRA adapters ($r = 16$, $\alpha = 16$, dropout $= 0.2$) on the Human Phenotype Project retinal images, using Automorph-derived morphometric features as regression targets. This task probes whether the learned representations could be adapted to capture clinically relevant vascular structure, beyond demographic covariates.

The demographic probe results (Fig. 8, top left) indicate that, unlike pretraining, fine-tuning partially *increases* demographic separability. Age AUC remains high ($\sim$0.86–0.87), essentially unchanged from the starting point, while sex improves steadily from $\sim$0.70 at epoch 0 to $\sim$0.73 by epoch 6, and BMI rises modestly from chance levels to $\sim$0.55. This suggests that the regression task encourages the backbone to re-expose demographic cues correlated with retinal vessel morphology. PCA projections after fine-tuning (Fig. 8, top right) corroborate this shift: embeddings show visible gradients for age and BMI, while sex classes, though overlapping, are more distinguishable compared to pretraining.

Training dynamics further support this interpretation. Mean squared error on both training and validation sets decreases smoothly across epochs (Fig. 8, bottom left), indicating stable optimization without overfitting. Analysis of individual Automorph features (Fig. 8, bottom right) reveals that vessel density and fractal dimension dominate the top-5 most predictable features (Zhou et al., 2022). Validation R^2 for artery vessel density, for example, rises from $\sim$0.15 to > 0.35 across epochs, while vein density and global vessel density follow similar trends. These features are among the most biologically interpretable and clinically relevant descriptors of vascular architecture, highlighting that fine-tuning drives the model to encode structural information aligned with expert-derived biomarkers (Zhou et al., 2022).

Overall, fine-tuning on Automorph targets succeeds in aligning the pretrained backbone with vessel-level structural features. Although this improves demographic decodability somewhat, it simultaneously increases the ability of the model to capture clinically relevant morphometrics, suggesting that fine-tuning acts as a domain adaptation step that reinforces the vascular signal at the cost of slightly higher demographic entanglement.

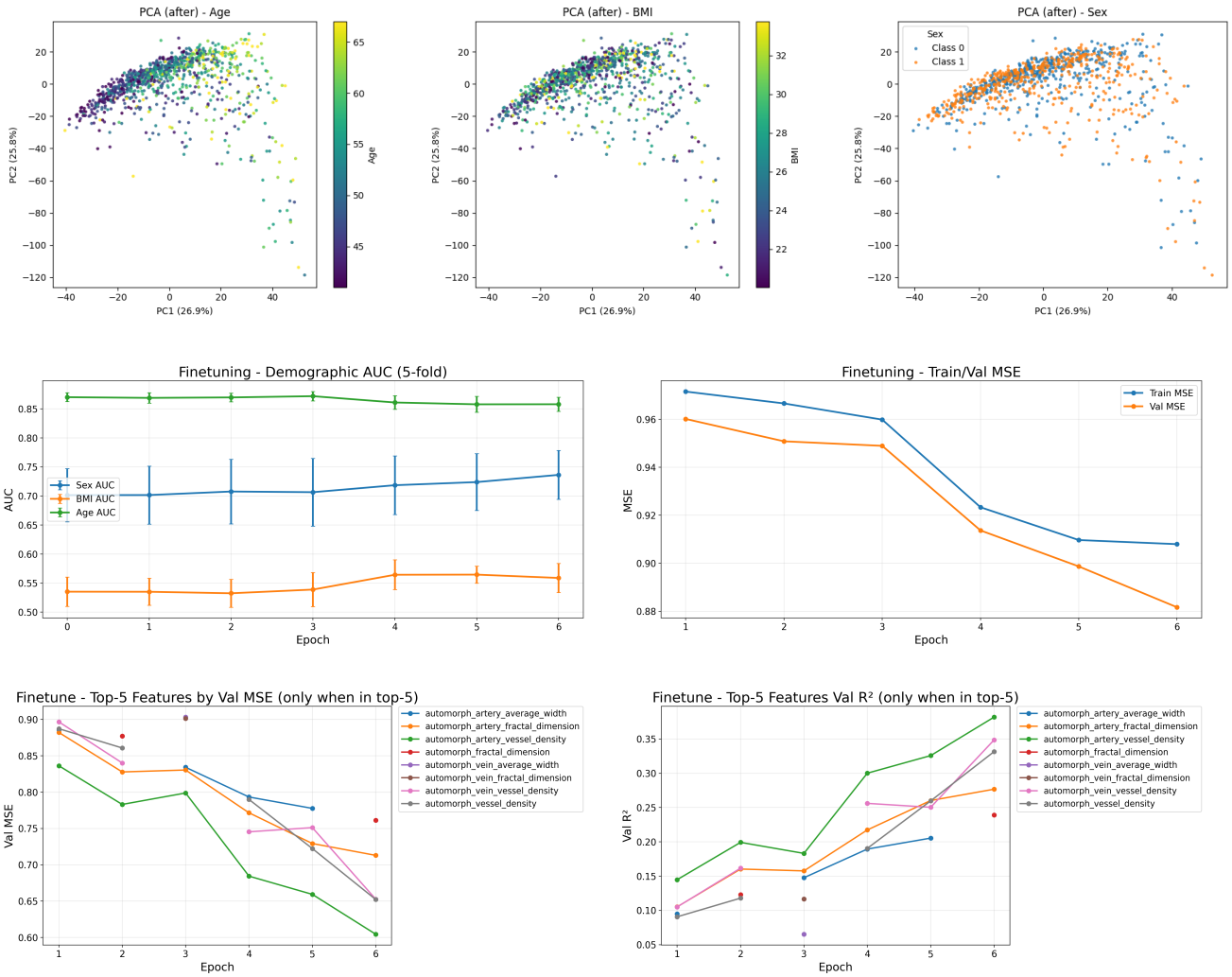

Figure 8. **Fine-tuning results on Automorph features.** Top row: PCA plots of embeddings after fine-tuning, colored by age, BMI, and sex. Compared to pretraining, demographic gradients are more visible and sex separability slightly improves. Middle row: probe AUC trajectories (left) and training/validation reconstruction loss (right) across epochs. Bottom row: top-5 Automorph features ranked by validation MSE (left) and R^2 (right), showing which retinal features are most reliably predicted during fine-tuning.

3.2. Cross-Dataset Evaluation

We evaluated the transfer performance of different adaptation strategies across external datasets. While PAPILA was initially considered, we excluded it due to its substantially lower image quality and inconsistent optic nerve visibility compared to the other datasets (Kovalyk et al., 2022). We therefore focus on two benchmarks: **IDRiD**, focusing on diabetic retinopathy, and **Messidor** (Porwal et al., 2018; Decencière et al., 2014). See Table 3 for numerical results.

Our fine-tuned I-JEPA with LoRA outperformed the ImageNet-only baseline on Messidor, achieving an AUC of ∼0.76 vs. ∼0.73. On IDRiD, however, all models, including the fine-tuned and ImageNet-only variants, overfit rapidly, with AUCs hovering near chance and failing to improve meaningfully even with hyperparameter sweeps. This was particularly evident from early stopping curves, where validation loss deteriorated almost immediately after a few epochs. Despite attempts at regularization and dropout tuning, IDRiD remained a challenging benchmark for our transfer pipeline.

Interestingly, the non-parametric **kNN baseline** on ImageNet features produced a relatively strong performance on IDRiD (∼0.78 AUC), suggesting that simple similarity-based classifiers can sometimes exploit dataset-specific biases more effectively than fine-tuned transformers. However, the same approach underperformed on Messidor, where LoRA fine-tuning of I-JEPA embeddings retained a clear

advantage.

When compared to recent state-of-the-art results, such as OphtAI's 0.989 AUC on Messidor and BhAFPN's 0.9901 AUC on IDRiD, our adapted I-JEPA models fall far short in absolute terms (Quellec et al., 2019; Mukherjee et al., 2025). Nonetheless, the relative improvements of LoRA fine-tuning over vanilla ImageNet initialization on Messidor, together with the promising IDRiD kNN baseline, highlight the flexibility of representation-based pipelines. These experiments suggest that I-JEPA embeddings retain latent capacity for transfer to ophthalmic tasks, even if our current finetuning setup does not yet unlock their full potential.

Parameter sweeps across learning rates, dropout values, and masking strategies yielded limited performance improvements. IDRiD particularly displayed rapid overfitting regardless of configuration, with validation AUCs clustering near chance ($\sim$0.5). This highlights the need for more sophisticated domain adaptation, augmentation strategies, or semi-supervised approaches to fully exploit I-JEPA's potential in small clinical datasets. Conversely, the Messidor results demonstrate that lightweight LoRA adaptation can achieve meaningful improvements over standard transfer learning.

These findings reveal both the promise and limitations of adapting I-JEPA for medical imaging. While the model shows encouraging performance on large, high-quality datasets like Messidor, it struggles with smaller, noisier cohorts like IDRiD. This underscores I-JEPA's potential as a medical imaging foundation model, contingent on developing stronger domain adaptation strategies and access to larger, clinically curated benchmarks.

Strategy	**IDRiD**	**Messidor**
HPP pretrain	0.503 ± 0.004	0.681 ± 0.002
HPP finetune	0.531 ± 0.012	$\mathbf{0.760 \pm 0.001}$
ImageNet	0.542 ± 0.003	0.731 ± 0.007
kNN ImageNet	$\mathbf{0.779 \pm 0.041}$	0.642 ± 0.019

Table 3. **Evaluation of transfer strategies across retinal benchmarks.** ROC AUC scores ($\pm$ standard error) for IDRiD and Messidor. PAPILA was excluded due to poor image quality. LoRA fine-tuning on HPP outperforms ImageNet-only initialization on Messidor, while IDRiD performance remains near chance, with the exception of a kNN baseline that achieved moderately higher scores. State-of-the-art DR-specific models report AUCs above 0.99, underscoring the gap to task-optimized methods.

To further probe how different pretraining and fine-tuning strategies influence model behavior, we visualized attention maps on representative Messidor images (Fig. 9). These qualitative comparisons highlight notable differences: the ImageNet baseline tends to produce diffuse or misplaced attention, often outside clinically informative regions. In contrast, HPP-pretrained models shift attention toward vascular structures and the optic disc, while the two-stage finetuning pipeline further consolidates focus around both the optic disc and macula. Such patterns suggest that I -JEPA pretraining not only improves quantitative performance on Messidor but also aligns model attention more closely with clinically meaningful retinal features.

Figure 9. **Attention visualization across fine-tuning strategies on Messidor.** Example fundus images with overlaid attention maps from three model variants: **(A)** ImageNet-pretrained baseline, **(B)** HPP pretrain strategy, **(C)** HPP finetune strategy Green highlights indicate regions that the model attends to when making predictions. The ImageNet baseline (**A**) shows diffuse and often misplaced attention outside clinically relevant areas. In contrast, pretraining on HPP (**B**) sharpens attention toward vascular and optic disc regions, while the two-stage fine-tuning strategy (**C**) further consolidates focus around the optic disc and macula. These qualitative differences suggest that I -JEPA pretraining yields representations more aligned with clinically meaningful retinal structures.

4. Conclusion

This thesis explored the fundamental question of whether we can unify diverse healthcare data modalities to predict chronic disease onset. The graph-based experiments, while not yielding strong predictive results despite extensive optimization efforts, provided crucial insights into the challenges of unifying sparse, heterogeneous clinical data within temporal frameworks. These findings illuminate critical obstacles that must be addressed before graph-based approaches can effectively integrate diverse healthcare modalities for early disease prediction.

The I-JEPA adaptation proved more promising as a foundation for unified healthcare modeling. While this proof-of-concept focused on retinal imaging, I-JEPA's inherent ability to learn joint representations across data modalities positions it as a compelling architecture for the broader goal of healthcare data unification. The successful capture of clinically relevant retinal features, despite not surpassing specialized models, demonstrates how self-supervised architectures can serve as building blocks for truly multi-modal systems that could eventually integrate imaging, physiological monitoring, and structured clinical data within a single predictive framework. The contrasting outcomes highlight a key insight: successful healthcare data unification may require architectures specifically designed for multi-modal learning rather than approaches that attempt to force diverse data types into uniform representations.

This work contributes to the broader challenge of developing clinically deployable models that can harness the full spectrum of healthcare data. The lessons learned here, both the limitations of graph-based integration and the promise of self-supervised multi-modal architectures, provide a roadmap for future research toward truly unified healthcare prediction systems capable of early disease detection across multiple data modalities.

The next steps for the JEPA project will be to start integrating new data into this task, for example, medical information, basic BMI, sex, and age, and other features of building a richer and more meaningful latent space.

Recent developments in unified foundation models, such as MedGemma's multimodal approach to medical text and imaging (Sellergren et al., 2025), demonstrate the growing momentum toward the vision explored in this thesis. As these foundation models mature, the architectural insights gained from I-JEPA's joint embedding approach may prove instrumental in developing the next generation of unified healthcare AI systems.

5. Code Availability

Due to data privacy regulations and ethical considerations, the Human Phenotype Project dataset used in this thesis cannot be shared publicly. However, most of the code developed during this research is available in the following repository: `https://github.com/gavrielhan/MSc_Thesis`. This includes implementations of the temporal graph neural network architectures, I-JEPA adaptation with LoRA, preprocessing pipelines, and evaluation scripts. The code is provided to facilitate reproducibility and to enable researchers to adapt these methods to their own datasets and clinical applications.

References

Alchihabi, A., Yan, H., and Guo, Y. Overcoming class imbalance: Unified gnn learning with structural and semantic connectivity representations. *arXiv*, 2024. URL https://arxiv.org/abs/2412.20656.

American Diabetes Association Professional Practice Committee. 2. diagnosis and classification of diabetes: Standards of care in diabetes—2025. *American Diabetes Association*, 2025. doi: 10.2337/dc25-S002. URL https://diabetesjournals.org/care/article/48/Supplement_1/S27/157566.

Armoundas, A. A. et al. Use of artificial intelligence in improving outcomes in heart failure. *American Heart Association*, 2024. doi: 10.1161/CIR.0000000000001201. URL https://www.ahajournals.org/doi/10.1161/CIR.0000000000001201.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv*, 2023. doi: 10.48550/ARXIV.2301.08243. URL https://arxiv.org/abs/2301.08243.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv*, 2016. doi: 10.48550/ARXIV.1607.06450. URL https://arxiv.org/abs/1607.06450.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv*, 2021. doi: 10.48550/ARXIV.2105.04906. URL https://arxiv.org/abs/2105.04906.

Bardes, A., Ponce, J., and LeCun, Y. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv*, 2023. doi: 10.48550/ARXIV.2307.12698. URL https://arxiv.org/abs/2307.12698.

Boll, H. O., Amirahmadi, A., Athiveeraramachandran, D., Chehri, A., and Sik-Lanyi, C. Graph neural networks for clinical risk prediction based on electronic health records: A survey. *Elsevier*, 2024a. doi: 10.1016/j.jbi.2024.104616. URL https://www.sciencedirect.com/science/article/pii/S1532046424000340.

Boll, H. O. et al. Graph neural networks for heart failure prediction on an ehr-based patient similarity graph. *arXiv*, 2024b. doi: 10.48550/arXiv.2411.19742. URL https://arxiv.org/abs/2411.19742.

Cao, D., Wang, Y., Duan, J., Zhang, C., Zhu, X., Huang, C., Tong, Y., Xu, B., Bai, J., Tong, J., and Zhang, Q. Spectral temporal graph neural network for multivariate time-series forecasting. *arXiv*, 2020. doi: 10.48550/arXiv.2005.10296. URL https://proceedings.neurips.cc/paper/2020/hash/cdf6581cb7aca4b7e19ef136c6e601a5-Abstract.html.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *AI Access Foundation*, 2002. doi: 10.1613/jair.953. URL https://www.jair.org/index.php/jair/article/view/10302.

Chen, R., Xie, Y., Cao, Y., Tang, F., Zhang, J., Li, X., Ji, Y., Liu, Q., Zhao, H., and Wang, F. Predictive modeling with temporal graphical representation on electronic health records. *arXiv*, 2024. doi: 10.48550/arXiv.2405.03943. URL https://arxiv.org/abs/2405.03943.

Cheng, J. and Hu, G. Extending cox proportional hazards model with symbolic non-linear log-risk functions for survival analysis. *arXiv*, 2025. doi: https://arxiv.org/html/2504.04353v2. URL https://arxiv.org/html/2504.04353v2.

Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. Gram: Graph-based attention model for healthcare representation learning. *Association for Computing Machinery*, 2017. doi: 10.1145/3097983.3098126. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7954122/.

Cox, D. R. Partial likelihood. *Oxford University Press*, 1975. doi: 10.1093/biomet/62.2.269. URL https://academic.oup.com/biomet/article/62/2/269/236568.

Daniore, P. et al. From wearable sensor data to digital biomarker development. *Nature Portfolio*, 2024. doi: 10.1038/s41746-024-01151-3. URL https://www.nature.com/articles/s41746-024-01151-3.

Decencière, E., Zhang, X., Cazuguel, G., Laÿ, B., Cochener, B., Trone, C., Gain, P., Ordóñez, J.-R., Massin, P., Erginay, A., Charton, B., and Klein, J.-C. Feedback on a publicly distributed image database: The messidor database. *Image Analysis & Stereology*, 2014. doi: 10.5566/ias.1155. URL https://www.ias-iss.org/ojs/IAS/article/view/1155.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *IEEE*, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848.

Dong, Y., Wu, X., Xiao, Y., Chong, J. S. X., Jin, Y., and Zhou, J. H. Lighted: An interpretable recurrent neural network with long- and short-term graph propagation for electronic health records. *Elsevier*, 2022. doi: 10.1016/j.artmed.2022.102439. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9394492/.

Dong, Z., Li, R., Wu, Y., Nguyen, T. T., Chong, J. S. X., Ji, F., Tong, N. R. J., Chen, C. L. H., and Zhou, J. H. Brainjepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *arXiv*, 2024. doi: 10.48550/ARXIV.2409.19407. URL https://arxiv.org/abs/2409.19407.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. doi: 10.48550/ARXIV.2010.11929. URL https://arxiv.org/abs/2010.11929.

Drozdov, K., Shwartz-Ziv, R., and LeCun, Y. Video representation learning with joint-embedding predictive architectures. *arXiv*, 2024. doi: 10.48550/ARXIV.2412.10925. URL https://arxiv.org/abs/2412.10925.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Portfolio*, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. *arXiv*, 2017. doi: 10.48550/ARXIV.1704.01212. URL https://arxiv.org/abs/1704.01212.

Grill, J.-B., Strub, F., Altche, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv*, 2020. doi: 10.48550/ARXIV.2006.07733. URL https://arxiv.org/abs/2006.07733.

Hamilton, W. L., Ying, R., and Leskovec, J. Inductive representation learning on large graphs. *arXiv*, 2017. doi: 10.48550/ARXIV.1706.02216. URL https://arxiv.org/abs/1706.02216.

Hancox, Z., Kingsbury, S. R., Clegg, A., Conaghan, P. G., and Relton, S. D. Developing the temporal graph convolutional neural network model to predict hip replacement using electronic health records. *arXiv*, 2024. doi: 10.48550/arXiv.2409.06585. URL https://arxiv.org/abs/2409.06585.

Harrell, F. E. Regression modeling strategies. *Springer Nature*, 2015. doi: 10.1007/978-3-319-19425-7. URL https://link.springer.com/book/10.1007/978-3-319-19425-7.

Heagerty, P. J., Lumley, T., and Pepe, M. S. Time-dependent roc curves for censored survival data and a diagnostic marker. *Oxford University Press*, 2000. doi: 10.1093/biostatistics/1.3.277. URL https://academic.oup.com/biostatistics/article/1/3/277/245074.

Holme, P. and Saramäki, J. Temporal networks. *Elsevier*, 2012. doi: 10.1016/j.physrep.2012.03.001. URL https://www.sciencedirect.com/science/article/pii/S0370157312000841.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv*, 2021. doi: 10.48550/ARXIV.2106.09685. URL https://arxiv.org/abs/2106.09685.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *arXiv*, 2020a. doi: 10.48550/ARXIV.2005.00687. URL https://arxiv.org/abs/2005.00687.

Hu, Z., Dong, Y., Wang, K., and Sun, Y. Heterogeneous graph transformer. *Association for Computing Machinery*, 2020b. doi: 10.1145/3366423.3380027. URL https://dl.acm.org/doi/10.1145/3366423.3380027.

Huang, S.-C., Shen, L., Lungren, M. P., and Yeung, S. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *Nature Portfolio*, 2023. doi: 10.1038/s41746-023-00811-0. URL https://www.nature.com/articles/s41746-023-00811-0.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Springer Nature*, 2016. doi: 10.1038/sdata.2016.35. URL https://www.nature.com/articles/sdata201635.

Kalapos, A. and Gyires-Tóth, B. Cnn-jepa: Self-supervised pretraining convolutional neural networks using joint embedding predictive architecture. *arXiv*, 2024. doi: 10.48550/ARXIV.2408.07514. URL https://arxiv.org/abs/2408.07514.

Kapoor, S. and Narayanan, A. Leakage and the reproducibility crisis in ml-based science. *arXiv*, 2022. doi:

10.48550/ARXIV.2207.07048. URL https://arxiv.org/abs/2207.07048.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC*, 2018. doi: 10.1186/s12874-018-0482-1. URL https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1.

Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., Wu, S., Smyth, C., Poupart, P., and Brubaker, M. Time2vec: Learning a vector representation of time. *arXiv*, 2019. doi: 10.48550/ARXIV.1907.05321. URL https://arxiv.org/abs/1907.05321.

Khera, R., Oikonomou, E. K., Nadkarni, G. N., Morley, J. R., Wiens, J., et al. Transforming cardiovascular care with artificial intelligence: From discovery to practice. *Elsevier*, 2024. doi: 10.1016/j.jacc.2024.05.003. URL https://www.jacc.org/doi/10.1016/j.jacc.2024.05.003.

Kim, M., Kim, J., Qu, J., Huang, H., Long, Q., Sohn, K.-A., Kim, D., and Shen, L. Interpretable temporal graph neural network for prognostic prediction of alzheimer's disease using longitudinal neuroimaging data. *IEEE*, 2021. doi: 10.1109/BIBM52615.2021.9669504. URL https://pubmed.ncbi.nlm.nih.gov/35299717/.

Kipf, T. N. and Welling, M. Variational graph auto-encoders. *arXiv*, 2016. doi: 10.48550/ARXIV.1611.07308. URL https://arxiv.org/abs/1611.07308.

Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., and Sancho-Gómez, J.-L. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Nature Portfolio*, 2022. doi: 10.1038/s41597-022-01388-1. URL https://www.nature.com/articles/s41597-022-01388-1.

LeCun, Y. A path towards autonomous machine intelligence. *OpenReview*, 2022. URL https://openreview.net/pdf?id=BZ5a1r-kVsf.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature Portfolio*, 1999. doi: 10.1038/44565. URL https://www.nature.com/articles/44565.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Oxford University Press*, 2019. doi: 10.1093/bioinformatics/btz682. URL https://arxiv.org/abs/1901.08746.

Li, Q., Han, Z., and Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. *arXiv*, 2018. doi: 10.48550/ARXIV.1801.07606. URL https://arxiv.org/abs/1801.07606.

Longa, A. et al. Graph neural networks for temporal graphs: State of the art and future directions. *arXiv*, 2023. doi: 10.48550/ARXIV.2305.14447. URL https://openreview.net/pdf?id=pHCdMat0gI.

Lu, C., Reddy, C. K., and Ning, Y. Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE*, 2023. doi: 10.1109/TCYB.2021.3109881. URL https://pubmed.ncbi.nlm.nih.gov/34546938/.

Luo, J., Wang, X., Fan, X., He, Y., Du, X., Chen, Y.-Q., and Zhao, Y. A novel graph neural network based approach for influenza-like illness nowcasting: exploring the interplay of temporal, geographical, and functional spatial features. *BMC Public Health*, 2025. doi: 10.1186/s12889-025-21618-6. URL https://pubmed.ncbi.nlm.nih.gov/39893390/.

Micikevicius, P., Narang, S., Alben, J., et al. Mixed precision training. *arXiv*, 2018. doi: 10.48550/ARXIV.1710.03740. URL https://arxiv.org/abs/1710.03740.

Mo, S. and Tong, S. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. *arXiv*, 2024. doi: 10.48550/ARXIV.2410.19560. URL https://arxiv.org/abs/2410.19560.

Mukherjee, N., Sengupta, S., Ahmed, M. N., Yaqoob, S. I., Hussain, M. R., and Zamani, A. T. Bi-directional hybrid attention feature pyramid network for detecting diabetic macular edema in retinal fundus images. *IEEE Access*, 2025. doi: 10.1109/ACCESS.2025.3545873. URL https://ieeexplore.ieee.org/document/10904230.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. *arXiv*, 2020. doi: 10.48550/ARXIV.1905.10947. URL https://arxiv.org/abs/1905.10947.

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., and Igel, C. U-sleep: resilient high-frequency sleep staging. *Nature Portfolio*, 2021. doi: 10.1038/s41746-021-00440-5. URL https://www.nature.com/articles/s41746-021-00440-5.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., and Webster, D. R. Prediction of cardiovascular risk factors from retinal fundus photographs via deep

learning. *Nature Portfolio*, 2018. doi: 10.1038/s41551-018-0195-0. URL https://www.nature.com/articles/s41551-018-0195-0.

Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, S.-H., Liu, Y., Wang, G., et al. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *MDPI*, 2018. doi: 10.3390/data3030025. URL https://www.mdpi.com/2306-5729/3/3/25.

Qian, Y., Li, L., Nakashima, Y., Nagahara, H., Nishida, K., and Kawasaki, R. Is cardiovascular risk profiling from uk biobank retinal images using explicit deep learning estimates of traditional risk factors equivalent to actual risk measurements? a prospective cohort study design. *BMJ*, 2024. doi: 10.1136/bmjopen-2023-078609. URL https://pubmed.ncbi.nlm.nih.gov/39384229/.

Quellec, G., Lamard, M., Lay, B., Le Guilcher, A., Erginay, A., Cochener, B., and Massin, P. Instant automatic diagnosis of diabetic retinopathy. *arXiv*, 2019. doi: 10.48550/arXiv.1906.11875. URL https://arxiv.org/abs/1906.11875.

Reicher, L., Shilo, S., Godneva, A., Lutsker, G., Zahavi, L., Shoer, S., Krongauz, D., Rein, M., Kohn, S., Segev, T., Schlesinger, Y., Barak, D., Levine, Z., Keshet, A., Shaulitch, R., Lotan-Pompan, M., Elkan, M., Talmor-Barkan, Y., Aviv, Y., Dadiani, M., Tsodyks, Y., Nili Gal-Yam, E., et al. Deep phenotyping of health–disease continuum in the human phenotype project. *Nature Portfolio*, 2025. doi: 10.1038/s41591-025-03790-9. URL https://pubmed.ncbi.nlm.nih.gov/40665053/.

Rim, T. H., Lee, C. J., Tham, Y.-C., Cheung, N., Yu, M., Lee, G., Kim, Y., Ting, D. S. W., Chong, C. C. Y., Choi, Y. S., Yoo, T. K., Ryu, I. H., Baik, S. J., Kim, Y. A., Kim, S. K., Lee, S.-H., Lee, B. K., Kang, S.-M., Wong, E. Y. M., Kim, H. C., Kim, S. S., Park, S., Cheng, C.-Y., and Wong, T. Y. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. *Elsevier*, 2021. doi: 10.1016/S2589-7500(21)00043-1. URL https://pubmed.ncbi.nlm.nih.gov/33890578/.

Rossi, E., Kenlay, H., Havakhor, S., Monti, F., and Bronstein, M. Temporal graph networks for deep learning on dynamic graphs. *arXiv*, 2020. doi: 10.48550/arXiv.2006.10637. URL https://arxiv.org/abs/2006.10637.

Saito, T. and Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *Public Library of Science*, 2015. doi: 10.1371/journal.pone.0118432. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432.

Schlichtkrull, M. S., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. Modeling relational data with graph convolutional networks. *Springer*, 2018. doi: 10.1007/978-3-319-93417-4_38. URL https://link.springer.com/chapter/10.1007/978-3-319-93417-4_38.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., Baby, S. A., Baby, S. M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Riviere, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-b., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J.-B., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J., Warkentin, T., Shlens, J., Fleet, D., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D. F., Kirmizibayrak, C., Pilgrim, R., Golden, D., and Yang, L. c. m. m. P. d.-c. r. Medgemma technical report. *arXiv*, 2025. URL https://arxiv.org/abs/2507.05201.

Shajari, S., Kuruvinashetti, K., Komeili, A., and Sundararaj, U. The emergence of ai-based wearable sensors for digital health technology: A review. *MDPI*, 2023. doi: 10.3390/s23239498. URL https://www.mdpi.com/1424-8220/23/23/9498.

Shilo, S., Bar, N., Keshet, A., Talmor-Barkan, Y., Rossman, H., Godneva, A., Aviv, Y., Edlitz, Y., Reicher, L., Kolobkov, D., Wolf, B. C., Lotan-Pompan, M., Levi, K., Cohen, O., Saranga, H., Weinberger, A., and Segal, E. 10k: a large-scale prospective longitudinal study in israel. *Springer Nature*, 2021. doi: 10.1007/s10654-021-00753-5. URL https://pubmed.ncbi.nlm.nih.gov/33993378/.

Skarding, J., Gabrys, B., and Musial, K. Foundations of graph neural networks for temporal graphs. *IEEE*, 2021. doi: 10.1109/ACCESS.2021.3082932. URL https://ieeexplore.ieee.org/document/9444729.

Thapa, R., He, B., Kjaer, M. R., Moore IV, H., Ganjoo, G., Mignot, E., and Zou, J. Y. Sleepfm: Multimodal representation learning for sleep across brain activity, ecg and respiratory signals. *arXiv*, 2024. doi:

10.48550/ARXIV.2405.17766. URL https://arxiv.org/abs/2405.17766.

Tipirneni, S. et al. Sparse and irregularly sampled multivariate clinical time series: Challenges and opportunities. *Association for Computing Machinery*, 2022. doi: 10.1145/3516367. URL https://dl.acm.org/doi/10.1145/3516367.

van Houwelingen, H. and Putter, H. *Dynamic Prediction in Clinical Survival Analysis*. 2011. doi: 10.1201/b11311. URL https://doi.org/10.1201/b11311.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. Attention is all you need. *arXiv*, 2017. doi: 10.48550/ARXIV.1706.03762. URL https://arxiv.org/abs/1706.03762.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. *arXiv*, 2018. doi: 10.48550/ARXIV.1710.10903. URL https://arxiv.org/abs/1710.10903.

Wang, X., Ji, H., Shi, C., Wang, B., Cui, P., Yu, P. S., and Ye, Y. Heterogeneous graph attention network. *Association for Computing Machinery*, 2019. doi: 10.1145/3308558.3313562. URL https://dl.acm.org/doi/10.1145/3308558.3313562.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. Graph convolutional neural networks for web-scale recommender systems. *Association for Computing Machinery*, 2018. doi: 10.1145/3219819.3219890. URL https://dl.acm.org/doi/10.1145/3219819.3219890.

Zeng, H., Zhou, H., Srivastava, A., Kannan, R., and Prasanna, V. Graphsaint: Graph sampling based inductive learning method. *arXiv*, 2020. doi: 10.48550/ARXIV.1907.04931. URL https://arxiv.org/abs/1907.04931.

Zhang, Y., Han, Y., Cao, S., Dai, G., Miao, Y., Cao, T., Yang, F., and Xu, N. Adam accumulation to reduce memory footprints of both activations and gradients for large-scale dnn training. *arXiv*, 2023. doi: https://arxiv.org/pdf/2305.19982. URL https://arxiv.org/pdf/2305.19982.

Zheng, Y., Jiang, W., Zhou, A., Hung, N. Q. V., Zhan, C., and Chen, T. Epidemiology-informed graph neural network for heterogeneity-aware epidemic forecasting (heatgnn). *arXiv*, 2024. doi: 10.48550/arXiv.2411.17372. URL https://arxiv.org/abs/2411.17372.

Zhou, Y., Wagner, S. K., Chia, M. A., Zhao, A., Woodward-Court, P., Xu, M., Struyven, R., Alexander, D. C., and Keane, P. A. Automorph: Automated retinal vascular morphology quantification via a deep learning pipeline. *ARVO*, 2022. doi: 10.1167/tvst.11.7.12. URL https://doi.org/10.1167/tvst.11.7.12.