MDPI

Article

Deep Learning Approaches to IBD Patient Classification: A Comparative Study of GNNs and Neural Networks on Olink/UK Biobank Data

Francisco J. Salamanca [1,2] and Jorge Morales [3]

[1] MSc Bioinformatics Student, Universidad Nacional de Colombia ; fsalamancar@unal.edu.co
[2] Institute of Clinical Molecular Biology, Kiel University, University Hospital Schleswig-Holstein, Kiel, Germany
[3] MSc System Engineering Student, Universidad Nacional de Colombia ; jomorales@unal.edu.co

Abstract: Inflammatory Bowel Disease (IBD) is a chronic condition with complex molecular underpinnings, for which early and accurate diagnosis is essential. In this study, we explore predictive models using proteomic profiling data from 22,625 individuals in the UK Biobank—18,999 healthy controls and 3,626 diagnosed with IBD. To enhance the biological relevance of the input features, we integrate protein-protein interaction (PPI) information from the STRING database with raw protein expression data obtained via Olink assays. This transformation enables both matrix-based neural network models and graph-based learning.

We compare the performance of classical dense neural networks with Graph Neural Networks (GNNs), using the PPI network as the structural backbone for GNN input. The dense models are optimized via grid search over learning rate, architecture, batch size, and epochs, while addressing class imbalance through SMOTETomek resampling and focal loss. The best dense network configuration achieved a test accuracy of 83.14% and an F1-score of 0.29 for the disease class, showing promising sensitivity to minority cases.

In parallel, GNNs were implemented to leverage the relational structure of the proteomic data, evaluating their capacity to capture topological dependencies in protein interactions. This comparative study offers insights into the trade-offs and complementary strengths of classical and graph-based deep learning methods for large-scale proteomics-driven disease prediction.

Keywords: Inflammatory Bowel Disease (IBD); Proteomics; UK Biobank; Olink protein assays; Protein-Protein Interaction (PPI); STRING database; Deep Learning; Neural Networks; Graph Neural Networks (GNNs)

Received:
Revised:
Accepted:
Published:

Citation: Salamanca, F.; Morales, J. Deep Learning Approaches to IBD Patient Classification: A Comparative Study of GNNs and Neural Networks on Olink/UK Biobank Data. *Journal Not Specified* **2025**, *1*, 0. https://doi.org/

1. Introduction

Inflammatory bowel disease (IBD), which includes Crohn's disease (CD) and ulcerative colitis (UC), is a chronic relapsing disorder characterized by an increasing global prevalence and marked clinical heterogeneity.

The diversity of disease manifestations, accompanied by progression and therapeutic response, complicates prognosis, treatment selection, and long-term monitoring, highlighting the urgent need to personalize approaches in patient management [1]).

Although advances in genomic research have identified more than 200 risk loci associated with susceptibility to IBD, these findings explain only a fraction of the disease variability, leaving critical gaps in understanding functional mechanisms and elaborable therapeutic targets [2,3]).

There is increasing evidence that circulating proteins are dynamic mediators of disease pathways, offering insights into the pathophysiology of IBD that genomics alone cannot capture. Unlike genetic variants, proteins reflect real-time biological processes, such as interactions with environmental triggers, immune dysregulation, and tissue damage. Despite extensive proteomic profiling in healthy populations, systematic analyzes of IBD remain scarce, limiting the translation of proteomic findings into clinical tools for diagnosis, disease monitoring, or targeted therapy [3,4]. Furthermore, the integration of proteomic data with genetic risk variants, called proteogenomic approaches, holds promise to elucidate the mechanisms underlying known loci, identify new biomarkers, and accelerate drug development [5,6].

This study uses proteomic information from the UK Biobank as a result of its influence in the detection of IBD. In addition to this information, it was decided to use the STRING database which contains protein-protein interaction data which could be useful to the models proposed.

2. Methods

2.1. Data Representation and Preprocessing

The dataset used in this study originates from the UK Biobank, comprising proteomic profiles for 22,625 individuals obtained via Olink® technology. The dataset includes 18,999 healthy (control) individuals and 3,626 with Inflammatory Bowel Disease (IBD), indicating a notable class imbalance.

To incorporate functional biological relationships, the original feature matrix—containing protein expression levels—was transformed using a protein-protein interaction (PPI) matrix derived from the STRING database. This matrix multiplication generated a new representation of each sample that reflects interaction-aware protein expression profiles, effectively introducing prior biological knowledge into the modeling process.

2.2. Graph Neural Network Architecture

The proposed model, named PatientClassifierGNN, is a hybrid deep learning architecture that integrates protein-protein interaction (PPI) network topology with patient-specific protein expression data to perform patient classification. The model consists of two main components: a Graph Neural Network (GNN) based on two layers of Graph Convolutional Networks (GCNs) or Graph Attention Network (GAT) and a fully connected (linear) layer that performs classification at the patient level.

2.2.1. *Protein Embedding Generation via Graph Convolutional Network (GCN)*

The initial phase of the PatientClassifierGNN focuses on generating context-rich protein embeddings. This is achieved using a two-layer Graph Convolutional Network (GCN), specifically employing the GCNConv operation.

Given an input feature matrix for proteins, $H^{(0)} \in \mathbb{R}^{N_{pr} \times N_p}$, where N_{pr} is the number of proteins and N_p is the number of patients (each protein's features are its expression levels across all patients), and a graph adjacency matrix A representing PPIs (with associated edge_weight values), the GCN layers perform a spectral graph convolution.

First GCN Layer: This layer transforms the initial protein features into an intermediate embedding space:

$$H^{(1)} = \sigma\left(\tilde{A}H^{(0)}W^{(1)}\right)$$

Here, $\tilde{A}$ denotes the symmetrically normalized adjacency matrix incorporating self-loops and edge weights, commonly defined as:

$$\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

where D is the degree matrix. $W^{(1)} \in \mathbb{R}^{N_p \times H_1}$ is a learnable weight matrix, and $\sigma(\cdot)$ represents the Rectified Linear Unit (ReLU) activation function. A dropout layer with a rate of 0.5 is subsequently applied to $H^{(1)}$ for regularization, resulting in $H_{\text{intermediate}} \in \mathbb{R}^{N_{pr} \times H_1}$.

Second GCN Layer: The intermediate protein embeddings are further processed to yield the final protein embeddings, mapped to a predefined protein embedding dimension:

$$H_{\text{proteins}} = \tilde{A} H_{\text{intermediate}} W^{(2)}$$

where $W^{(2)} \in \mathbb{R}^{H_1 \times D_{emb}}$ is the learnable weight matrix for this layer, and D_{emb} is the dimensionality of the final protein embedding. No activation function or dropout is applied after this layer, as $H_{\text{proteins}} \in \mathbb{R}^{N_{pr} \times D_{emb}}$ serves as the final dense representation for each protein.

2.2.2. *Patient Embedding Construction*

Following the generation of protein embeddings, a patient-specific embedding is constructed for each patient. This is achieved by taking a weighted sum of the learned protein embeddings, where the weights correspond to the original patient-specific protein expression levels.

Given the original patient expression data, $E_{\text{patient}} \in \mathbb{R}^{N_p \times N_{pr}}$, and the learned protein embeddings, $H_{\text{proteins}} \in \mathbb{R}^{N_{pr} \times D_{\text{emb}}}$, the final patient embeddings, $E_{\text{patients}} \in \mathbb{R}^{N_p \times D_{\text{emb}}}$, are computed as:

$$E_{\text{patients}} = E_{\text{patient}} \cdot H_{\text{proteins}}$$

This matrix multiplication effectively aggregates the network-aware protein features into a consolidated representation for each patient, reflecting their unique protein expression profile within the context of the protein-protein interaction (PPI) network.

2.2.3. *Patient Classification Sub-network*

After learning the final protein embeddings through the two-layer GNN, patient representations E_{patient} are computed by multiplying the original patient expression matrix with the protein embeddings. These aggregated patient embeddings are then passed through a single fully connected (Dense) layer to perform binary classification between "IBD" and "not IBD".

The classification layer outputs raw logits, which are suitable for further evaluation using cross-entropy loss during training. The architecture of this classification sub-network is defined as:

$$O = \text{Dense}_{\text{final}}(E_{\text{patients}})$$

Where:

- $E_{\text{patients}} \in \mathbb{R}^{N_p \times d}$ is the matrix of patient embeddings, with N_p patients and embedding dimension d.
- $\text{Dense}_{\text{final}}$ is a fully connected (linear) layer that outputs logits for each class.
- $O \in \mathbb{R}^{N_p \times 2}$ are the classification logits for each patient (for two classes).

2.3. Training and Evaluation Protocol

2.3.1. *Loss Function*

To address the inherent class imbalance between IBD and non-IBD patients, a **Weighted Cross-Entropy Loss** was employed as the training objective. This loss function introduces class-specific weights inversely proportional to class frequencies, ensuring that the minority class (IBD patients) has a proportionally larger influence during optimization.

Formally, for a two-class classification problem with class labels $y_i \in \{0, 1\}$ and predicted probabilities $\hat{y}_i \in [0, 1]^2$, the weighted cross-entropy loss is defined as:

$$\mathcal{L} = -\sum_{i=1}^{N} w_{y_i} \cdot \log(\hat{y}_{i,y_i})$$

where:

- N is the number of training samples,
- w_{y_i} is the weight assigned to class y_i,
- $\hat{y}_{i,y_i}$ is the predicted probability for the true class.

In this implementation, the class weights are computed as:

$$w_c = \frac{N}{C \cdot N_c}$$

where:

- $C = 2$ is the number of classes,
- N_c is the number of samples in class c,
- N is the total number of samples.

This weighting strategy helps mitigate the dominance of the majority class during optimization, improving the model's ability to detect the minority (IBD) class.

2.3.2. *Optimization*

The Adam optimizer was utilized for model parameter updates. The learning rate for the Adam optimizer was a key hyperparameter subject to tuning.

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

Where θ represents the model's parameters, η is the learning rate, $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates of the gradients, and ϵ is a small term for numerical stability.

2.4. Neural Network Architecture

We implemented a classical fully connected neural network (FCNN) using the Keras library. The architecture consists of:

- An input layer matching the dimensionality of the PPI-transformed proteomic data.
- Multiple dense hidden layers with varying widths (256, 512, or 1024 neurons), followed by layers with 128, 64, and 32 neurons respectively.
- Each dense layer uses the ReLU activation function.
- Dropout layers (rate = 0.5) were inserted after each hidden layer to mitigate overfitting.
- A final sigmoid-activated output layer for binary classification.

Hyperparameters such as learning rate, number of neurons, batch size, and number of training epochs were optimized through an exhaustive grid search, evaluating all combinations of:

- Learning rates: [0.0001, 0.0005, 0.001]

- Neurons in the first dense layer: [256, 512, 1024]
- Batch sizes: [32, 64]
- Epochs: [10, 20]

2.5. Imbalance Handling with SMOTETomek

To address the class imbalance inherent in the dataset, we applied the SMOTETomek technique on the training data. This hybrid method combines:

- SMOTE (Synthetic Minority Over-sampling Technique): Generates synthetic examples for the minority class (IBD).
- Tomek links removal: Cleans borderline or overlapping examples between classes to improve class separability.

This ensures a more balanced and cleaner training distribution, leading to improved generalization on the minority class.

2.6. Loss Function: Focal Loss

Instead of standard binary cross-entropy, we employed Focal Loss to further address class imbalance during training. Focal Loss reshapes the loss contribution so that misclassified and minority class samples receive more attention. Its formulation includes a tunable focusing parameter γ (gamma) and balancing factor α, with default values of 2.0 and 0.25, respectively.

$$\mathrm{FL}(p_t) = -\alpha_t \, (1 - p_t)^{\gamma} \log(p_t)$$

This is particularly useful in cases where the number of easy negatives (controls) vastly outnumbers harder positives (IBD patients).

3. Results

A comprehensive grid search was conducted to evaluate the impact of different hyperparameter configurations on the performance of a neural network model for classifying patients with Inflammatory Bowel Disease (IBD) using protein expression data and protein-protein interaction (PPI) integration. The evaluated parameters included learning rates (`0.0001`, `0.0005`, `0.001`), number of neurons in the first dense layer (`256`, `512`, `1024`), batch sizes (`32`, `64`), and training epochs (`10`, `20`, `40`).

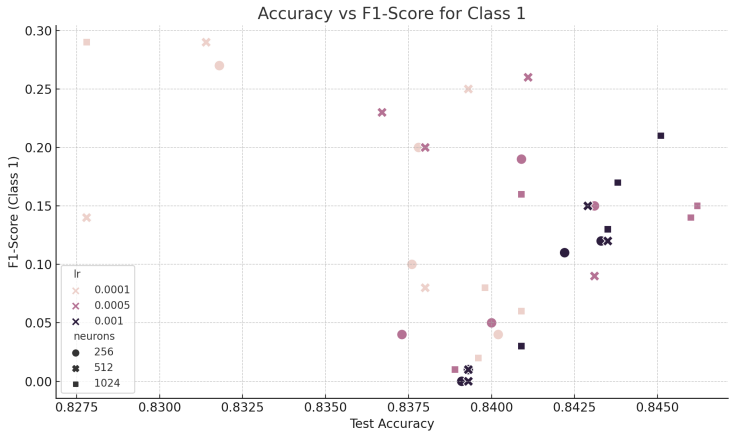

Figure 1. Accuracy vs F1-score for class 1 across different configurations for classical NN.

The best performing configuration used a learning rate of `0.0001`, `512` neurons, a batch size of `32`, and `40` training epochs. This model achieved a test accuracy of **83.14%**, with an F1-score for the disease class (class 1) of **0.29**. The model also achieved a precision of **0.38**

and a recall of **0.23** for class 1, highlighting the challenge in detecting the minority class despite overall high accuracy.

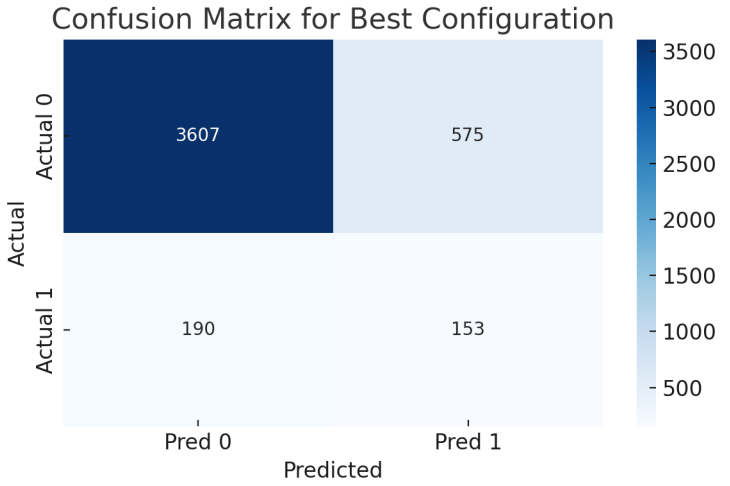

Figure 2. Confusion matrix for the best performing configuration for classical NN.

To address class imbalance, we applied the SMOTETomek method and employed focal loss with $\alpha = 0.25$ and $\gamma = 2.0$. Dropout layers were added to reduce overfitting.

The GCN model demonstrated a higher capacity to correctly identify IBD patients, achieving a recall of 0.51 for Class 1, indicating that over half of the true IBD cases were successfully detected. However, this came at the expense of precision (0.20), suggesting a high number of false positives. The overall F1-score for Class 1 was 0.29. In contrast, the control class (Class 0) achieved a precision of 0.86 and recall of 0.60. The model's overall accuracy was 0.58, with a macro-average F1-score of 0.50 and a weighted-average F1-score of 0.64, indicating a moderately balanced performance with a stronger bias toward the majority class.

On the other hand, the GAT model yielded a lower recall (0.43) and precision (0.14) for Class 1, resulting in a lower F1-score of 0.21 for the IBD class. The control class showed slightly weaker performance compared to the GCN, with a precision of 0.82 and a recall of 0.48. Overall accuracy for the GAT model was 0.47, with macro and weighted F1-scores of 0.40 and 0.54, respectively.

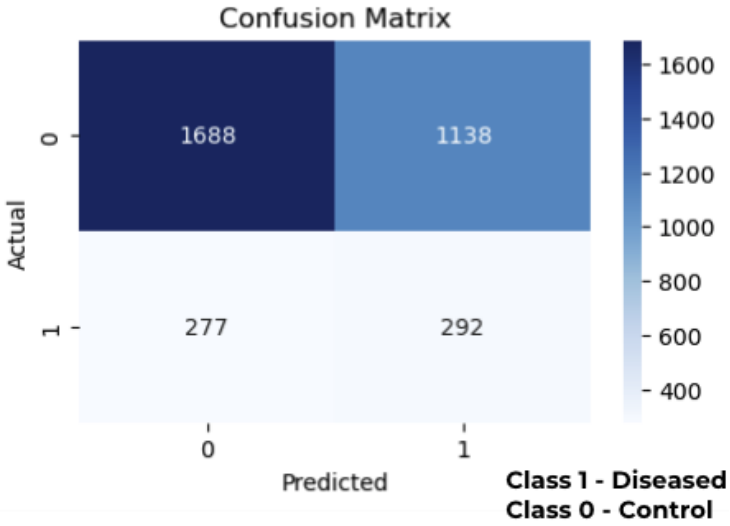

Figure 3. Confusion matrix for Graph NN using GCN.

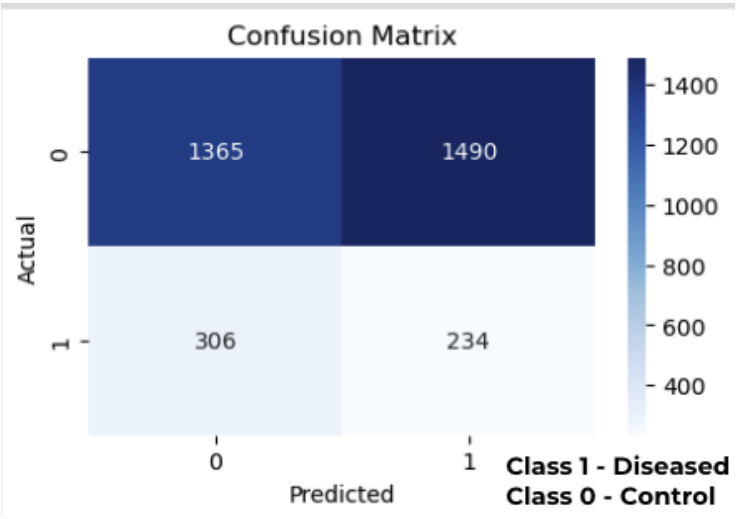

Figure 4. Confusion matrix for Graph NN using GAT.

4. Discussion

The results indicate that traditional deep neural networks can perform well in distinguishing between healthy individuals and IBD patients when enhanced with protein-protein interaction context. Incorporating focal loss helped mitigate the effects of class imbalance by reducing the impact of easy negative samples during training. The use of SMOTETomek further improved the representativeness of the minority class in the training dataset.

However, the relatively low recall and F1-score for the disease class suggest that the model still struggles to capture nuanced patterns associated with IBD. This could be due to the biological complexity of the disease, the presence of noise in high-dimensional proteomics data, or limitations in the interaction matrix used.

Future work could explore hybrid models, feature selection methods, or integrate clinical metadata to further enhance prediction performance.

5. Conclusions

In this model comparison, neither approach achieves strong performance in classifying the minority class, corresponding to sick patients, as reflected by a low and identical F1-score of 0.29. The traditional neural network exhibits higher precision (0.45), indicating fewer false positives when identifying positive cases. However, it suffers from a low recall (0.21), meaning it fails to detect a significant proportion of truly sick individuals—an important limitation in clinical settings where comprehensive detection is critical.

In contrast, the graph neural network (GCN) demonstrates superior ability to identify truly sick individuals, with a higher recall of 0.51, albeit at the cost of lower precision (0.20), resulting in more false positives. This trade-off may be acceptable or even desirable in medical contexts where missing a diagnosis poses a higher risk than over-diagnosis. Additionally, the GCN's higher AUC suggests a more reliable probability distribution for distinguishing between classes. Therefore, model selection should be guided by the clinical priorities of the task—whether minimizing false positives or maximizing the detection of true cases is more critical.

Youtube Video: https://youtu.be/O4Gf38GTIqM

References

1. Colombel, J.F.; Narula, N.; Peyrin-Biroulet, L. Management strategies to improve outcomes of patients with inflammatory bowel diseases. *Gastroenterology* **2017**, *152*, 351–361.e5. https://doi.org/10.1053/j.gastro.2016.09.046.
2. de Lange, K.M.; Moutsianas, L.; Lee, J.C.; et al.. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics* **2017**, *49*, 256–261.
3. Liu, J.Z.; van Sommeren, S.; Huang, H.; et al.. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics* **2015**, *47*, 979–986. International Multiple Sclerosis Genetics Consortium; International IBD Genetics Consortium.
4. Sun, B.B.; Maranville, J.C.; Peters, J.E.; et al.. Genomic atlas of the human plasma proteome. *Nature* **2018**, *558*, 73–79.
5. Zhernakova, D.V.; Le, T.H.; Kurilshikov, A.; et al.. Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nature Genetics* **2018**, *50*, 1524–1532. LifeLines cohort study; BIOS consortium.
6. Bourgonje, A.R.; von Martels, J.Z.H.; Gabriëls, R.Y.; et al.. A combined set of four serum inflammatory biomarkers reliably predicts endoscopic disease activity in inflammatory bowel disease. *Frontiers in Medicine (Lausanne)* **2019**, *6*, 251.