IDENTIFYING CRITICAL PHASES FOR DISEASE ONSET WITH SPARSE HAEMATOLOGICAL BIOMARKERS

Andrea Zerio[1], Maya Bechler-Speicher[2,3], Tine Jess[1,4] & Aleksejs Sazonovs[1]
[1] Center of Excellence for Molecular Prediction of Inflammatory Bowel Disease, PREDICT, Department of Clinical Medicine, Aalborg University
[2] Blavatnik School of Computer Science, Tel-Aviv University
[3] Meta
[4] Department of Gastroenterology & Hepatology, Aalborg University Hospital
{anze, jess, alesaz}@dcm.aau.dk mayab4@mail.tau.ac.il

1 TEMPORALLY SPARSE BIOMARKER DATA

Haematological biomarkers from clinical chemistry tests are widely used in medical practice, generating large-scale molecular data that can support health and disease research (Uttley et al., 2016; Foy et al., 2025). Many of these biomarker values and their dynamics are known to be strong indicators of health-related traits. Emerging evidence indicates that many complex diseases, such as immune-mediated diseases (IMIDs), exhibit pre-diagnostic stages that can be inferred from these biomarkers (Vestergaard et al., 2023; Deane et al., 2010). A pre-diagnostic stage refers to a phase where a patient has not yet met the clinical criteria for diagnosis, yet subtle, systemic changes in their biomarker profiles suggest an elevated risk of disease onset.

Detecting early dysregulation in biomarker patterns is crucial for enabling timely preventative interventions. Unfortunately, routine clinical sampling is guided by medical needs rather than standardized research protocols, introducing confounding noise. As such, similar biomarker values may be observed across multiple conditions, making it difficult to distinguish disease-specific patterns and reducing predictive specificity. More importantly, it results in irregular sampling intervals, introducing sparsity into the data's temporal dimension. All this makes it difficult to apply standard time-series models, which often rely on interpolation or imputation to fill in missing data (Herbers et al., 2021; Ahmed et al., 2023). Such preprocessing can obscure true biological signals, distorting learning patterns, reducing predictive accuracy, and compromising interpretability.

2 DETECTING DISEASE ONSET FROM SPARSE BIOMARKERS WITH INTERPRETABLE GRAPH LEARNING

Our ultimate goal is to identify biomarker dysregulation periods predictive of disease onset. To address the sampling challenges, we model biomarker trajectories as time-weighted directed graphs, preserving temporal structure without imputation or zero-inflation. In this framework, detecting dysregulation reduces to identifying the key nodes that correspond to critical periods.

For each individual, we define a graph composed of multiple longitudinal trajectories, one for each of the biomarkers measured throughout their history. These trajectories are represented as directed line graphs $G_k = (V, E)$, where $V = \{v_1, v_2, \ldots, v_T\}$ is the set of nodes, with each node v_t representing a sampling event at time t, and where $E = \{(v_1, v_2), (v_2, v_3), \ldots, (v_{T-1}, v_T)\}$ is the set of edges. To encode the temporal structure of the data, each edge $(v_t, v_{t+1}) \in E$ is assigned a weight $w_t = \rho(\Delta_t)$, computed as a function of the time interval $\Delta_t = t_{t+1} - t_t$ between consecutive sampling events, where ρ is a weighting function that maps the time interval Δ_t to a scalar.

Since our problem formulation aims to detect important nodes, we leverage and extend the recently proposed interpretable GNAN (Bechler-Speicher et al., 2024). Unlike black-box deep learning models, GNANs provide intrinsically interpretable predictions by constraining the use of feature cross-products and graph topology, resulting in a transparent architecture which provides node and feature importance metrics. This allows us to trace which biomarkers and time points contribute most to a classification decision, shifting the focus to explaining when and how disease-related changes

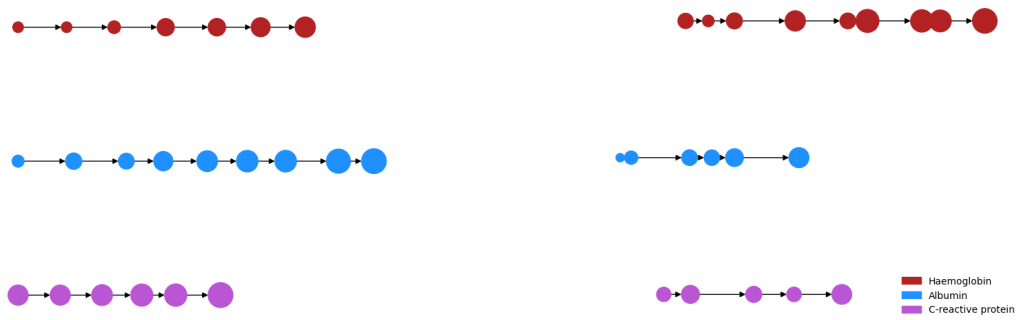


Figure 1: Diagram illustrating node-level interpretability of synthetically generated data. The size of each biomarker node corresponds to the importance that the model assigned to it. Individual-level trajectories cannot be provided due to Danish data protection laws.

emerge. We extend the original GNAN formulation to generate node representations as shown in Figure 1:

$$[\mathbf{h}_i]_k = \sum_{j \in V} \rho \left(\frac{1}{0.1 + \Delta t_{ji}} \right) f_k(\mathbf{x}_j^{(k)}), \tag{1}$$

where $[\mathbf{h}_i]_k$ is the k-th entry of the representation of node i, denoted as $\mathbf{h}_i$. This formulation simplifies the distance function of the original GNAN by leveraging symmetries to make the computation of node distances efficient. Additionally, we incorporate a one-dimensional representation of the biomarker's one-hot encoding to capture categorical information alongside continuous trajectories. For the full formulation see Appendix A.1.

We demonstrate our approach using binary classification of the pre-diagnostic trajectories of patients and age-matched controls. We select a set of 2,500 Crohn's Disease (CD) patients and 2,500 controls, sampling the trajectories of three routine clinical biomarkers: haemoglobin, albumin, and C-reactive protein. These biomarkers were chosen for their widespread use in routine clinical testing and their statistical association with the pre-CD disease state (Vestergaard et al., 2023).

As a work-in-progress, the model's current performance is not yet sufficient for practical application, indicating the need for further development and optimization. However, early results suggest that the model is learning to capture at least some meaningful signals, as shown in Appendix A.3. In order to understand whether the representations of the graphs that the model learns are biologically coherent, we freeze the model and compute the importance of each node in patients' graphs after a single forward pass. To comply with privacy and data-sharing regulations we provide diagrams from synthetic data of node-level interpretability plots in Figure 1.

Our initial results suggest that the model assigns varying importance to nodes, potentially distinguishing between biomarkers and across time. Notably, there is anecdotal evidence that it prioritizes clusters near the trajectory's end, implying that recent biomarker dynamics could matter more than distant history. This aligns with previous findings that the association strength of the pre-diagnostic signal peaks in the 1–2 years preceding diagnosis (Vestergaard et al., 2023).

3 Conclusion

We introduce a novel GNAN-based representation for sparse, temporal biomarker trajectories. Our approach, still in early development, shows promise in learning sparse representations of haematological biomarkers and providing insights into node-level feature importance. This is crucial for clinical and biological applications, where interpretability and efficient representation are key for decision-making and detecting early disease signs. Future work will explore extending the GNAN model further by using recurrent architectures to model temporal nodes, expand support for a broader set of biomarkers, integrate non-biomarker features (e.g., comorbidities), and model interactions across multiple omics layers.

MEANINGFULNESS STATEMENT

Understanding health and disease requires models that extract biologically relevant patterns from complex, sparse, and available data. Patient medical history contains an incomplete yet invaluable imprint of one of the most important aspects of life: our health. In this work, we propose using data from routine blood tests, something that each of us has experienced in our lifetime, to learn biomarker trajectories representative of our health. While we demonstrate our approach on CD, its potential extends far beyond, offering a versatile framework for uncovering critical health patterns in any longitudinal setting.

REFERENCES

Khandakar Tanvir Ahmed, Sze Cheng, Qian Li, Jeongsik Yong, and Wei Zhang. Incomplete time-series gene expression in integrative study for islet autoimmunity prediction. *Briefings in Bioinformatics*, 24(1):bbac537, 2023.

Maya Bechler-Speicher, Amir Globerson, and Ran Gilad-Bachrach. The intelligible and effective graph neural additive network. *Advances in Neural Information Processing Systems*, 37:90552–90578, 2024.

Kevin D Deane, Jill M Norris, and V Michael Holers. Preclinical rheumatoid arthritis: identification, evaluation, and future directions for investigation. *Rheumatic Disease Clinics*, 36(2):213–241, 2010.

Brody H Foy, Rachel Petherbridge, Maxwell T Roth, Cindy Zhang, Daniel C De Souza, Christopher Mow, Hasmukh R Patel, Chhaya H Patel, Samantha N Ho, Evie Lam, et al. Haematological setpoints are a stable and patient-specific deep phenotype. *Nature*, 637(8045):430–438, 2025.

Judith Herbers, Robert Miller, Andreas Walther, Lena Schindler, Kornelius Schmidt, Wei Gao, and Florian Rupprecht. How to deal with non-detectable and outlying values in biomarker research: Best practices and recommendations for univariate imputation approaches. *Comprehensive Psychoneuroendocrinology*, 7:100052, 2021.

Aleksejs Sazonovs[*], Kirsten Schut[*], Nikolas Plevris, Filip A Ottosson, Tine Jess[#], Jeffrey C Barrett[#], and Charlie W Lees[#]. OP19 Pre-and post-diagnostic metabolomic biomarker profiling of over 700,000 individuals in three national biobanks enables prediction of inflammatory bowel disease onset and complications. *Journal of Crohn's and Colitis*, 19(Supplement_1):i38–i41, 2025. [*] Co-first authors. [#] Co-senior author.

Lesley Uttley, Becky L Whiteman, Helen Buckley Woods, Susan Harnan, Sian Taylor Philips, and Ian A Cree. Building the evidence base of blood-based biomarkers for early detection of cancer: a rapid systematic mapping review. *EBioMedicine*, 10:164–173, 2016.

Marie Vibeke Vestergaard, Kristine H Allin, Gry J Poulsen, James C Lee, and Tine Jess. Characterizing the pre-clinical phase of inflammatory bowel disease. *Cell Reports Medicine*, 4(11), 2023.

A APPENDIX

A.1 TIME-DELTA GNAN FORMULATION

Given a node feature vector consisting of d individual features, each feature is treated as a separate univariate series and processed by its corresponding function f_k. For node i, the feature representation is:

$$[\mathbf{h}_i]_k = \sum_{j \in V} \rho\left(\frac{1}{0.1 + \Delta t_{ji}}\right) f_k(\mathbf{x}_j^{(k)}), \tag{2}$$

where ρ is a distance weighting function and f_k processes the corresponding feature set S_k.

This simplifies the original GNAN formulation in two ways:

1. The original $\#\text{dist}_i(j)$ function is replaced with Δt_{ji}, which is computationally more efficient to calculate.
2. Since Δt_{ji} represents the difference in time between two dates in a directed graph, every valid path between two nodes is equal in length. As such, we can remove the normalisation term $\frac{1}{\#\text{dist}_i(j)}$.

In addition to processing each biomarker feature independently, we introduce an additional one-dimensional, node-level feature derived from the one-hot encoding of the biomarker. This feature is computed using a multivariate function F_{oh}, which takes as input the one-hot encoded representation of the biomarker across nodes and aggregates it accordingly. Specifically, for node i, this additional feature is given by:

$$[\mathbf{h}_i]_{\text{oh}} = \sum_{j \in V} \rho\left(\frac{1}{0.1 + \Delta t_{ji}}\right) F_{\text{oh}}(\mathbf{x}_j^{(\text{oh})}), \tag{3}$$

where $\mathbf{x}_j^{(\text{oh})}$ represents the one-hot encoding of the biomarker at node j. This ensures that GNAN captures categorical biomarker information alongside continuous biomarker trajectories.

The final node representation is:

$$\mathbf{h}_i = \big([\mathbf{h}_i]_1, [\mathbf{h}_i]_2, \ldots, [\mathbf{h}_i]_d, [\mathbf{h}_i]_{\text{oh}}\big). \tag{4}$$

For graph-level tasks, we apply sum pooling:

$$\mathbf{h}_G = \sum_{i \in V} \mathbf{h}_i, \tag{5}$$

followed by a readout function for prediction:

$$\sigma\left(\sum_{k=1}^{d} [\mathbf{h}_G]_k\right), \tag{6}$$

where σ is an activation function (e.g., sigmoid for classification). Importantly, GNAN enables interpretability by quantifying the influence of node j on feature set S_k:

$$\text{Influence}(j, S_k, G) = f_k(\mathbf{x}_j^{(k)}) \sum_{i \in V} \rho\left(\frac{1}{0.1 + \Delta t_{ji}}\right), \tag{7}$$

allowing us to identify critical time points that contribute most to the prediction. The total contribution of node i to the graph-level decision is:

$$\text{TotalContribution}(i) = \sum_{j \in V} \rho\left(\frac{1}{0.1 + \Delta t_{ji}}\right) \sum_{k=1}^{d} f_k(\mathbf{x}_j^{(k)}). \tag{8}$$

This formulation ensures that GNAN not only models sparse biomarker trajectories effectively, but also provides an interpretable framework for identifying critical phases in disease progression.

A.2 Experimental Details

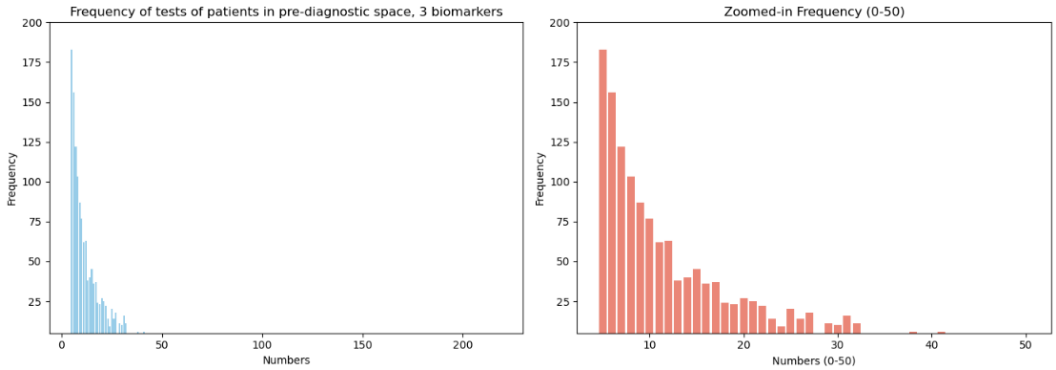
A.2.1 Model Training

We trained a series of FeatureGroupGNAN models, all with ReLU activations, layers in the {3,5} and hidden channels in the {100, 64} range. We used an Adam optimizer over 10 epochs with a custom CosineAnnealingWarmRestartsDecay scheduler, with T0 in the {10, 50} range, decay factor in the {0.3, 0.8} range, learning rate in the {1e-5, 5e-5} range and minimum learning rate in the {5e-8, 5e-9} range.

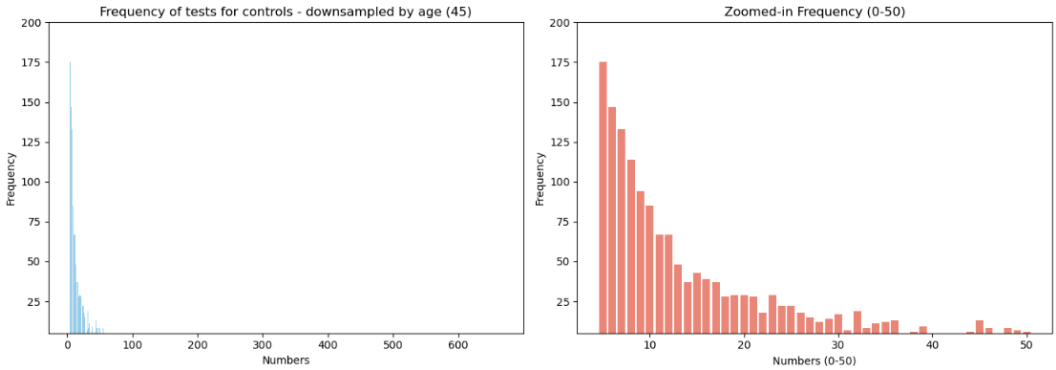
All models were trained on a single NVIDIA Tesla V100-PCIE-16GB GPU.

A.2.2 Dataset

We train and experiment on a dataset of 2,500 individuals who eventually develop CD and 2,500 controls. The patients are sampled at random from a larger nation-wide cohort of CD patients. Similarly the controls are sampled at random from a cohort of around 9 million individuals who were never diagnosed with CD. All data referenced has been obtained from the Danish healthcare registries. We select only the pre-diagnostic trajectories of patients, meaning the biomarkers that were sampled before a formal diagnosis. We then downsampled blood tests from controls to align with the typical age of onset in CD, ensuring a comparable testing frequency between controls and patients. The resulting distributions over the number of tests per-person is displayed in Figure 2. To comply with privacy and data-sharing regulation we exclude all samples with less than 5 measurements before computing the adjusted distributions and producing the plots.



(a) Patient Data Distribution

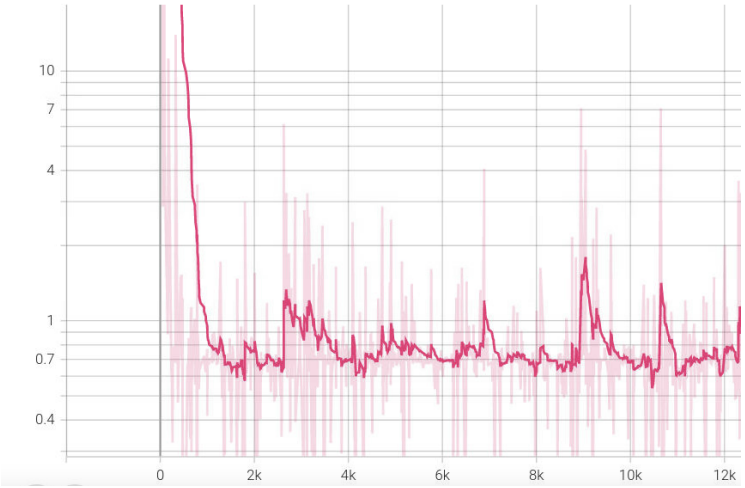


(b) Control Data Distribution

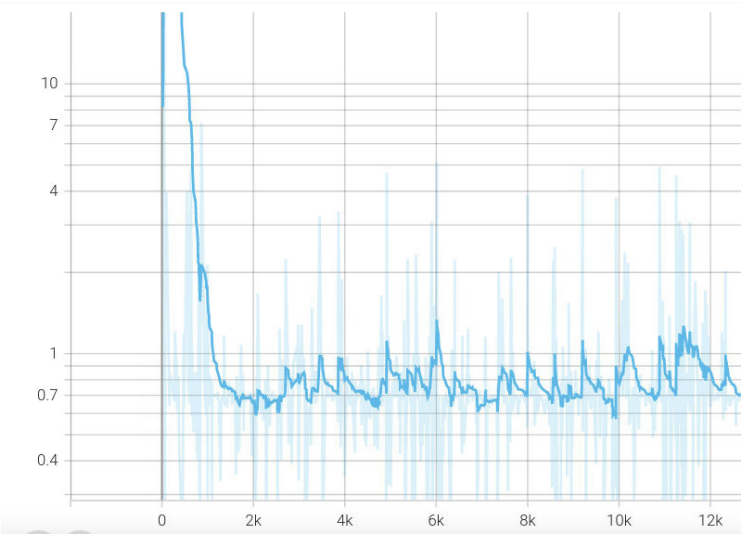
Figure 2: Comparison of patient and control data distributions after downsampling controls by age. Values below n=5 were excluded due to Danish data protection rules.

A.3 Initial Performance Details

Although still in early development, our model achieves performance comparable to baseline methods reported in UK Biobank (UKBB) studies for similar biomarkers (Sazonovs* et al., 2025). However, a key distinction is that these baseline models were evaluated on data with homogeneous time-point sampling, whereas our model was developed and tested on temporally sparse trajectories. Furthermore, UKBB data is subject to higher informational bias, as control participants were generally healthier due to the requirement to attend one of the recruitment facilities. In contrast, control data in our Danish cohort was hospital-sampled, likely representing a population with a worse overall health profile.

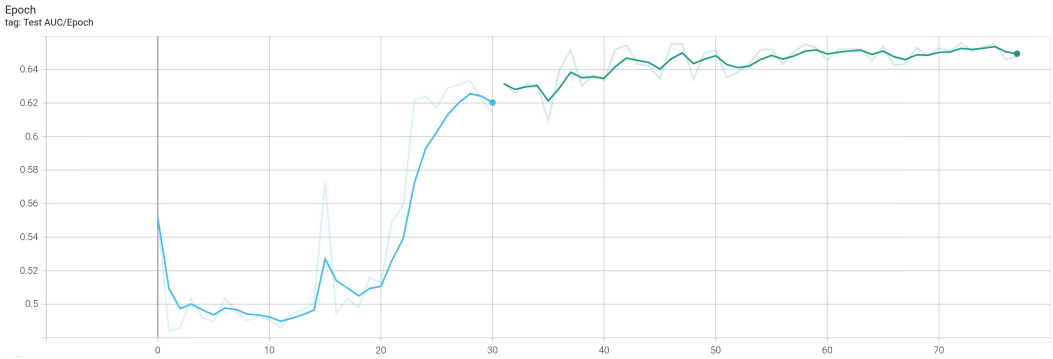


(a) Higher learning rate decay and faster restart

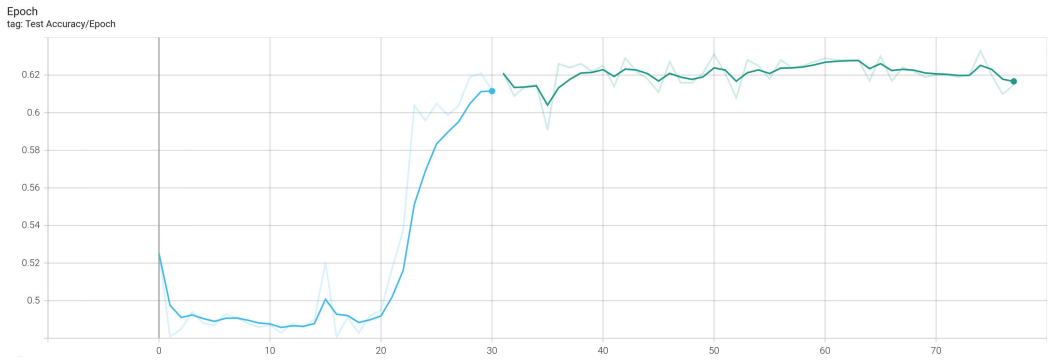


(b) Lower learning rate decay and slower restart

Figure 3: Logspace of batch training loss (BCEWithLogitsLoss). The plots demonstrate the model is capable of learning some initial signal, before finding a local minimum early on and stabilising.

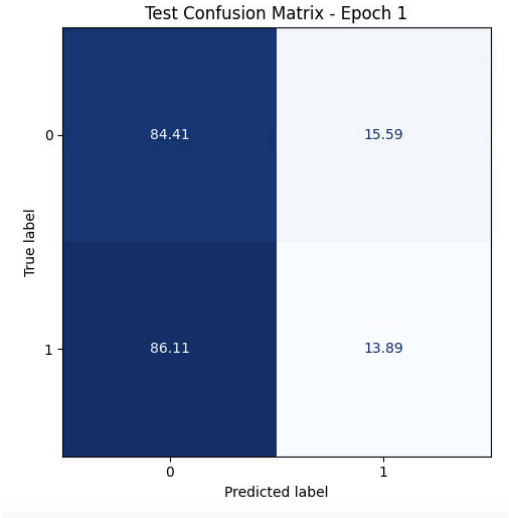


(a) Test Area under the curve (AUC)

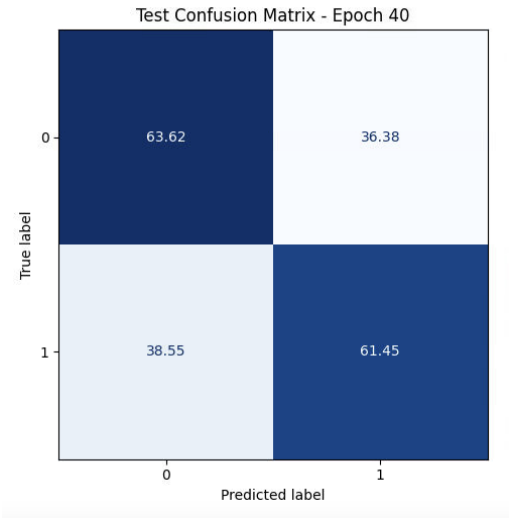


(b) Test Accuracy

Figure 4: Test AUC and Accuracy across training. The two lines in each plot correspond to different segments of the same training run, where training was resumed from a checkpoint after reaching the initial stopping point



(a) Test Confusion Matrix after 1 epoch



(b) Test Confusion Matrix after 40 epochs

Figure 5: Comparison of test set confusion matrices. Despite not yet being performant enough to be clinically relevant, the model does seem to learn some initial signal that is discriminative of CD patients and controls.

A.4 Funding and Acknowledgement

The work was supported by grant DNRF148 from the Danish National Research Foundation Center of Excellence and grant NNF23OC0087616 from the Novo Nordisk Foundation. We would like to thank Marie Vibeke Vestergaard for her input and advice with regards to this manuscript.