Enhancing patient stratification and interpretability through class-contrastive and feature attribution techniques

Sharday Olowu
University of Cambridge
Cambridge, UK
sylo2@cantab.ac.uk

Neil Lawrence
University of Cambridge
Cambridge, UK
ndl21@cam.ac.uk

Soumya Banerjee
University of Cambridge
Cambridge, UK
sb2333@cam.ac.uk

Abstract

A crucial component of treating genetic disorders is identifying the genes and gene modules that drive disease processes. While Next-Generation Sequencing (NGS) provides rich data for this task, current machine learning approaches often lack explainability and fail to account for gene correlations. We develop a comprehensive framework of machine learning techniques for explainable patient stratification in inflammatory bowel disease, focusing on Crohn's disease (CD) subtypes: CD with deep ulcer, CD without deep ulcer and IBD-controls. Our approach combines Gaussian Mixture Modelling for patient stratification, a modified kernelSHAP algorithm accounting for gene co-expression, systematic identification of gene modules, and class-contrastive analysis for explaining individual patient phenotypes. This framework confirms known disease-associated genes while unveiling novel genetic factors potentially underlying CD heterogeneity. Gene Ontology enrichment analysis validates the biological relevance of identified gene modules and associated pathways. Our methods offer a versatile toolkit for analysing high-dimensional, correlated biological data across diverse disease contexts.

1 Introduction

The wealth of RNA-Seq data from Next-Generation Sequencing has created new opportunities for analysing the genetic basis of disease, but current machine learning approaches often lack interpretability and overlook gene co-expression patterns. We introduce an explainable machine learning framework for uncovering the genetic aetiology of Crohn's disease (CD) subtypes, accounting for key gene correlation patterns. Our approach is summarised in Figure 1.

2 Data and Methods

2.1 Patient stratification

The study used a publicly available transcriptomic dataset [1, 2] containing RNA-Seq data from ileal tissue samples of paediatric subjects. The subjects were divided into subtypes: CD with deep ulcer, CD without deep ulcer and IBD-controls. Dimensionality reduction was performed using an autoencoder and PCA, followed by clustering with Gaussian Mixture Models (GMMs) and K-means; a post-processing algorithm was developed for subsequent classification into CD subtypes. Please see Supp. Material Section 4.1 for more essential details.

2.2 Clustering explainability

2.2.1 Modifying kernelSHAP to identify risk genes

Feature importance ranking provides crucial insights into machine learning model predictions. We developed an extension of kernelSHAP, a leading algorithm for this task, to identify genes most influential in predicting CD subtypes of patients from RNA-Seq data, which we demonstrate using Gaussian Mixture Models. While the original kernelSHAP [3] assumes feature independence, our method accounts for the correlated nature of gene expression by incorporating inter-feature dependence. Building upon work by Aas et al. [4], we approximate conditional feature distributions using a multivariate Gaussian distribution derived from training data, adjusting for each input instance to maintain gene relationships across coalition scenarios (see Supp. Material Section 4.2.1). This generates more realistic synthetic samples, yielding more accurate SHAP values and reliable explanations of disease subtypes at both patient and cluster levels. In practice, this can enable the systematic prioritisation of genes for further analysis, based on their predicted influence on CD subtype presentation.

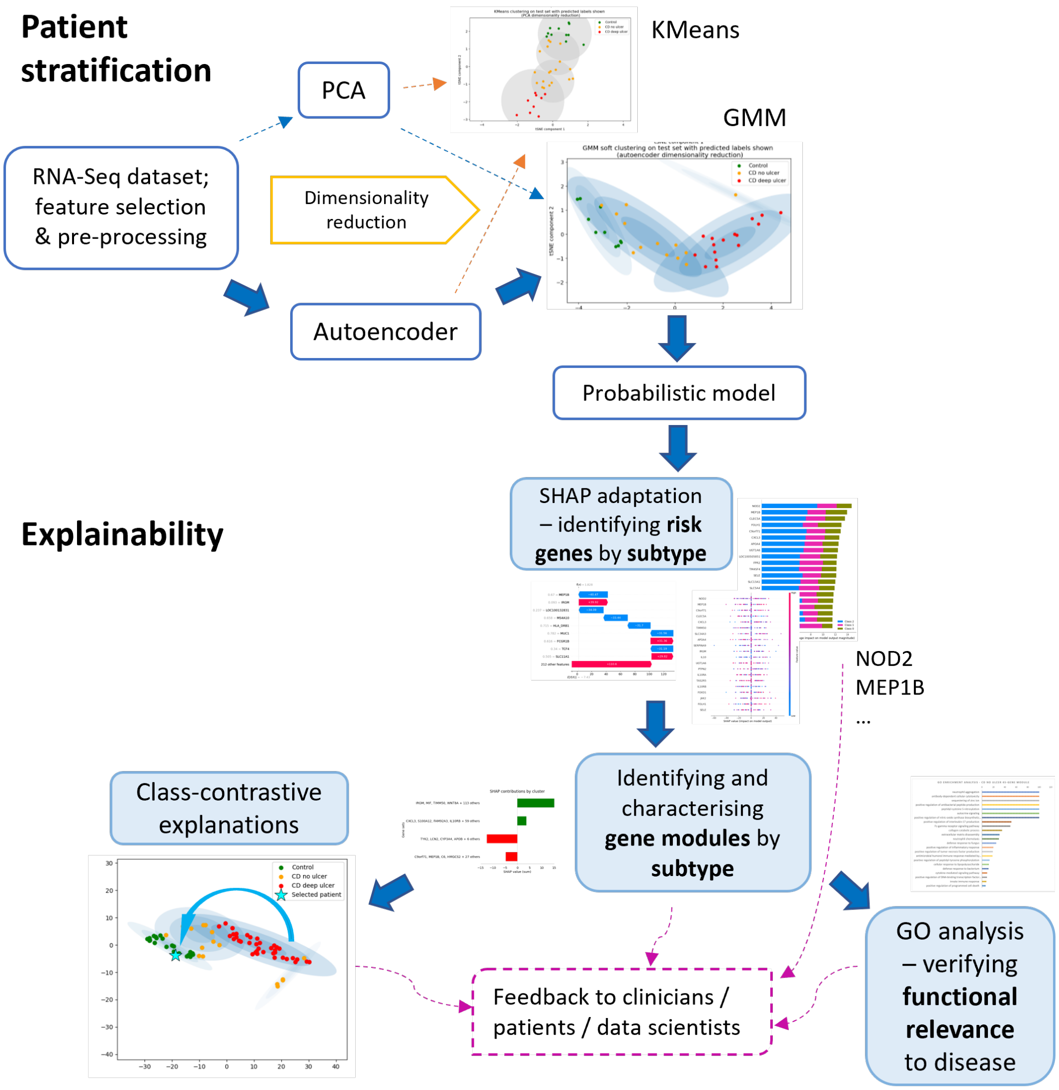

Patient stratification
PCA
KMeans
GMM
RNA-Seq dataset; feature selection & pre-processing
Dimensionality reduction
Autoencoder
Probabilistic model
Explainability
SHAP adaptation – identifying risk genes by subtype
NOD2
MEP1B
...
Class-contrastive explanations
Identifying and characterising gene modules by subtype
GO analysis – verifying functional relevance to disease
Feedback to clinicians / patients / data scientists
Control
CD no ulcer
CD deep ulcer
Selected patient

Figure 1: Overview of our computational framework. We identify genes and gene modules implicated in Crohn's disease (CD) subtypes. Starting with RNA-Seq data, reduce dimensionality using PCA and an autoencoder (adapted from [5]). We then use a probabilistic model (GMM) and post-processing to cluster and classify patients into disease subtypes (CD with deep ulcer, CD without deep ulcer and control). In order to explain our model we develop: 1. an extension of Shapley Additive Explanations (SHAP) to account for gene correlations, and 2. a class-contrastive technique to visually demonstrate the effect of changing gene expression on individual patients. Potential gene modules are identified using novel data integration and WECR clustering [6]. We confirm our findings by referencing peer-reviewed studies and conducting a Gene Ontology (GO) enrichment analysis.

2.2.2 Identification and characterisation of potential gene modules

We introduce a novel approach to identify disease-related gene modules by integrating the resulting SHAP values (Section 2.2.1) with the gene expression (RNA-Seq) data. The main idea is to use both gene activity level *and* SHAP gene importance to determine those genes which are likely to be working in concert as "modules" wrt. disease processes. The integration formula is as follows: $v_{pg} = x_{pg}\frac{\sum_i abs(s_{ig})c_i}{n}$ where v_{pg} represents the integrated value for patient p and gene g, x_{pg} is the gene expression, s_{ig} is the SHAP value, c_i indicates membership in a disease subtype, and n is the total number of patients in that subtype. This formula uniquely combines gene importance (via SHAP values) with expression levels, providing a more comprehensive view of gene influence on disease subtypes. The integrated data undergoes Weighted Ensemble Consensus of Random (WECR) K-Means clustering [6], with the optimal number of clusters determined using four validation metrics: Bayesian Information Criterion, Davies-Bouldin Index, Silhouette Score, and Calinski-Harabasz Index. We also demonstrate a technique for characterising the resulting gene modules by a sum of SHAP values across each gene module for a given disease subtype, as shown in Figure 3, to indicate the type and magnitude of influence of each identified module. The results are verified through Gene Ontology enrichment analysis, offering insights into the functional relevance of each gene module wrt. specific biological pathways, and potential impacts on disease subtype. Further essential details of the approach can be found in Supp. Material Section 4.2.2.

2.2.3 Class-contrastive technique for patient-specific explainability

A novel class-contrastive technique was developed to explain clusters specific to a patient by generating explanations that provide a contrast to another class.

For patients with Crohn's Disease (CD), class-contrastive explanations can be generated by artificially modifying the expression of genes in a given module to more closely resemble those of control subjects. We achieved this by assigning a new expression value (v) for each chosen gene, calculated as the mean value for the expression of this gene across all control individuals, as shown in the equation: $\forall g \in G, \quad v_{pg} = \frac{1}{N}\sum_i c_i x_{ig}$, where x represents an expression value, p is the selected patient, g is the selected gene, and G is the full set of genes in the module. The summation is performed over all control individuals i, where c_i serves as the indicator variable for the control group and N represents the total number of control individuals. After modifying these expression values, the patient's data is rerun through the clustering model to observe whether their cluster assignment changes - for instance, whether a patient originally clustered in the 'severe disease' group might shift to a 'mild disease' or even 'healthy control' cluster, thereby revealing the importance of those modified genes in disease classification. More comprehensive details can be found in Supp. Material Section 4.2.3.

3 Results and Discussion

3.1 Gaussian Mixture Model (GMM) and KMeans clustering

GMM clustering with autoencoder and tSNE outperformed other methods in stratifying and classifying patients into Chrohn's disease subtypes, as shown in Table 1. This approach was selected for downstream analysis. More details are in Supp. Material Section 5.1.

Table 1: Clustering and classification evaluation results for novel classifiers based on Gaussian Mixture Model (GMM) and KMeans models, using autoencoder and PCA dimensionality reduction methods. Results shown for binary classification (controls and all CD patients) and multi-class classification (control, CD no ulcer and CD deep ulcer) of disease subtype.

		Binary (control & CD)		**Multi-class (all labels)**	
		Autoencoder	PCA	Autoencoder	PCA
GMM	Accuracy / %	94.9	92.3	71.8	64.1
	F1-Score / %	96.7	94.9	71.5	62.6
	Silh. score	0.382	0.410	0.320	0.317
KMeans	Accuracy / %	84.6	82.1	64.1	59.0
	F1-Score / %	89.3	88.1	61.9	58.3
	Silh. score	0.556	0.409	0.469	0.334

3.2 Cluster explanation using kernelSHAP adapted for feature dependence

We coupled our GMMs to kernelSHAP [3] to generate explainability for each cluster, including visualisations [7]. Our modification of the kernelSHAP method incorporates feature dependence to more accurately model gene correlations and regulatory relationships, enabling identification of the most influential genes in predicting disease subtypes. Figure 2 shows a summary of the top 20 most influential genes, with bars depicting their influence on "CD deep ulcer" (blue), "CD no ulcer" (pink) and "control" (green) classifications.

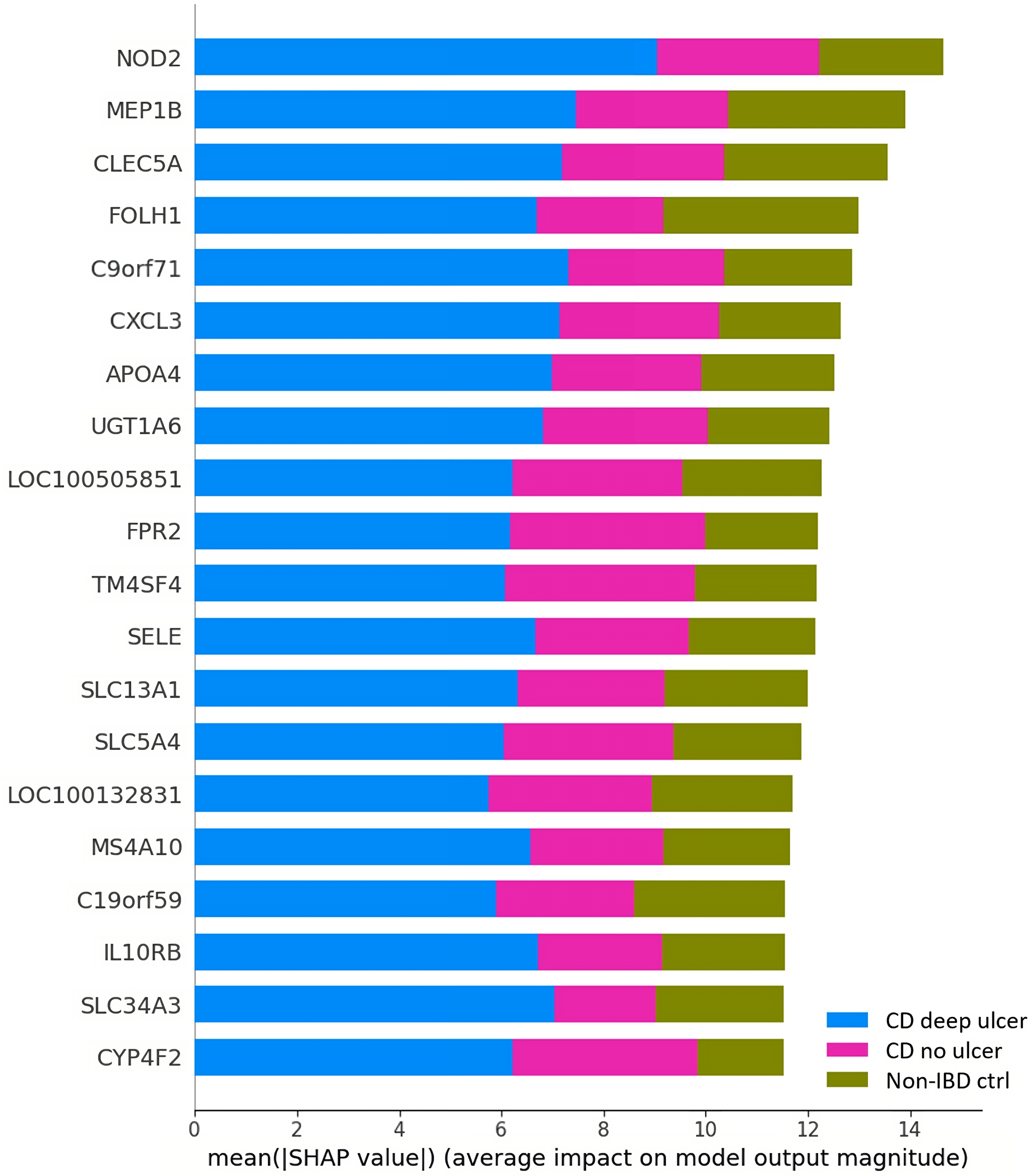


Figure 2: Summary plot showing top 20 genes in terms of their average impact on class predictions across all patients. This is for a model which accounts for feature dependence. The blue, pink and green bars depict the magnitude of influence of a gene on the “CD deep ulcer”, “CD no ulcer” and “control” classes respectively. The greatest proportion of influence of genes is attributed to the “CD deep ulcer” cluster. Genes like NOD2, MEP1B and FOLH1 are within the top 5 and known as susceptibility genes for IBD. The most significant gene identified overall is NOD2, which is known to be strongly associated with IBD.

Compared to the original kernelSHAP method (Supp. Material Fig. 6), where 4 of the top 5 genes (BPIFB1, FOXD1, C19orf59 and NAT8B) had no known IBD associations, our SHAP adaptation ranked established IBD genes significantly higher, with NOD2, MEP1B, and FOLH1 appearing in the top 5. These are examples of known susceptibility genes for IBD. For example, NOD2 plays a crucial role in bacterial sensing by recognising muramyl dipeptide on bacterial cell envelopes, leading to oligomerisation and RICK kinase binding. This activates the NF-κB pathway [8, 9], resulting in pro-inflammatory cytokine accumulation and tissue inflammation [10]. Our method also identified additional IBD-implicated genes not detected by the original algorithm, including IL10RB, CXCL3, APOA4, SLC13A1 and SLC5A4. Chemokines such as CXCL3 and cytokines such as IL10 are associated with inflammatory processes in IBD [11]. For example, Interleukin-10 maintains intestinal haemostasis by suppressing pro-inflammatory cytokines like TNF and IL-12, with its disruption leading to IBD symptoms [12]. Interestingly, our analysis also highlighted two uncharacterised "LOC" genes [13] that may represent novel therapeutic targets, although further validation is needed, as some identified genes such as SELE currently show limited evidence of IBD association.

Compared to state-of-the-art approaches that commonly apply SHAP to neural networks [14, 15, 16, 17], our GMM-based probabilistic model coupled with the kernelSHAP adaptation provides improved interpretability while accounting for gene dependencies. The method generates both local explanations for individual patients and global insights across the dataset, with the incorporation of inter-feature dependence resulting in improved alignment with established IBD literature. Please see Supp. Material Section 5.2 for more details.

3.3 Identification and characterisation of potential gene modules

To identify disease-related gene modules for each CD subtype, our novel approach integrates SHAP values with gene expression data, before applying Weighted Ensemble Consensus of Random K-Means clustering [6]. In analysing the CD deep ulcer subtype, for example, it reveals four modules containing known IBD genes, like IRGM [18], CXCL3 [11], and IL10RB [12]. Each module's type and magnitude of influence on disease predictions is visualised in Figure 3.

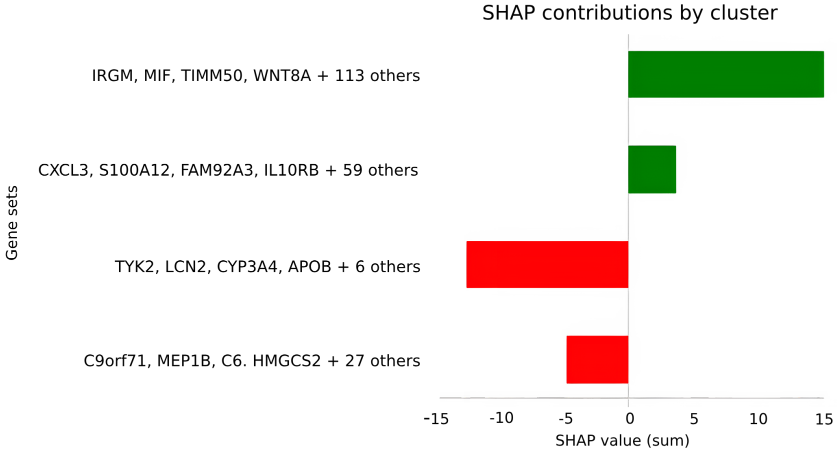


Figure 3: Final gene modules identified as being associated with severe disease (CD deep ulcer), alongside relative contributions determined using SHAP values. Shown is a bar plot in which each gene module is characterised in terms of its influence on the model predicting the "CD deep ulcer" cluster. Each bar represents the sum of mean SHAP values associated with the "CD deep ulcer" cluster, across all genes in the module. We show the top 4 most influential genes of each module. More positive values (green) indicate greater confidence for that module predicting "CD deep ulcer". More negative values (red) reduce our confidence in a "CD deep ulcer" prediction. Consistent with the literature on IBD, we find genes like IRGM, CXCL3 and IL10RB are present in influential modules and have a significant effect on disease severity.

3.3.1 Gene Ontology (GO) enrichment analysis

We verified the biological relevance of identified gene modules through Gene Ontology enrichment analysis [19, 20, 21]. The CD deep ulcer 117-gene module (Supp. Material Fig. 8) showed enrichment in IBD-relevant processes including transport mechanisms, reactive oxygen species regulation, and immune responses, with statistically significant results (FDR < 0.05) and high fold enrichment (>100) for many processes. This aligns with IBD's characteristic dysregulated immune response to pathogens [22], showing inflammatory processes regulated through various cytokines and signaling pathways [12]. Notably, this was the only module showing enriched regulation of reactive oxygen species, specifically associated with CD deep ulcers [1]. The 63-gene module (Supp. Material Fig. 9) revealed additional pathways involving TLR6 and TLR2 recognition of bacterial patterns [23], along with processes related to embryonic digestive tract development and extracellular matrix organisation [24, 25]. The 45-gene module (Supp. Material Fig. 10), associated with CD no ulcer, showed broader immune responses including fungal response and fc-gamma receptor signaling [26]. Our findings align with current literature [27, 28, 29, 30] while uncovering novel pathways, particularly in embryonic development, potentially advancing our understanding of IBD mechanisms and CD subtype differentiation. Though smaller modules sometimes lacked enriched processes due to insufficient gene counts, the overall results strongly correspond with established IBD literature.

3.4 Class-contrastive explainability

We developed a class-contrastive method to visually demonstrate the impact of identified gene modules on disease subtypes. We had initially observed that gene expression across disease subtypes follows Gaussian distributions, with some genes showing clear up- or down-regulation in CD patients compared to controls, as illustrated by CXCL3 and MEP1B (Supp. Material Fig. 11). Exploring the effects of expression strength led to the development of our technique. To demonstrate our method, we analysed Patient 46 (CD deep ulcer; Fig. 4) by modifying their gene expression values to match control group means. When adjusting the 117-gene module, which showed the strongest positive association with CD deep ulcer (Figure 3), the patient's classification shifted from CD deep ulcer to CD no ulcer (Figure 5a). Furthermore, modifying both the 117-gene and 63-gene modules (180 genes total) resulted in reclassification to the control cluster (Figure 5b), suggesting these modules' significant role in severe disease manifestation. Similar effects were observed when analysing numerous other patients across the dataset. This technique provides intuitive patient-specific explanations. While serving as a proof-of-concept with potential for improvement through larger datasets, the method effectively demonstrates how specific gene modules influence disease classification and severity.

3.5 Validation with an independent cohort

To validate the generalisability of our approach, we applied our methods to an independent cohort of 210 pediatric Crohn's disease patients and 35 non-IBD controls, featuring different disease subtypes ("CD no complication" and "CD with complication"). The initial stages involved pre-processing and filtering, from which we identified 278 differentially expressed genes and applied our computational framework using a newly trained autoencoder adapted for this dataset. Results (Supp. Material Table 3; Figure 14) showed strong performance comparable to our original study. Binary classification (control vs. CD) achieved 89.2% accuracy and 93.3% F1-score using GMM with autoencoder and tSNE. Multi-class classification showed slightly lower but respectable scores (64.9% accuracy, 69.9% F1-score), reflecting the real-world challenge of distinguishing between related CD subtypes. Subsequent kernelSHAP analysis identified the most influential genes, with CXCL5, IL1B, and CXCL6 among the top three, as shown in Supp. Material Figure 15. These genes, known for their role in intestinal inflammation, showed strong influence on model predictions [31]. Several other identified genes have established links to IBD, such as FMO5's role in mucus barrier formation [32] and ALDH1A2's involvement in vitamin A signaling disruption during active IBD [33]. We also identified potentially novel targets, including WNT5A, which has recently been recognised as a potential therapeutic target. These results demonstrate our framework's ability to generalise across datasets and effectively identify both known and novel genes associated with CD subtypes.

4 Conclusions and discussion

Compared to the state-of-the-art, one of our key contributions is in adapting kernelSHAP to account for gene correlations, enabling more accurate identification of risk genes in CD subtypes through our GMM-derived probabilistic model. This framework not only stratifies patients effectively but also captures complex relationships between expression profiles and disease subtypes.

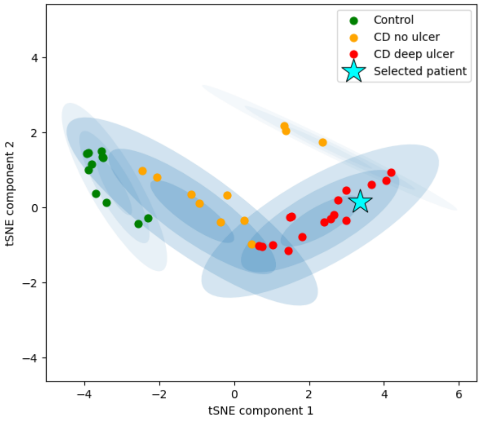
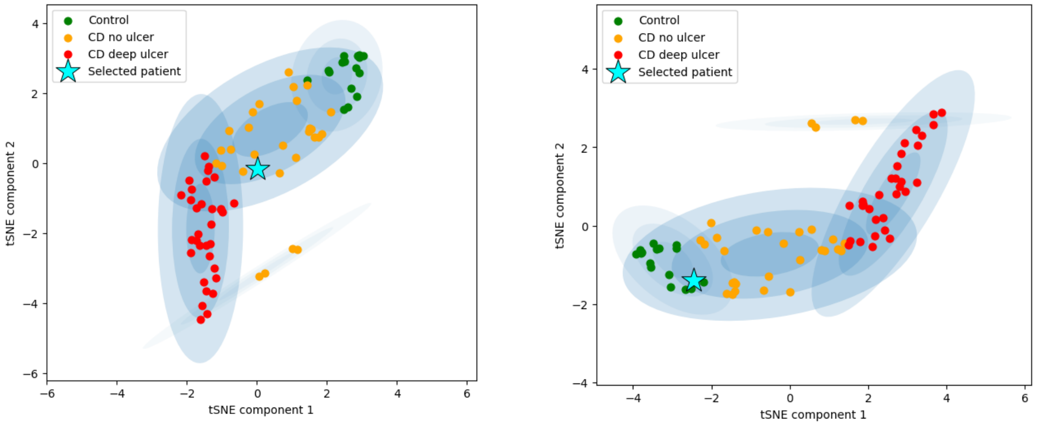


Figure 4: Initial position of Patient 46 (CD deep ulcer) within the clustering model.



(a) 117-gene module alteration

(b) 117-gene and 63-gene module alteration

Figure 5: Visual explanations of the effect of modules on disease in a patient. In comparison to Figure 4, we show the position of Patient 46 within clustering model after modifying the 117-gene module (a) and both 117-gene and 63-gene modules (b), which together were found to account for all positive contribution to CD deep ulcer predictions (Figure 3). In (a) modifying the 117 genes using the class-contrastive technique results in Patient 46 (with CD deep ulcer [a severe form of the disease], Figure 4) being assigned to the "CD no ulcer" cluster [a less severe form of the disease]. In (b) modifying both the 117-gene and 63-gene modules using the class-contrastive technique results in Patient 46 moving from the CD deep ulcer cluster to the control cluster. This suggests that these modules may be involved in a severe form of CD that leads to deep ulcers.

We identify both established IBD genes (NOD2, IRGM, IL10) and potentially novel targets (including uncharacterised LOC genes). While some identified genes, such as SELE, lack strong established links to IBD, these could represent either novel findings or limitations of our approach. Through a novel data integration technique and WECR consensus clustering [6], we identify disease-relevant gene modules, with GO enrichment analysis validating their biological significance. Our class-contrastive technique also provides intuitive visual explanations of how gene modules influence disease classification at the individual patient level.

While our approach is not intended for de-novo risk gene identification, it effectively identifies disease-relevant genes and gene modules from expression data, and characterises their influence. We validate our techniques with an independent cohort in Supp. Material Section 5.5. The model-agnostic nature of our methods also makes them applicable to other domains where explainable analysis of correlated features is crucial (e.g. fraud detection, climate modelling etc.). This framework has significant potential to enhance clinical decision-making by providing interpretable insights into disease mechanisms at both population and individual levels.

References

[1] Yael Haberman, Timothy L. Tickle, Phillip J. Dexheimer, Mi Ok Kim, Dora Tang, Rebekah Karns, Robert N. Baldassano, Joshua D. Noe, Joel Rosh, James Markowitz, Melvin B. Heyman, Anne M. Griffiths, Wallace V. Crandall, David R. Mack, Susan S. Baker, Curtis Huttenhower, David J. Keljo, Jeffrey S. Hyams, Subra Kugathasan, Thomas D. Walters, Bruce Aronow, Ramnik J. Xavier, Dirk Gevers, and Lee A. Denson. Erratum: Pediatric crohn disease patients exhibit specific ileal transcriptome and microbiome signature (journal of clinical investigation (2014) 124: 8 (3617-3633) doi: 10.1172/jci75436). *Journal of Clinical Investigation*, 125, 2015.

[2] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(Database issue):D991–995, January 2013.

[3] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[4] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, 2021.

[5] Satyam Kumar. Improve your Model Performance with Auto-Encoders, December 2021.

[6] Yongxuan Lai, Songyao He, Zhijie Lin, Fan Yang, Qifeng Zhou, and Xiaofang Zhou. An adaptive robust semi-supervised clustering framework using weighted consensus of random k-means ensemble. *IEEE Transactions on Knowledge and Data Engineering*, 33(5):1877–1890, 2021.

[7] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018.

[8] Soichiro Yamamoto and Xiaojing Ma. Role of Nod2 in the development of Crohn's disease. *Microbes and infection / Institut Pasteur*, 11(12):912–918, October 2009.

[9] I. Atreya, R. Atreya, and M. F. Neurath. NF-κB in inflammatory bowel disease. *Journal of Internal Medicine*, 263(6):591–596, 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2796.2008.01953.x.

[10] B. M. Fournier and C. A. Parkos. The role of neutrophils during intestinal inflammation. *Mucosal Immunology*, 5(4):354–366, July 2012. Number: 4 Publisher: Nature Publishing Group.

[11] J. Puleston, M. Cooper, S. Murch, K. Bid, S. Makh, P. Ashwood, A. H. Bingham, H. Green, P. Moss, A. Dhillon, R. Morris, S. Strobel, R. Gelinas, R. E. Pounder, and A. Platt. A distinct subset of chemokines dominates the mucosal chemokine response in inflammatory bowel disease. *Alimentary Pharmacology & Therapeutics*, 21(2):109–120, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2036.2004.02262.x.

[12] Reza Yazdani, Bobak Moazzami, Seyedeh Panid Madani, Nasrin Behniafard, Gholamreza Azizi, Majid Aflatoonian, Hassan Abolhassani, and Asghar Aghamohammadi. Candidiasis associated with very early onset inflammatory bowel disease: First IL10RB deficient case from the National Iranian Registry and review of the literature. *Clinical Immunology*, 205:35–42, August 2019.

[13] LOC100505851 uncharacterized LOC100505851 [Homo sapiens (human)] - Gene - NCBI.

[14] Melvyn Yap, Rebecca L. Johnston, Helena Foley, Samual MacDonald, Olga Kondrashova, Khoa A. Tran, Katia Nones, Lambros T. Koufariotis, Cameron Bean, John V. Pearson, Maciej Trzaskowski, and Nicola Waddell. Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Scientific Reports*, 11(1):2641, January 2021.

[15] Jin Hayakawa, Tomohisa Seki, Yoshimasa Kawazoe, and Kazuhiko Ohe. Pathway importance by graph convolutional network and shapley additive explanations in gene expression phenotype of diffuse large b-cell lymphoma. *PLOS ONE*, 17:e0269570, 6 2022.

[16] Yang Yu, Pathum Kossinna, Wenyuan Liao, and Qingrun Zhang. Explainable autoencoder-based representation learning for gene expression data. 12 2021.

[17] M. Pavageau, L. Rebaud, D. Morel, S. Christodoulidis, E. Deutsch, C. Massard, H. Vanacker, and L. Verlingue. DeepOS: pan-cancer prognosis estimation from RNA-sequencing data. preprint, Oncology, July 2021.

[18] Subhash Mehto, Kautilya Kumar Jena, Parej Nath, Swati Chauhan, Srinivasa Prasad Kolapalli, Saroj Kumar Das, Pradyumna Kumar Sahoo, Ashish Jain, Gregory A. Taylor, and Santosh Chauhan. The Crohn's Disease Risk Factor IRGM Limits NLRP3 Inflammasome Activation by Impeding Its Assembly and by Mediating Its Selective Autophagy. *Molecular Cell*, 73(3):429–445.e7, February 2019.

[19] Seth Carbon and Chris Mungall. Gene Ontology Data Archive, March 2023.

[20] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, May 2000.

[21] Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, January 2021.

[22] Bani Ahluwalia, Luiza Moraes, Maria K. Magnusson, and Lena Öhman. Immunopathogenesis of inflammatory bowel disease and mechanisms of biological therapies. *Scandinavian Journal of Gastroenterology*, 53(4):379–389, April 2018. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00365521.2018.1447597.

[23] Takumi Kawasaki and Taro Kawai. Toll-Like Receptor Signaling Pathways. *Frontiers in Immunology*, 5, 2014.

[24] Alicja Derkacz, Paweł Olczyk, Krystyna Olczyk, and Katarzyna Komosinska-Vassev. The Role of Extracellular Matrix Components in Inflammatory Bowel Diseases. *Journal of Clinical Medicine*, 10(5):1122, March 2021.

[25] Cristiano Pagnini and Fabio Cominelli. Tumor Necrosis Factor's Pathway in Crohn's Disease: Potential for Intervention. *International Journal of Molecular Sciences*, 22(19):10273, September 2021.

[26] Fabian Junker, John Gordon, and Omar Qureshi. Fc Gamma Receptors and Their Role in Antigen Uptake, Presentation, and T Cell Activation. *Frontiers in Immunology*, 11, 2020.

[27] Weitao Hu, Taiyong Fang, and Xiaoqing Chen. Identification of differentially expressed genes and mirnas for ulcerative colitis using bioinformatics analysis. *Frontiers in Genetics*, 13, 2022.

[28] Mohammad Elahimanesh and Mohammad Najafi. Cross talk between bacterial and human gene networks enriched using ncRNAs in IBD disease. *Scientific Reports*, 13(1):7704, May 2023. Number: 1 Publisher: Nature Publishing Group.

[29] Chunwei Cheng, Juan Hua, Jun Tan, Wei Qian, Lei Zhang, and Xiaohua Hou. Identification of differentially expressed genes, associated functional terms pathways, and candidate diagnostic biomarkers in inflammatory bowel diseases by bioinformatics analysis. *Experimental and Therapeutic Medicine*, 18(1):278–288, July 2019.

[30] Xiaoli Pang, Hongxiao Song, Xiaolu Li, Fengchao Xu, Bingxun Lei, Fei Wang, Jing Xu, Lingli Qi, Libo Wang, and Guangyun Tan. Transcriptomic analyses of treatment-naïve pediatric ulcerative colitis patients and exploration of underlying disease pathogenesis. *Journal of Translational Medicine*, 21(1):30, January 2023.

[31] Łukasz Kopiasz, Katarzyna Dziendzikowska, and Joanna Gromadzka-Ostrowska. Colon expression of chemokines and their receptors depending on the stage of colitis and oat beta-glucan

dietary intervention—crohn's disease model study. *International Journal of Molecular Sciences*, 23(3):1406, January 2022.

[32] Megan L. Schaller, Madeline L. Sykes, Joy Mecano, Sumeet Solanki, Wesley Huang, Ryan J. Rebernick, Safa Beydoun, Emily Wang, Amara Bugarin-Lapuz, Yatrik M. Shah, and Scott F. Leiser. Fmo5 plays a sex-specific role in goblet cell maturation and mucus barrier formation. April 2024.

[33] Siri Sæterstad, Ann Elisabet Østvik, Elin Synnøve Røyset, Ingunn Bakke, Arne Kristian Sandvik, and Atle van Beelen Granlund. Profound gene expression changes in the epithelial monolayer of active ulcerative colitis and crohn's disease. *PLOS ONE*, 17(3):e0265189, March 2022.

Supplementary Material: Enhancing patient stratification and interpretability through class-contrastive and feature attribution techniques

Sharday Olowu
University of Cambridge
Cambridge, UK
sylo2@cantab.ac.uk

Neil Lawrence
University of Cambridge
Cambridge, UK
ndl21@cam.ac.uk

Soumya Banerjee
University of Cambridge
Cambridge, UK
sb2333@cam.ac.uk

1 Introduction

In recent years, vast amounts of genomic and transcriptomic data have become publicly available, driven by the development of next-generation sequencing technologies such as RNA-Seq [1]. These allow us to study patterns of gene expression within tissue samples, which can be used to analyse the genetic component of various diseases.

Genetic diseases can be characterised by the activity of certain genes, which often work in concert as modules, driving specific biological processes. The objective of this work is to develop explainable techniques for the identification of genes and gene modules associated with disease. Advanced machine learning techniques have greater ability to capture nuanced relationships between genes and, despite high dimensionality and noise, enable clustering or classification of disease subtype. However, many of these machine learning models are difficult to explain.

An active area of research is in developing methods to explain machine learning model predictions: Explainable AI. This is increasingly important in sensitive domains like healthcare, recruitment and adjudication. For example, in clinical settings, any predictions drawn must be grounded in sound reasoning, given the risks involved.

In this work, we develop a technique for explainability, to identify genes and gene modules associated with disease. We focus on improving the explainability of patient stratification from transcriptomic data. We develop explainability techniques to account for key characteristics of gene expression such as gene distributions and correlations, which current machine learning approaches often overlook.

Inflammatory Bowel Disease (IBD) can be categorised into Crohn's Disease (CD) and Ulcerative Colitis. IBD is a chronic digestive disorder characterised by inflammation of the gastrointestinal tract, causing symptoms such as abdominal pain, diarrhoea, and weight loss. There is known to be a strong genetic component, but the risk factors are not fully understood [2].

We use our methods to analyse bulk RNA-Seq data and explore the genetic basis of CD subtypes. We use our techniques to analyse transcriptomic profiles and identify genes associated with subtypes of CD. Patient stratification can enable targeted treatment. Identifying therapeutic gene targets may inform the development of more effective treatments.

1.1 Overview of approach

Our contributions are summarised below:

- We predict Crohn's disease (CD) subtype based on gene expression data. We use a machine learning model called a Gaussian Mixture Model which performs soft clustering. The disease subtypes predicted are Crohn's disease with deep ulcer (a severe form of the disease), Crohn's disease with no ulcer, and IBD-control.

- We adapt an explainable AI technique called kernelSHAP to account for gene correlations. We couple kernelSHAP to a probabilistic model we derive from the Gaussian Mixture Model, to quantify the influence of genes for each Crohn's disease subtype.
- We then use consensus clustering [3] to identify potential gene modules by Crohn's disease subtype.
- For these gene modules, we determine the type and relative magnitude of influence on disease subtype. These are verified using Gene Ontology enrichment analysis.
- Finally, we develop a class-contrastive technique to visually explain the impact of gene modules on the disease subtype for an individual patient.

We identify known IBD risk genes such as NOD2, IRGM, JAK2 and IL10. Furthermore, our analysis suggests a role for uncharacterised genes such as LOC100505851 and LOC100132831. We show using Gene Ontology enrichment analysis that the identified gene modules are relevant to IBD. We visually demonstrate the impact of these genes on each patient using a class-contrastive technique.

Our techniques may aid in the diagnosis and treatment of genetic diseases. Our approach may also be broadly applicable in other domains where explainability and feature correlations are important.

2 Background

We highlight key theoretical background to our work in this section.

2.1 Gaussian Mixture Models (GMMs)

Gaussian Mixture Models (GMMs) can be used to perform soft clustering of data. They help organise data into groups, which is useful for finding disease subtypes based on our transcriptomic dataset.

GMMs work by using a mix of bell-shaped curves (Gaussian distributions) to represent the groups. These groups are called mixture components and have associated parameters such as mean μ_k and covariance Σ_k. The mix of these groups and how important they are in the model is determined by mixing coefficients (π_k). These coefficients show how much each group contributes to understanding the overall data.

To make these groups fit the data best, we adjust them iteratively using a process called maximum-likelihood estimation. This process fine-tunes the groups' characteristics to match the observed data distribution.

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{1}$$

This equation helps us evaluate how well our groups fit our observed data. We aim to adjust our groups to maximise this measurement, ensuring our model accurately represents the data.

2.2 SHAPley Additive exPlanations (SHAP)

SHAPley Additive exPlanations (SHAP) [4] is a method to explain the predictions of machine learning models. It aims to find the contribution of each feature to the model output by using a game-theoretic approach.

KernelSHAP provides an efficient approximation to SHAP values using weighted linear regression based on sampling. Rather than retraining a new model for each coalition as with classic SHAP, we marginalise the missing features out of the model. In Eq 2, we define a fidelity function L that measures how unfaithful is a surrogate model g in approximating the model f, in the feature subspace defined by z', across all models. Here, we use $z' \in \{0, 1\}^M$ to define the coalition of features, where M is the number of input features.

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} [f(h_x(z')) - g(z')]^2 \pi_{x'}(z') \tag{2}$$

We generate synthetic samples for each model, where each baseline sample z is drawn from the same probability distribution as the input features. We can compute the model output $f(h_x(z'))$ as

$E[f(z)|z_S] = E_{z_{\bar{S}}|z_S}[f(z)]$. However, kernelSHAP assumes feature independence so $f(h_x(z')) \approx E_{z_{\bar{S}}}[f(z)] \approx f([z_S, E[z_{\bar{S}}]])$. This means we simulate the subset of missing features $\bar{S}$ using expectation values, to show that these features carry no information. We use z' to represent a perturbed version of the sample z, where the included features S take their value from the input instance we are analysing. The h_x function is used to map the samples to a potentially higher-dimensional space. A kernel weighting function is also used to emphasise the independent and global effects of features.

We then perform linear regression to minimise the fidelity function L, which gives rise to Eq 3. ϕ_0 is the expected model output when no features are present and the remaining coefficients ϕ_{1-M} are the SHAP values of the corresponding features.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{3}$$

We aim to find the SHAP value of each feature, as this represents the type and extent of influence that a feature has on the model prediction. This explainability technique can be applied to estimate the influence of particular genes on patient outcomes.

3 Related work

3.1 Explainability

3.1.1 Class-contrastive techniques

Class-contrastive techniques can be used to find the impact of particular features on the model output. This can be very useful for improving the transparency of a model. For example, Banerjee et al. generate explainability for mortality predictions of patients in [5], as predicted by deep learning models. By setting the presence or absence of binary features such as “suffering with depression” or “lack of family support”, they were able to find the isolated effect of each feature on the risk of patient mortality (as predicted by a black-box model). This technique is usually used with categorical features. The class-contrastive approach has been used for self-supervised clustering of RNA-Seq data [6], but not for generating explanations of disease subtype based on genomic data. In this work, we extend the class-contrastive approach to demonstrate the impact of gene modules on disease subtype, taking into account the underlying gene expression distributions.

3.1.2 Applications of SHAP

SHapley Additive exPlanations (SHAP) [4] is another state-of-the-art explainability technique. Fast approximations of SHAP have been applied to analyse gene expression data [7, 8, 9, 10, 11], such as kernelExplainer, treeExplainer and gradientExplainer [12]. Yu et al. use a deep autoencoder [9] to learn gene expression representations, applying treeExplainer SHAP to measure the contributions of genes to each of the latent variables.

Although SHAP has shown success in this line of work, one significant problem is that it assumes feature independence. This means that when applying SHAP to find the contribution of genes, it is assumed that there are no correlations between genes. This is unrealistic because genes are often correlated and/or regulated by other genes; this is governed by complex gene regulatory networks [13].

There have been some attempts to include feature dependence, for example in the linearExplainer and treeExplainer [14] SHAP variants. However, linear models are not suitable for modelling complex gene-gene and genotype-phenotype relationships. The treeExplainer is also limited to tree-based ensemble methods, which can be difficult to visualise and interpret. We aim to address this by incorporating inter-feature dependence for kernelSHAP (kernelExplainer), which is model-agnostic and therefore applicable in many more contexts. By incorporating inter-feature dependence, we can more accurately identify potential genes, gene modules, and associated biological pathways implicated in disease processes.

Aas et al. propose to achieve inter-feature dependence for kernelSHAP in [15]. Here, a multivariate Gaussian distribution is constructed using a sample mean vector and covariance matrix, calculated from the training data. For each input instance (corresponding to a coalition of features), this model is updated under a Bayesian framework and used to generate synthetic samples for the calculation of conditional expectations required by the kernelSHAP algorithm. In the context of RNA-Seq analysis, this method is not appropriate because the model of relationships between genes can differ significantly between input instances. More specifically, the Bayesian framework leads to the modification of feature correlations and the means of marginalised features to varying degrees.

In this work, we use an alternative approach to address the need to take account of consistent relationships between features. We construct a multivariate Gaussian distribution to model these relationships. However, when updating this model between input instances, we only modify the mean and variance of those features present in the current model's coalition, leaving feature correlations intact. In this way, we preserve our knowledge of the underlying dataset, including the relationships between genes originally captured in the training data. This results in a more representative multivariate Gaussian distribution. Since we are aiming to draw insights about consistent relationships between genes, this promotes more realistic SHAP values and therefore cluster explanations.

3.2 Cluster analysis and gene module identification

Identifying gene modules is a crucial step in characterising the genetic component of disease. Current techniques tend to include a clustering aspect and/or network construction [16, 17, 18] to organise genes, such as Weighted Gene Co-expression Network Analysis [19, 20]. However, they can be sensitive to noise, with high computational complexity that can limit scalability to larger datasets. Our approach also uses clustering, but reduces complexity and the impact of noise by implicitly capturing gene and sample relationships. We achieve this by using Gaussian Mixture Modelling and a deep autoencoder that can infer both linear and non-linear relationships. We adapt our mixture-based clustering model for classifying disease subtype based on RNA-Seq data.

Our approach explicitly accounts for inter-feature dependence by analysing the underlying data distributions and correlations between genes, using data from real patients.

4 Data and Methods

Our work is comprised of two main stages. The first stage involved using dimensionality reduction and clustering techniques for patient stratification. Our data included patients with Crohn's disease (CD) and controls. The subjects were classified into subtypes within the dataset: CD with deep ulcer (the most severe form of the disease), CD without deep ulcer and controls. The goal of patient stratification was to discover these subtype groups.

The second stage involved adapting state-of-the-art methods for explainability, and identifying and characterising genes and gene modules associated with each disease subtype.

Patient stratification

1. Sampling from a publicly available genomic dataset
2. Dimensionality reduction using an autoencoder and PCA
3. Clustering and classification with GMMs and Kmeans

Explainability

4. Adapting kernelSHAP to identify genes that are involved in disease
5. Identification of potential gene modules by disease subtype
6. Class-contrastive technique for patient-specific explainability

4.1 Patient stratification

4.1.1 Sampling from a public RNA-Seq dataset

We performed our analyses on a publicly available transcriptomic dataset called RISK [21]. This contains RNA-Seq data from ileal tissue samples. The samples were taken from children: non-IBD controls as well as patients diagnosed with IBD who had not yet undergone treatment.

The RNA was sequenced as described in [21], resulting in normalised counts in RPKM (Reads Per Kilobase of transcript per Million mapped reads). We removed data on Ulcerative Colitis to focus on Crohn's disease (CD) subtypes. We used data from 260 individuals who were classified into three groups: CD with deep ulcer, CD without deep ulcer and non-IBD controls. Non-IBD controls were those with "suspected IBD, but with no microscopic or macroscopic inflammation and normal radiographic, endoscopic, and histologic findings". The goal of patient stratification was to develop a robust method for recovering these subtypes.

The first task was to sample a relevant selection of genes from the RISK dataset [21, 22, 23]. Many genes in the dataset did not vary much in expression across the sample of patients. We decided to analyse genes with a variance of at least 0.01 (normalised) RPKM across the sample, to help us identify those that could reasonably affect disease subtype.

In addition to the RISK dataset, a supplementary dataset from [21, 22, 23] contained 1,281 genes that were found to be differentially expressed with a fold change of at least 1.5, between two independent CD and control groups. For each independent group, we identified the top 60 most upregulated genes and top 60 most downregulated genes. This resulted in 240 genes, from which we identified 130 matches in the RISK dataset. Further exploration of the literature identified more genes associated with IBD, such as JAK2, NOD2 and LTA. 41 of these genes were discovered in the RISK dataset and added to the sample. To promote more diversity, we also added a random sample of 50 genes from the RISK dataset not linked to IBD. This was done to provide potential models with a broader range of gene types that could be used to differentiate between disease subtypes. However, this also requires a model to be robust to random noise. Our selection process resulted in a total of 221 genes.

4.1.2 Dimensionality reduction approach

An autoencoder model (adapted from [24]) was used to reduce dimensionality, using the given sample of 221 genes (and their associated expression values) in the RISK dataset. The dataset was randomly shuffled and split 70:30 into training and test sets. We then performed feature scaling [25]. The model was implemented using the Keras Functional API [26]. We experimented with different layers, layer sizes and activation functions to achieve a good performance. The architecture for the encoder and decoder sections is detailed in Table 1. 32 neurons were used in the bottleneck layer to produce a 32-dimensional latent representation.

To tune the model, we used the Keras 'GridSearchCV' function for a 5-fold cross-validated grid search over the hyperparameters: epochs, batch size, optimiser, learning rate and weight initialisation function. The final model had the following hyperparameters: batch size of 32, 150 epochs, Adam optimiser, 0.001 learning rate and Xavier normal initialisation function. We used 80% of the training data for this training, keeping the remaining 20% for validation. Mean Squared Error (MSE) was used as a performance metric.

In this work, we applied both PCA and an autoencoder for dimensionality reduction. In both cases, after reducing to 32 latent variables, t-SNE (t-distributed Stochastic Neighbour Embedding) was used for visualisation.

4.1.3 Gaussian Mixture Modelling (GMM) and KMeans clustering

We explored both GMM and KMeans clustering algorithms for patient stratification. We first divided the gene expression dataset into training, validation and test sets using a 70/15/15 split. After reducing dimensionality using the PCA-tSNE and autoencoder-tSNE methods explained previously, the GMM and KMeans algorithms were trained on the training set. We used 4 clusters over 3 disease subtype classes to allow for the discovery of more potential subtypes. Each component in the GMM was set to have an individual covariance matrix, to fit the model closely to the data. As the perplexity used for t-SNE can have a large impact on performance, we tuned this hyperparameter on the validation set, using accuracy, F1-score and silhouette analysis.

For classification, we designed a post-processing algorithm to assign each of the four clusters to a disease subtype (CD with deep ulcer, CD with no ulcer, or control). This is summarised in Algorithm 1. In essence, each cluster was assigned to the class with the greatest density of its datapoints present. For GMMs, the mixture component probability density functions were used in this estimation. This was replaced with a simple datapoint count for KMeans. We then managed possible duplicate assignments in $handle\_duplicates()$. More specifically, where two classes were initially assigned to the same cluster, the class with the greatest density present took precedence. The remaining class was then assigned to one of the remaining clusters with the greatest density of its points present.

The next step was to explore methods for the classification of disease subtype. Realising the power of GMMs, we implemented a function to generate a probability distribution across the disease subtype classes for any new data sample, given our trained clustering model and class assignments generated by Algorithm 1. We first evaluated the probability density of each mixture component with respect to the given point. For the class with two clusters assigned, we took the highest probability density of the two components, discounting the other component. For each datapoint, we normalised the resulting values by dividing each by the total sum, to ensure that the probabilities added up to one. Therefore, the probability of a datapoint x being classified as a, the class of interest, would be $\frac{p(x|C_a)}{\sum_{i=1}^{K} p(x|C_i)}$ where C represents a disease subtype class and $K = 3$, the number of disease subtypes. Therefore, we adapted the GMM into a probabilistic model for the classification of CD subtype. KMeans subtype predictions for the test set were based on the closest cluster center and its given subtype assignment, since no probability density function was available. The models were visualised [27] and evaluated on the test set using accuracy, F1-score and silhouette analysis [28, 29]. This allowed us to assess clustering quality and classification ability.

4.2 Clustering explainability

4.2.1 Modifying kernelSHAP to identify risk genes

SHAPley Additive exPlanations (SHAP) [4] is a state-of-the-art method for machine learning model explainability. KernelSHAP is a model-agnostic fast approximation of SHAP. This is commonly applied to regression or classification models. In this work, we developed a post-processing technique for Gaussian Mixture Models (GMMs). We produced a probability distribution over the disease subtypes for each patient. Therefore, we were able to couple our mixture models to kernelSHAP for patient-specific and global cluster explainability.

However, a major limitation of kernelSHAP in this context is that it assumes feature independence i.e. gene independence. In reality, gene expression in biological systems can be highly correlated between genes, and many genes are regulated by other genes within a complex network. Therefore, we extended kernelSHAP to incorporate inter-feature dependence, to enable more accurate cluster explanations.

During the calculation of SHAP values, we perform linear regression which involves calculating model output expectations for each coalition of features. If we denote S as the subset of features being included in the coalition for a model f, and $\overline{S}$ as the set of missing features, the expectation E for the model output can be calculated as follows:

$$E[f(\mathbf{x})|\mathbf{x}_S = \mathbf{x}_S^*] = E[f(\mathbf{x}_{\overline{S}}, \mathbf{x}_S)|\mathbf{x}_S = \mathbf{x}_S^*] = \int f(\mathbf{x}_{\overline{S}}, \mathbf{x}_S^*) p(\mathbf{x}_{\overline{S}}|\mathbf{x}_S = \mathbf{x}_S^*)\, dx_{\overline{S}} \tag{4}$$

as explained in [30], where $p(\mathbf{x}_{\overline{S}}|\mathbf{x}_S = \mathbf{x}_S^*)$ is the conditional distribution over the missing feature values, given a set of known values $\mathbf{x}_S$ for the subset of features S included in the given coalition, obtained from the current input of interest **x***. To simplify the process, the original kernelSHAP implementation assumes feature independence and instead draws from the marginal distribution [4]. Building upon work by Aas et al. in [30], we propose an adaptation to approximate the conditional distribution, resulting in more representative synthetic samples and therefore more realistic SHAP values.

We can approximate the conditional distribution by modelling the underlying data distributions of the training set using a multivariate Gaussian distribution. We calculate a mean vector and sample covariance across the training data to construct this distribution. For each new input instance, we modify the distribution according to the features present in the coalition, before drawing synthetic samples for the expectation calculation.

Because we aim to uncover the relationships between genes, we preserve the relationships captured by the sample covariance and expectation on the training data. We reduce the variance of a particular feature to zero if it is included in the coalition. The corresponding values in the mean vector are also updated to be equal to values given in the input instance. In this way, we maintain the important underlying distributions and correlations between features (genes).

We then adjust the matrix to make it positive definite by adding a small multiple of the identity matrix: $C = C + \epsilon I_p$, where I is a p x p identity matrix and $\epsilon = abs(\lambda_{min}) + b$, where λ_{min} is the smallest eigenvalue of C and $b = 1.5$. This ensures that we obtain a valid probability density function for sampling. The constant b can be tuned to affect ϵ; a larger ϵ value will result in greater overall strength of covariance between features. However, a value too large may distort the distribution. From the resulting distribution, we generate samples which are used to evaluate the expectations in Eq. 4. We must rescale each sample to account for the variances modified when making the covariance matrix positive definite, by applying the transformation $x = \boldsymbol{\sigma}(x - \boldsymbol{\mu}) + \boldsymbol{\mu}$. We also clip the values between 0 and 1 as the data is in normalised form.

Taking gene relationships into account results in synthetic samples with realistic background expression values for marginalised genes. We can therefore obtain more realistic SHAP values to explain the phenotypes predicted by the Gaussian Mixture Model. SHAP values are calculated on the test set, following the calculation of expected values using the training set. These calculations use the probabilistic model over Crohn's disease subtypes, derived from the Gaussian Mixture Model. We then generate individualised patient-specific plots and summary plots to explain the clusters.

4.2.2 Identification and characterisation of potential gene modules

We propose a method for identifying potential gene modules that explain disease subtype. The method relies on our kernelSHAP feature dependence adaptation and involves the following processes:

- Integration of SHAP values and gene expression data
- Consensus clustering
- Characterisation of gene modules
- Verification of gene modules using Gene Ontology enrichment analysis

Firstly, we prepared the data by combining different sources of information in a useful way. For each gene, we found how much it affects the model's prediction for a specific disease subtype. This was done by averaging how important the gene is across all patients with that disease subtype. Then, we multiplied this average importance by the gene's activity level across all patients.

The process is shown in Equation 5. This equation calculates representative values for each patient and gene in a specific disease type. Here, x represents a gene's activity level, and s represents its importance according to the model (SHAP value). We add up the importance values (using c_i to identify patients in a particular disease subtype) and divide by the total number of patients (n) to find the average importance for each gene.

The resulting dataset shows how active each gene is and how much it influences the model's prediction. This prediction represents how confident we are about assigning a particular disease subtype to a patient.

$$v_{pg} = x_{pg} \frac{\sum_i abs(s_{ig}) c_i}{n} \tag{5}$$

We then performed consensus clustering on the integrated data, using Weighted Ensemble Consensus of Random (WECR) K-Means, proposed by Yongxuan et al. in [3]. To summarise the WECR algorithm, we first run the K-means algorithm several times on random samples of the data using random subspaces of features, to generate an ensemble of base partitions. The value of k is also randomised for each run. In this work, we run K-Means 100 times, where each single run draws 80% of the sample and 50% of the features. To obtain the final clustering, we evaluate each base partition and form a co-association matrix. We then apply cluster-based similarity partitioning and spectral clustering [31]. Please see [3] for more details about the WECR algorithm.

As the number of gene modules in the data is unknown and dimensionality is high, applying this consensus clustering is suitable to obtain more stable clusters. For the same reason, we integrated different types of data and used four different validation metrics to select the optimal number of clusters k from 2 to 9. These were the Bayesian Information Criterion (BIC), Davies-Bouldin (DB) Index, Silhouette Score (SIL) and Calinski-Harabasz (CH) Index, where BIC and DB should be minimised, and SIL and CH should be maximised. In this way, we obtained representative modules informed by similarities in expression pattern and influence on disease subtype. We then displayed aggregate bar plots to show the relative contributions made by each gene module in predicting a particular disease subtype. This is represented using a sum of the mean SHAP values across all genes in the module (where the mean is calculated across all patients of the given disease subtype).

Incorporating inter-feature dependence provides valuable insights into sets of genes which could be working in concert: this includes the type and magnitude of their influence on the model's prediction of disease subtype.

Finally, we confirm the functional relevance of the gene modules using Gene Ontology enrichment analysis.

4.2.3 Class-contrastive technique for patient-specific explainability

We develop a class-contrastive technique for explaining clusters specific to a patient. Class-contrastive reasoning generates an explanation by providing a contrast to another class. An example of a class-contrastive explanation is: "The selected patient is predicted to be in a severe disease subtype (CD with deep ulcer) because all genes in a particular module were overexpressed. If these genes had abundance similar to genes in the control group, then the patient would be predicted to be in a less severe disease category (CD without deep ulcer)."

The expression of each gene approximately follows a normal distribution. Some genes were upregulated or downregulated in patients with CD (Crohn's disease) compared to controls. Therefore, for a specific patient with CD, we can generate a class-contrastive explanation in the following way: we can modify the expression of genes in a given module so that they are more similar to controls. We did this by assigning a new expression value v for each chosen gene, as the mean value for the expression of this gene across all control individuals, as shown in Eq 6.

$$\forall g \in G, \quad v_{pg} = \frac{1}{N} \sum_{i} c_i x_{ig} \tag{6}$$

where x is an expression value, p is the selected patient, g is the selected gene and G is the full set of genes in the module. We sum over all control individuals i in finding the mean, where c_i is the indicator variable for the control group and N is the total number of control individuals.

We can reduce the dimensionality, as described in Section 4.1.2, and generate the GMM clustering for the dataset before and after this correction. If the patient with CD moves into a different CD cluster after correction, we can infer how the genes in the module may be affecting the disease. In Section 5.4 we apply this technique for intuitive patient-specific explainability, showing how gene modules can contribute to CD subtype. As the identification of gene modules relies on the feature dependence extension proposed in Section 4.2.1, this also takes gene correlations into account across all patients. This method combines the strengths of feature attribution, consensus clustering and class-contrastive reasoning.

4.3 Strategies to mitigate overfitting

To mitigate overfitting, we employed several strategies. Initially, we reduced the dimensionality of the high-dimensional RNA-Seq data using the autoencoder and t-SNE. This step helped minimise the model's capacity to overfit by focusing on fewer and more informative features in a lower-dimensional latent space. We split our dataset into training, validation, and test sets to ensure proper model evaluation and generalisation. This approach allowed us to fine-tune our models on the validation set and evaluate their performance on the unseen test set.

For t-SNE dimensionality reduction, we tuned the perplexity parameter based on silhouette score, accuracy, and F1 score on the validation set for the application of our GMM and post-processing algorithm. This tuning ensured that the reduced dimensionality representations were optimal for subsequent clustering and classification tasks on the test set.

During autoencoder training, we applied batch normalisation to stabilise and enhance the training process. Additionally, we conducted a grid search over various hyperparameters, including optimiser, epochs, batch size, learning rate, and initialisation function. This comprehensive search helped to identify the optimal configuration, reducing the risk of overfitting.

The loss plot (Figure 1) demonstrates that the autoencoder is not overfitting. The training and validation loss curves decrease steadily and converge, indicating that the model performs well on both the training and validation datasets without significant divergence. This suggests that the model is learning the underlying patterns in the data rather than memorising the training set.

5 Results and Discussion

In this section, we evaluate and discuss the significance of our results.

5.1 Gaussian Mixture Model (GMM) and KMeans clustering

We used both PCA and an autoencoder for dimensionality reduction of the data, each alongside tSNE for visualisation in two dimensions. We then apply the clustering techniques. The steps are as follows:

1. PCA $\rightarrow$ tSNE $\rightarrow$ KMeans
2. Autoencoder $\rightarrow$ tSNE $\rightarrow$ KMeans
3. PCA $\rightarrow$ tSNE $\rightarrow$ GMM
4. Autoencoder $\rightarrow$ tSNE $\rightarrow$ GMM

KMeans and Gaussian Mixture Models were implemented to cluster the samples into 4 groups. We then added a post-processing step to transform the models into classifiers of disease phenotype for Crohn's disease (CD). The three classes/disease subtypes were "CD with deep ulcer", "CD no ulcer" and "control". We used both PCA and an autoencoder for dimensionality reduction of the data, each alongside tSNE for visualisation in two dimensions. We then apply the clustering techniques.

Our autoencoder model shows good performance, achieving a MSE of 0.0143 on the test set, despite its simplicity compared to the state-of-the-art [32, 33]. This suggests good capabilities in reducing dimensionality of the RNA-Seq data, while retaining important information.

The final results for GMM clustering on the test set are shown in Figure 2, using dimensionality reduction by the autoencoder (left) and PCA (right). The final evaluation results are summarised in Table 2.

The results show a good overall performance (Table 2). Clustering provides an informative visual representation of relationships between patients in terms of disease subtype. In particular, the GMM provides an effective density estimation which is useful for inferring an accurate disease subtype during post-processing. Our autoencoder also performs better than PCA, particularly in the context of multi-class classification of disease subtype. Here, accuracy and F1-score were higher by 7.7% and 8.9% respectively when using our autoencoder compared to PCA (Table 2) (when using GMMs). In addition, by using a greater number of clusters than disease subtypes, we may discover additional substructures which could correspond to potentially new subtypes of CD.

5.2 Cluster explanation using kernelSHAP adapted for feature dependence

We can couple our GMMs to kernelSHAP [4] to generate explainability for each cluster, including visualisations [34]. As each feature corresponds to a gene and each cluster class corresponds to a disease subtype, the resulting SHAP values represent the importance of each gene in predicting a particular disease subtype for a given patient. Additionally, we modify the original kernelSHAP method to incorporate feature dependence. This more accurately models the living system, as many genes are highly correlated and/or regulated by other genes. Using this method, we can therefore identify the genes that are the most influential in predicting given disease subtypes.

5.2.1 Waterfall plot

A waterfall plot can be used to analyse a single patient, as shown in Figure 4. This explains the model output for the CD deep ulcer cluster class. It shows a quantification of the contributions made from the top genes identified, alongside that of the remaining genes. In this way we can see how the model output has been shifted from the expected value $E[f(z)]$, to the actual output, $f(x)$. In the former, the model f is provided with a baseline sample z and no information about the features, whereas in the latter we provide our actual data sample x as input to the model. These values are given in the log-odds space.

The genes shown are highly relevant. For example, IRGM is a negative regulator of IL1B as it suppresses NLRP3 inflammasome activation by hindering its assembly. In this way, it has a protective effect against inflammatory cell death and gut inflammation in Crohn's disease [35]. In the plot we can see that IRGM has a relatively low normalised expression level of 0.093 for this patient, which would have little protective effect. This rightfully leads the model to predict a greater probability of CD with deep ulcer. In comparison to the plot resulting from feature independence using the original kernelSHAP method (Fig. 5), we obtain more genes specifically related to IBD for the same patient (who is diagnosed with CD with a deep ulcer). For example, in addition to IRGM, we also obtain HLA_DRB1, MEP1B, MUC1 and SLC11A1, which all have established links to IBD [36, 37, 38, 39].

5.3 Identification and characterisation of potential gene modules

We proposed a novel method to identify potential gene modules. This integrates SHAP values with gene expression values, before performing Weighted Ensemble Consensus of Random consensus clustering [3]. We utilise SHAP values calculated using our kernelSHAP adaptation. This incorporates dependence between genes. We then use the SHAP values to characterise each gene module, verifying our findings with Gene Ontology enrichment analysis.

We first applied our technique to identify gene modules associated with the most severe form of the disease: Crohn's disease with a deep ulcer. This was achieved by integrating the CD deep ulcer SHAP values with the expression values across all patients. We then perform Weighted Ensemble Consensus of Random (WECR) KMeans clustering. We select $k = 4$ as it achieves a good balance in terms of the validation scores. Using a variety of metrics in the selection of k helps us achieve more stable clusters.

Figure 7 shows a bar plot in which each final gene module is characterised in terms of its influence on model predictions. Each bar represents the sum of mean SHAP values associated with the "CD deep ulcer" cluster, across all genes in the module. We show the top 4 most influential genes of each module. More positive values (green) indicate that the given module increases our confidence in a "CD deep ulcer" prediction and more negative values (red) reduce our confidence in a "CD deep ulcer" prediction.

5.4 Class-contrastive explainability

Finally, we propose a class-contrastive method as an additional approach to cluster explanation. After identifying gene modules relevant to particular clusters (Section 5.3), we can demonstrate their impact on disease subtype (CD deep ulcer, CD no ulcer and control) in a visual way. We again utilise the autoencoder for dimensionality reduction, as this leads to better classification performance in comparison to PCA.

The expression of each gene (normalised counts), across patients and disease subtypes, follows a Gaussian distribution (Figure 11). We observe that some genes are downregulated or upregulated in patients with CD compared to controls. For example, in Figure 11a, the distributions corresponding to "CD no ulcer" and "CD deep ulcer" are shifted higher compared to the control distribution: this shows upregulation of CXCL3, with higher mean expression levels. However, in Figure 11b, the distribution of the gene MEP1B in patients with CD is shifted lower, compared to controls: this shows downregulation of MEP1B compared to controls.

After selecting a patient with CD, we generate a class-contrastive explanation by modifying the expression of particular genes to make the genetic profile more similar to those of controls (patients without CD). This is achieved by changing the expression value of each gene to the mean expression value across control individuals (explained in Section 4.2.3 and Eq 6). We can discover the effect of various gene modules, such as those we identified, by modifying their expression in this way. We will demonstrate the process with Patient 46, who is a CD patient with a deep ulcer. Their initial position in the clustering model is shown in Figure 12.

We first select the 117-gene module, which was found to make the greatest positive contribution to CD deep ulcer predictions (Figure 7). We modify this set of genes for the patient (using the class-contrastive technique explained above) and refit the model with the same set of parameters. This results in the patient (Patient 46 with CD deep ulcer [a severe form of the disease], Figure 12) being assigned to the "CD no ulcer" cluster [a less severe form of the disease], as shown in Figure 13a. Note that the GMM can change slightly each time we refit the model due to stochasticity in the tSNE algorithm used for visualisation. However, the general structure of the model remains the same. Modifying these genes resulted in the model predicting a less severe form of the disease (CD without deep ulcer). This may suggest that this module contributes to a severe form of the disease (CD with deep ulcer).

Alternatively, we can select the 117-gene module and 63-gene module, which together were found to account for all positive contributions to CD deep ulcer predictions (Figure 7). Modifying these 180 genes and refitting the model results in Patient 46 being assigned to the control cluster (Figure 13b). As the patient has moved from the deep ulcer cluster directly to the control cluster, we can infer that some or all of these 180 genes play a role in a severe form of the disease.

When this class-contrastive method is coupled with our gene module identification method, we can visually explore the effect of identified gene modules on the patient's disease subtype. Although the GMM provides an excellent general representation for stratifying patients, we note that the classifier is not 100% accurate. Our work provides a proof-of-concept; scaling up to larger datasets in future work would likely further improve accuracy and F1-score.

The class-contrastive technique provides intuitive patient-specific explainability. As the identification of gene modules relied on our kernelSHAP extension, all patients and genes were taken into account (including correlations between genes).

5.5 Validation with an independent cohort

To validate how well our approach generalises, we applied our methods to an independent cohort [40, 23]. This dataset contains ileal gene expression profiles for 210 treatment-naïve paediatric Crohn's disease patients and 35 non-IBD controls at diagnosis. Importantly, the disease subtypes in this cohort differ from our original dataset. Instead of "CD no ulcer" and "CD with deep ulcer", this cohort includes the subtypes of "CD no complication" and "CD with complication".

We pre-processed the raw RNA-seq counts using upper quartile normalisation. 13,769 genes were filtered using an expression threshold of 0.1 RPKM and variance threshold of 0.01, required in at least half of the samples. This resulted in 13,661 genes. We then performed differential expression analysis using Welch's t-test, with volcano plot thresholds of 0.9 and -0.9 for p-value and 25 for log-fold change. This identified 278 differentially expressed genes, which were used for subsequent analysis.

We trained a new autoencoder with a similar architecture to our original model, adjusting the layer sizes to account for the new number of genes in this dataset. Specifically, the encoder has 556, 278, and 32 units, while the decoder has 278, 556, and 278 units. These changes increased the total number of parameters to 322,234 in the encoder. We then applied our full computational framework to this independent cohort, including:

1. Dimensionality reduction using the autoencoder and t-SNE.
2. Training a Gaussian Mixture Model (GMM).
3. Applying post-processing to obtain a classifier.
4. Coupling to kernelSHAP to identify key risk genes.

This validation step allowed us to assess whether our methodology could be successfully applied to a new dataset with a different set of genes.

5.5.1 Clustering and classification in an independent cohort

Table 3 shows our clustering and classification results after applying our dimensionality reduction and GMM followed by post-processing. These scores are calculated using the test set, with visualisations shown in Figure 14. We achieve very similar results to those of the original study. For example, we achieve excellent scores for binary classification, in which we discriminate between control and CD patients, with 89.2% for accuracy and 93.3% for F1-score when utilising the GMM following the autoencoder and tSNE. These are marginally lower than the corresponding scores for the original study. The silhouette score is also very similar at 0.366 (compared to 0.382).

As with the original study, the scores for multi-class classification are slightly lower. With the autoencoder and GMM setup, we obtained a 64.9% accuracy and 69.9% F1-score and 0.255 silhouette score. These are similar to the corresponding results of the original study but slightly less compelling when compared to the binary setting. This reflects the real-world challenge of distinguishing between subtypes of Crohn's disease i.e. those who will go on to develop a complication and those who will not. It must be noted that these subtypes are different to the subtypes of the original dataset (deep ulcer, no deep ulcer), which may result in lower performance as it may be more difficult to distinguish between these subtypes using only RNA-Seq data. As with the original study, using PCA for dimensionality reduction resulted in much lower scores in nearly all cases, as compared with our trained autoencoder, which speaks to the strength of our autoencoder.

The main limitation seems to be in the clustering quality, as there is clearly room for improvement in the silhouette scores. From the visualisations in Figure 14, we can see that the clusters could be separated more definitively. This may be due to the pre-processing techniques, in which we apply both our autoencoder and tSNE for dimensionality reduction before clustering. In future work, rather than clustering in two dimensions, it may be beneficial to train the GMM in a higher dimensional space e.g. after reducing to 32 dimensions using the autoencoder, and perhaps using a higher cluster count. This increases the complexity but may lead to more defined and separated clusters which could result in higher classification scores, especially as the post-processing algorithm uses the density estimation to classify individual patients.

Overall, we achieve very similar performance to that of the original dataset, which shows that our pre-processing, clustering and classification pipeline can be applied successfully to a new dataset to discriminate between subtypes of Crohn's disease. In the next section, we will explore the identification of risk genes using this dataset.

5.5.2 Gene identification in an independent cohort

After obtaining our GMM classifier, we coupled this to kernelSHAP, as explained in Section 4.2.1, to identify influential genes with regard to the model's subtype predictions.

Figure 15 presents a summary plot highlighting the top 20 most influential genes, many of which play a significant role in the pathogenesis of IBD. Among these, the top three genes—CXCL5, IL1B, and CXCL6—are particularly noteworthy for their high expression levels in IBD patients. Greater expression of these genes increases the circulating levels of cytokines and chemokines which are crucial drivers of inflammation in the intestinal environment [41]. Other genes in the plot also have strong links with IBD, further emphasising their potential roles in associated biological pathways. For example, FMO5 has been found to play a role in goblet cell maturation and mucus barrier formation, having a protective effect against IBD [42], while ALDH1A2 has been associated with a disruption of vitamin A signalling during active IBD [43].

From the length of each bar, we can see that each gene influences the predictions of each class to varying degrees. For example, CXCL5 is the gene with most influence over "CD with complication" predictions, which shows that it may be a significant factor and/or indicator that intestinal complications may develop following the initial diagnosis of Crohn's disease. This is plausible given that CXCL5 is a known inflammatory cytokine.

One potential limitation of our study is that some of these genes have weaker connections to IBD, such as ALPL. This may be due to the performance of the classifier for multi-class classification. However, these may represent novel findings. For example, WNT5A (ranked 12th on the list) has been found to regulate inflammatory cytokines and has been identified as a potential novel therapeutic target for IBD [44]. Our system therefore may have applications in the advanced discovery of novel risk genes. Future work may confirm or disprove the findings. Note that a slightly different set of genes was used in this validation study as compared to the original study, with different subtypes under examination, so we would not necessarily be expecting the same set of resulting genes to appear in the summary plot.

In summary, our machine learning framework is very effective in identifying both known genes and novel genes which are indicative of different subtypes of Crohn's disease. Our techniques can be easily applied to a new dataset, achieving similar overall performance. We believe that our adaptation of kernelSHAP has been instrumental in developing a robust technique for the identification of the most influential genes in disease pathogenesis, by taking account of key gene correlations and dependencies. In future work, we could explore the application of these techniques on different RNA-Seq datasets for various genetic disorders to assess their efficacy in a broader context. The applications could be very impactful; for example, identified genes could be used to aid diagnosis and potentially as therapeutic targets for the treatment of genetic disorders such as Crohn's disease.

6 Conclusions and discussion

6.1 Summary

We develop techniques to improve the interpretability of models for patient stratification. We adapted and applied machine learning techniques to transcriptomic data from patients with Crohn's disease (CD), and identified genes and gene modules that are functionally relevant to the disease. We confirm these results using a range of peer-reviewed research, as well as Gene Ontology enrichment analysis. Our novel contributions are summarised below:

- Mixture-based patient stratification and classification into Crohn's disease (CD) subtypes based on gene expression.
- Adaptation of kernelSHAP for inter-feature dependence; application to Gaussian Mixture Models (GMM) for identification and ranking of genes by disease subtype.
- Data integration technique and consensus clustering [3] to identify potential gene modules by disease subtype. Characterisation of gene modules and confirmation using Gene Ontology (GO) enrichment analysis.
- A class-contrastive technique to visually explain the impact of gene modules on disease subtype for each patient.

Ethical considerations

Our study uses RNA-Seq data from paediatric patients, raising important ethical considerations about data use and informed consent. As the data is already anonymised and publicly available, formal ethics approval was not required for our analysis. However, it remains crucial to respect the original consent provided by participants, especially since the data is being reused in AI research. Transparency about how the data is utilised and ensuring that it aligns with the participants' expectations are key to maintaining trust and meeting ethical standards. The original dataset is available at at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57945.

We also took into account the ethical principle of the duty of easy rescue, as discussed by Mann et al. [45], when weighing the risks and benefits of our research. This principle suggests that when the cost to individuals, such as contributing anonymised data, is small, and the potential benefit to society is significant, there is a moral obligation to proceed. In this case, while the risks to individuals are minimal, the potential benefits, such as improved disease classification and treatment, are considerable. Additionally, following the guidance from Banerjee et al. [46], involving patients in the research process builds trust and ensures that our work meets public expectations, supporting the ethical use of this sensitive data.

Acknowldegements

We would like to express our gratitude to the Cambridge Trust and Google DeepMind for their generous support through the DeepMind Cambridge Scholarship, awarded to SO for her postgraduate studies at the University of Cambridge. SB acknowledges funding from the Accelerate Programme for Scientific Discovery Research Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the authors and not necessarily those of the funders.

We would also like to thank the paediatric study participants and the National Center for Biotechnology Information (NCBI) for their support in making the IBD RNA-Seq dataset freely accessible to the public.

Algorithm 1 Post-processing algorithm for Gaussian Mixture Model (GMM) to classify disease subtype

```
  Input: GMM, X_train, true_labels
  Output: Cluster-to-class assignments
amounts ← 3x4 matrix
for class in classes do                                   ▷ Datapoint density estimation
   for component in GMM.components do
      X ← X_train[true_labels==class]
      contribution ← sum(component.PDF(X)) × component.weight
      amounts[class][component] ← contribution
   end for
end for
cls_assignments ← handle_duplicates(argmax(amounts, axis=1))
class_amounts ← max(amounts, axis=1)
assignments ← 1x3 matrix
assigned ← 0
while assigned < 3 do                            ▷ Assign clusters in descending fashion
   curr_max_class ← argmax(class_amounts)
   assigned_cluster ← cls_assignments[curr_max_class]
   if assignments[curr_max_class] is None then
      assignments[curr_max_class] ← [assigned_cluster]
   else
    assignments[curr_max_class].append(assigned_cluster)
   end if
   class_amounts[curr_max_class] ← -1
   assigned ← assigned + 1
end while
rem_cluster ← setdiff(arange(4), cls_assignments)        ▷ Assign remaining cluster
rem_cls_assignment ← argmax(amounts[:,rem_cluster], axis=0)
assignments[rem_cls_assignment].append(rem_cluster)
```

Table 1: Autoencoder architecture.

Layer type	Output size	# Params
Encoder		
Dense	442	98124
Batch Normalisation	442	1768
LeakyReLU	442	0
Dense	221	97903
Batch Normalisation	221	884
LeakyReLU	221	0
Dense	32	7104
Decoder		
Dense	221	7293
Batch Normalisation	221	884
LeakyReLU	221	0
Dense	442	98124
Batch Normalisation	442	1768
LeakyReLU	442	0
Dense	32	97903

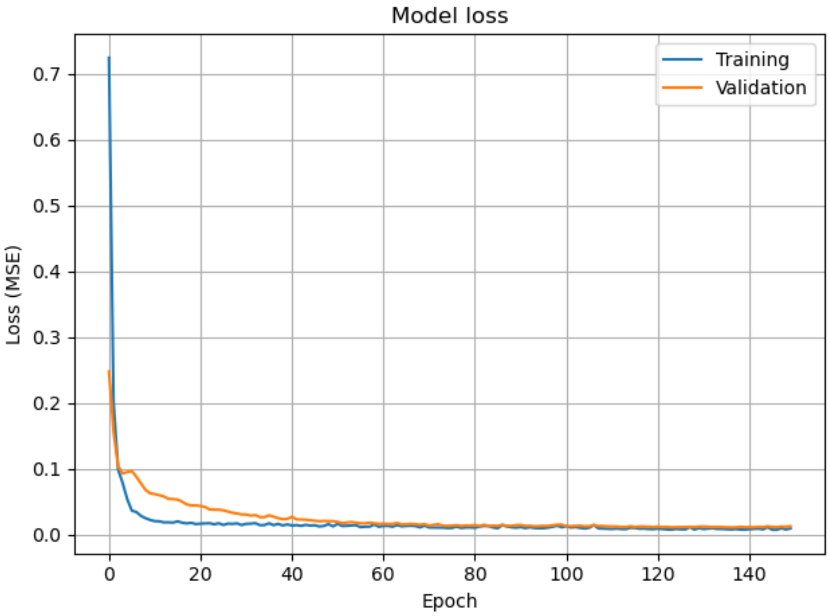


Figure 1: Autoencoder: Training and validation loss curves - MSE recorded across 150 epochs. The steady and converging loss curves show that the autoencoder is not overfitting.

Table 2: Clustering and classification evaluation results for novel classifiers based on Gaussian Mixture Model (GMM) and KMeans models, using autoencoder and PCA dimensionality reduction methods. Results shown for binary classification (controls and all CD patients) and multi-class classification (control, CD no ulcer and CD deep ulcer) of disease subtype.

		Binary (control & CD)		**Multi-class (all labels)**	
		Autoencoder	PCA	Autoencoder	PCA
GMM	Accuracy / %	94.9	92.3	71.8	64.1
	F1-Score / %	96.7	94.9	71.5	62.6
	Silh. score	0.382	0.410	0.320	0.317
KMeans	Accuracy / %	84.6	82.1	64.1	59.0
	F1-Score / %	89.3	88.1	61.9	58.3
	Silh. score	0.556	0.409	0.469	0.334

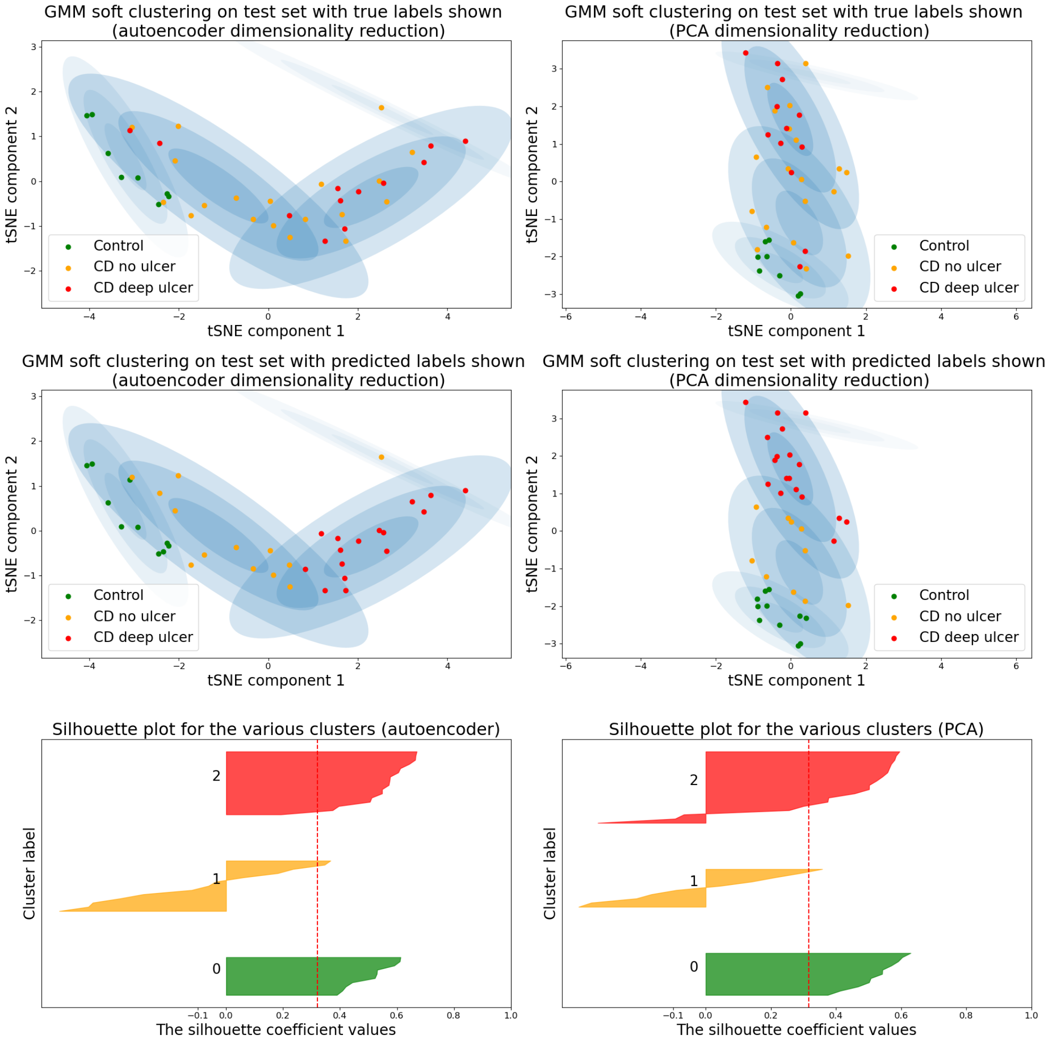


Figure 2: Gaussian Mixture Model (GMM) clustering model results after applying dimensionality reduction using autoencoder and tSNE (perplexity=130) (left) and PCA and tSNE (perplexity=150) (right). Deployed on the test set with true labels shown (top third) and predicted labels shown (middle third). Silhouette plots are shown for GMM clusters after applying autoencoder-tSNE (left) and PCA-tSNE (right) methods, with clusters 0, 1 and 2 corresponding to “control”, “CD no ulcer” and “CD deep ulcer” respectively. These are the disease subtypes and CD deep ulcer is the most severe form of the disease.

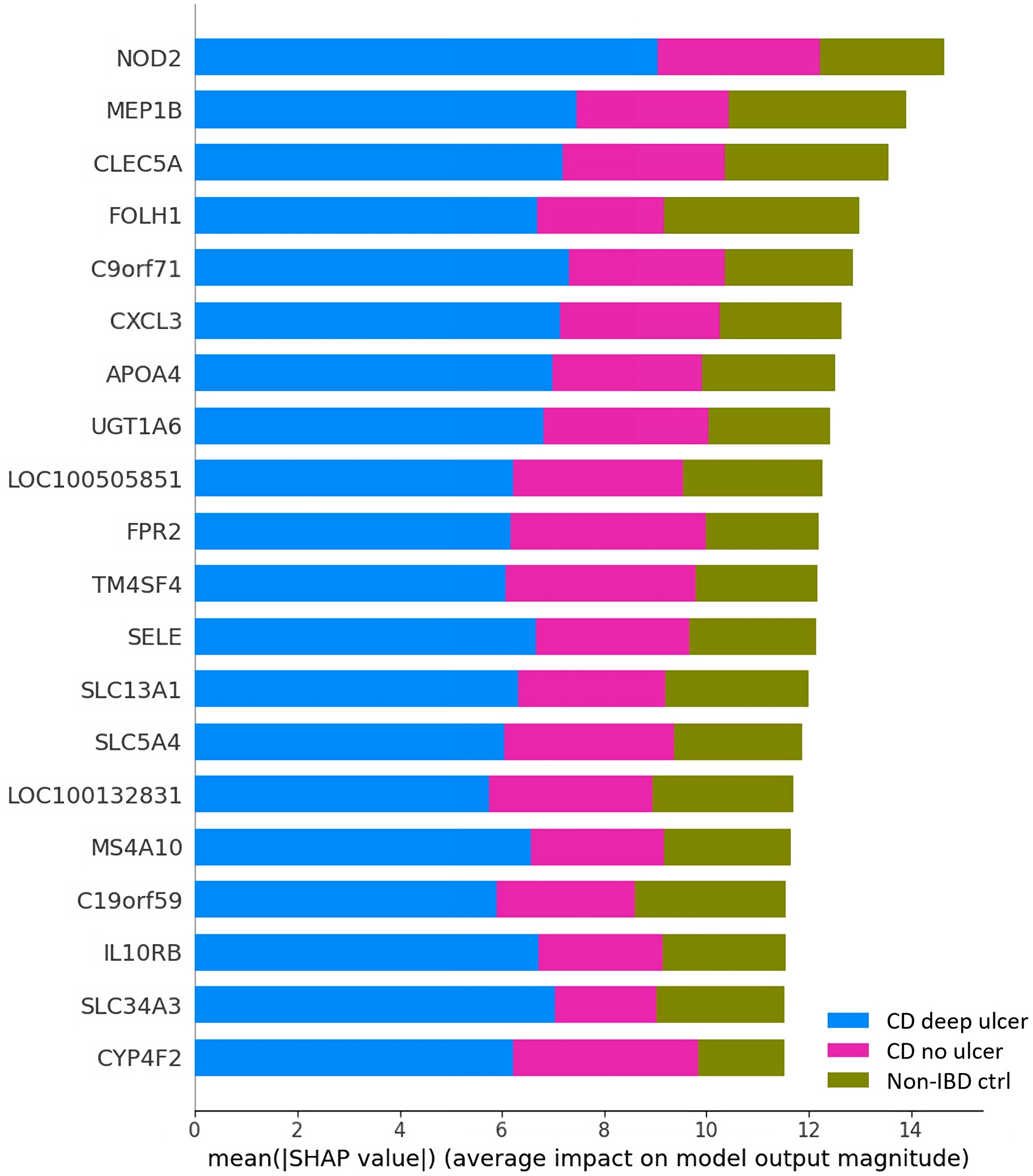


Figure 3: Summary plot showing top 20 genes in terms of their average impact on class predictions across all patients. This is for a model which accounts for feature dependence. The blue, pink and green bars depict the magnitude of influence of a gene on the "CD deep ulcer", "CD no ulcer" and "control" classes respectively. The greatest proportion of influence of genes is attributed to the "CD deep ulcer" cluster. Genes like NOD2, MEP1B and FOLH1 are within the top 5 and known as susceptibility genes for IBD. The most significant gene identified overall is NOD2, which is known to be strongly associated with IBD.

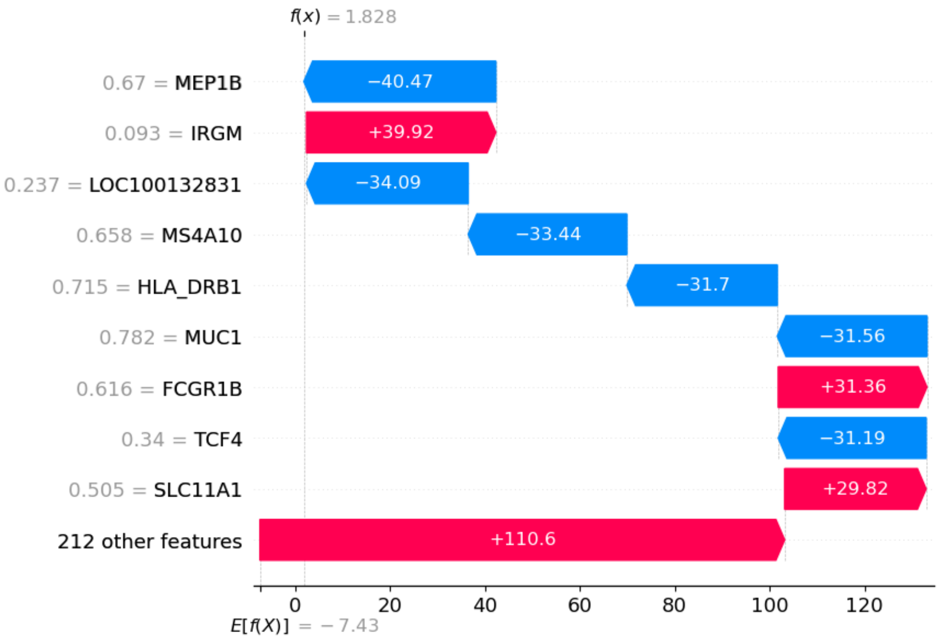


Figure 4: Waterfall plot for analysing model output for a single patient (Patient 260) for the "CD deep ulcer" disease subtype (using dependent features). Shown are contributions made from the top genes identified, alongside that of the remaining genes. The plot shows how the model output has shifted from the expected value $E[f(z)]$, where the model has no information about the features, to the actual output, $f(x)$ (values are in log-odds space). Shown are some highly relevant genes, such as IRGM [35] which is a negative regulator of IL1B as it suppresses NLRP3 inflammasome activation. It has a protective effect against gut inflammation in CD. IRGM has a relatively low normalised expression level of 0.093 for this patient, which may have little protective effect. The analysis shows that IRGM leads the model to predict a greater probability of CD with deep ulcer in this case.

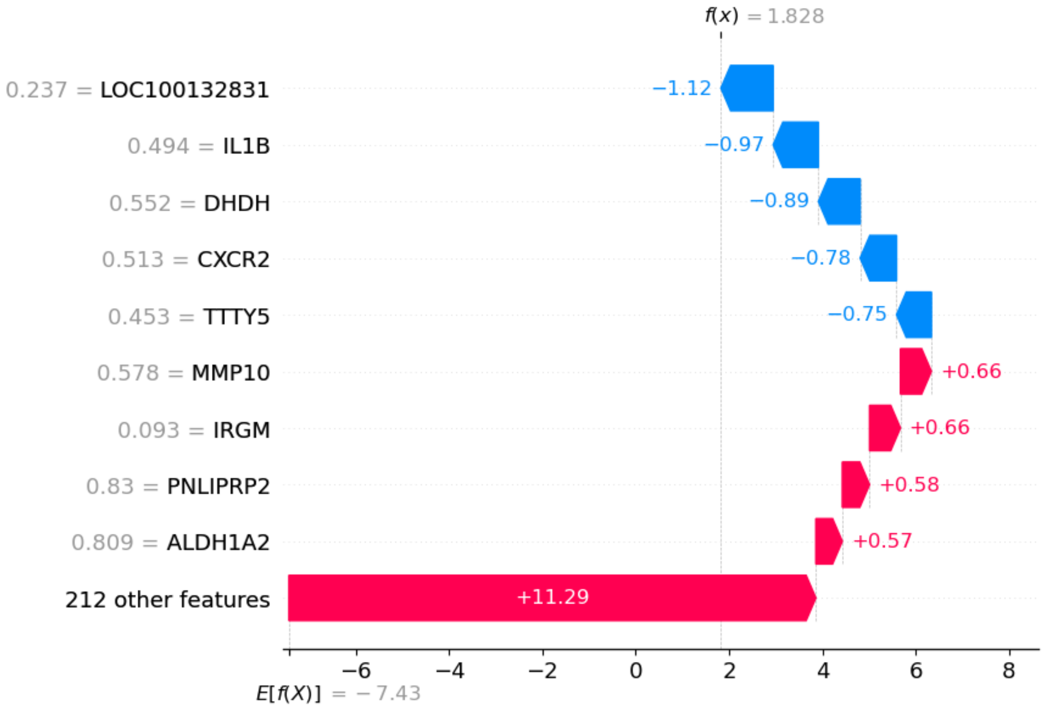


Figure 5: Waterfall plot of Patient 260 for "CD deep ulcer" class - independent features.

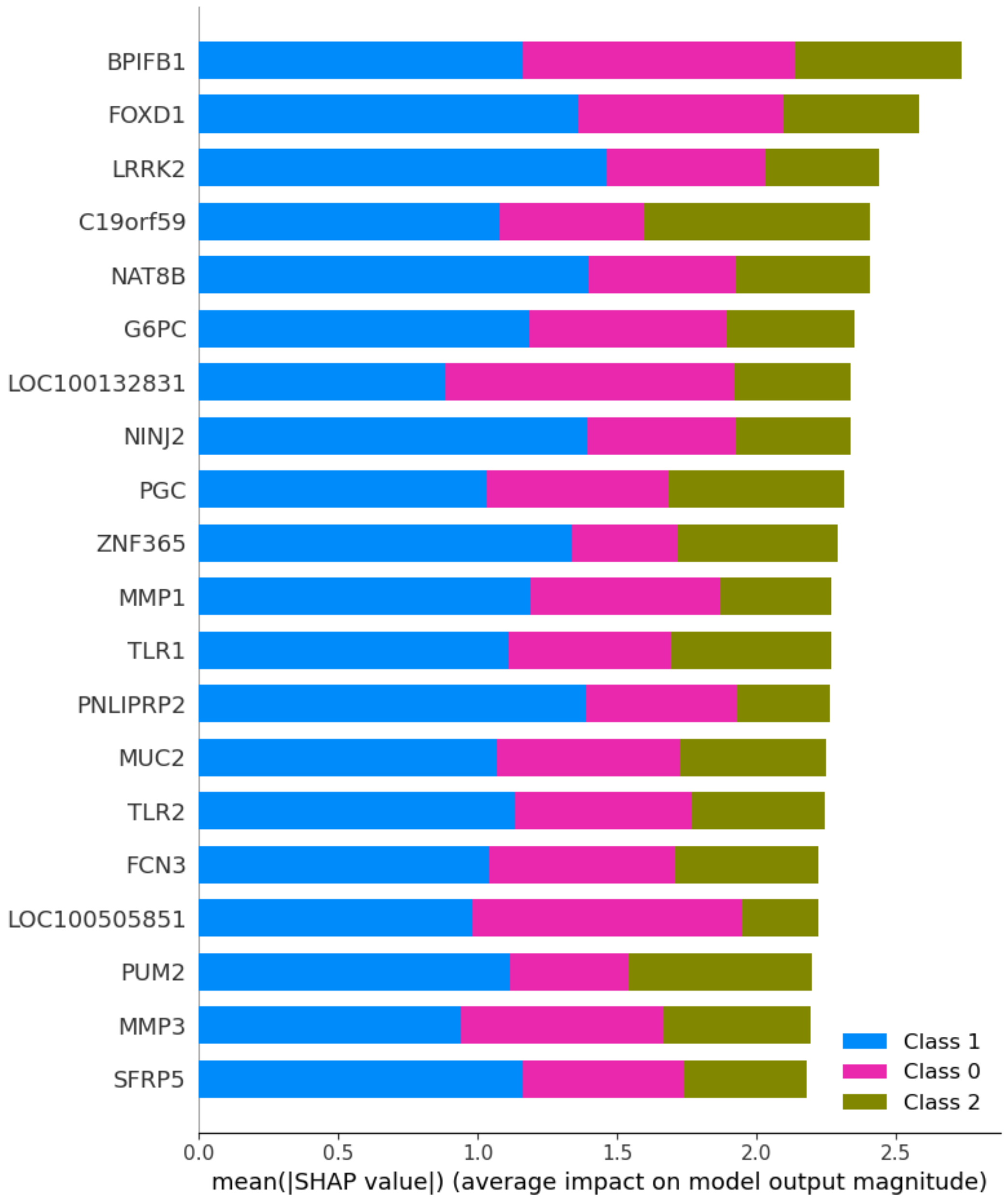


Figure 6: Summary plot showing top 20 genes in terms of their average impact on class predictions across all patients - independent features. Genes ranked by importance.

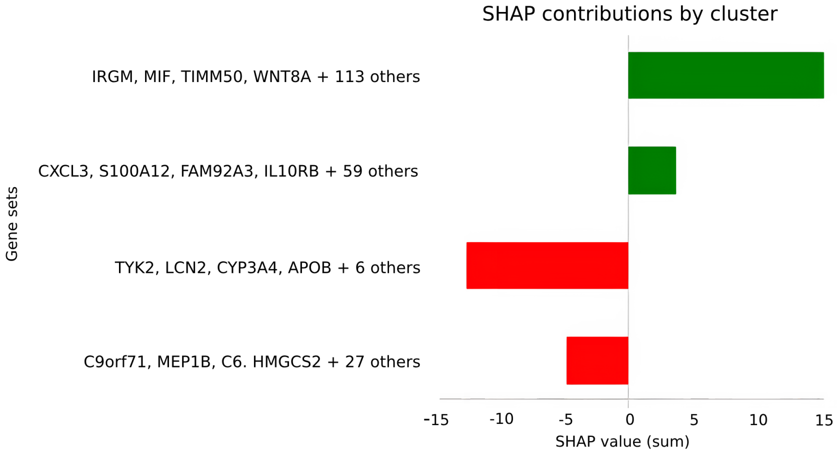


Figure 7: Final gene modules identified as being associated with severe disease (CD deep ulcer), alongside relative contributions determined using SHAP values. Shown is a bar plot in which each gene module is characterised in terms of its influence on the model predicting the "CD deep ulcer" cluster. Each bar represents the sum of mean SHAP values associated with the "CD deep ulcer" cluster, across all genes in the module. We show the top 4 most influential genes of each module. More positive values (green) indicate greater confidence for that module predicting "CD deep ulcer". More negative values (red) reduce our confidence in a "CD deep ulcer" prediction. Consistent with the literature on IBD, we find genes like IRGM, CXCL3 and IL10RB are present in the modules and have a significant effect on disease severity.

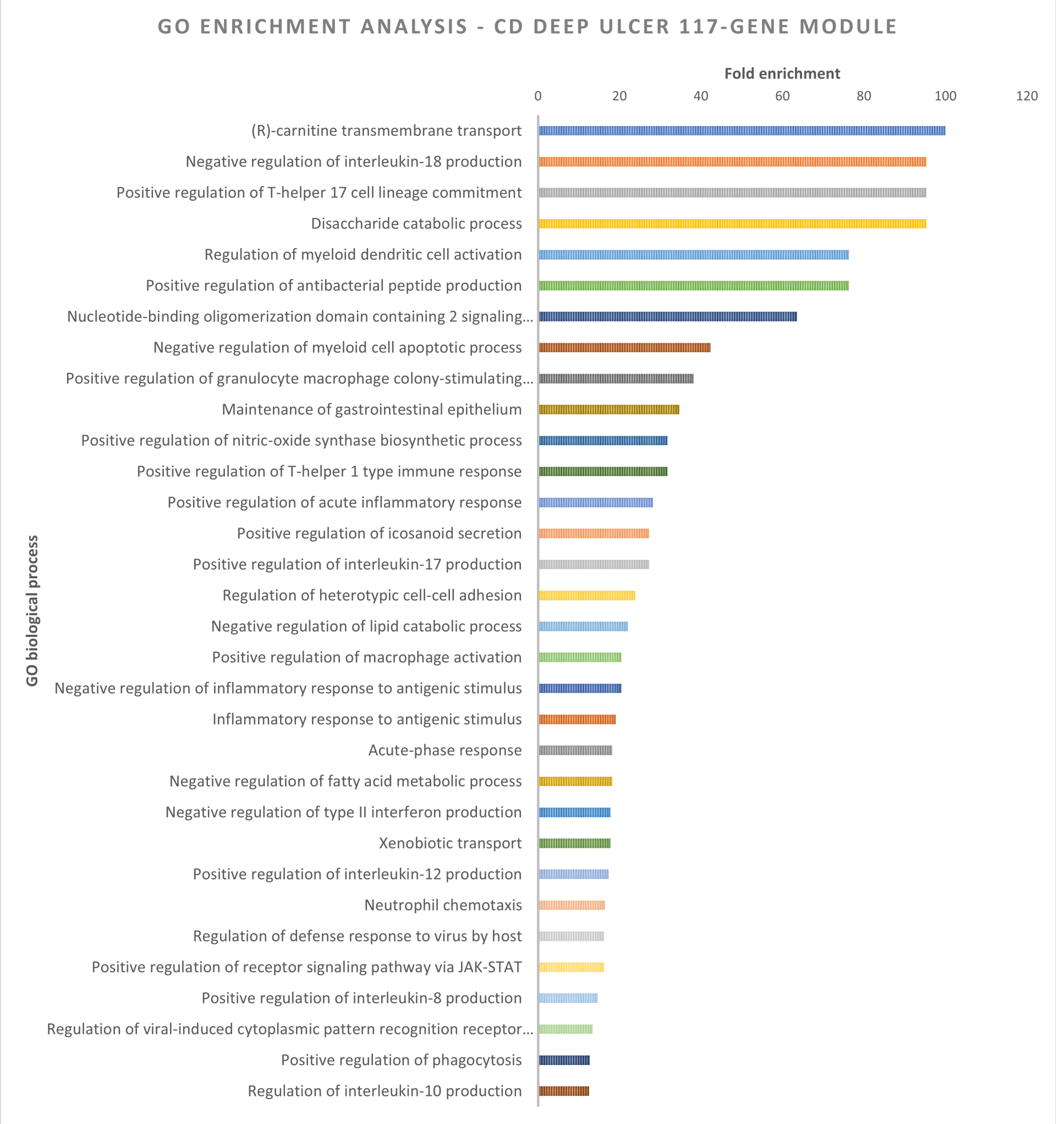


Figure 8: Gene Ontology enrichment analysis. Shown are most enriched biological processes associated with a 117-gene module which was found to be strongly associated with CD with deep ulcer. The GO processes include transport mechanisms, regulation of reactive oxygen species, cell adhesion, and regulation of immune response to viruses and Gram-negative bacteria, which is in line with our current knowledge of IBD. Our results are statistically significant with a false discovery rate (FDR) of less than 0.05 and fold enrichment of up to or exceeding 100 for many of the processes. The most enriched processes are related to T-cell proliferation, transmembrane transport, signalling pathway activation, and production of antibacterial peptides.

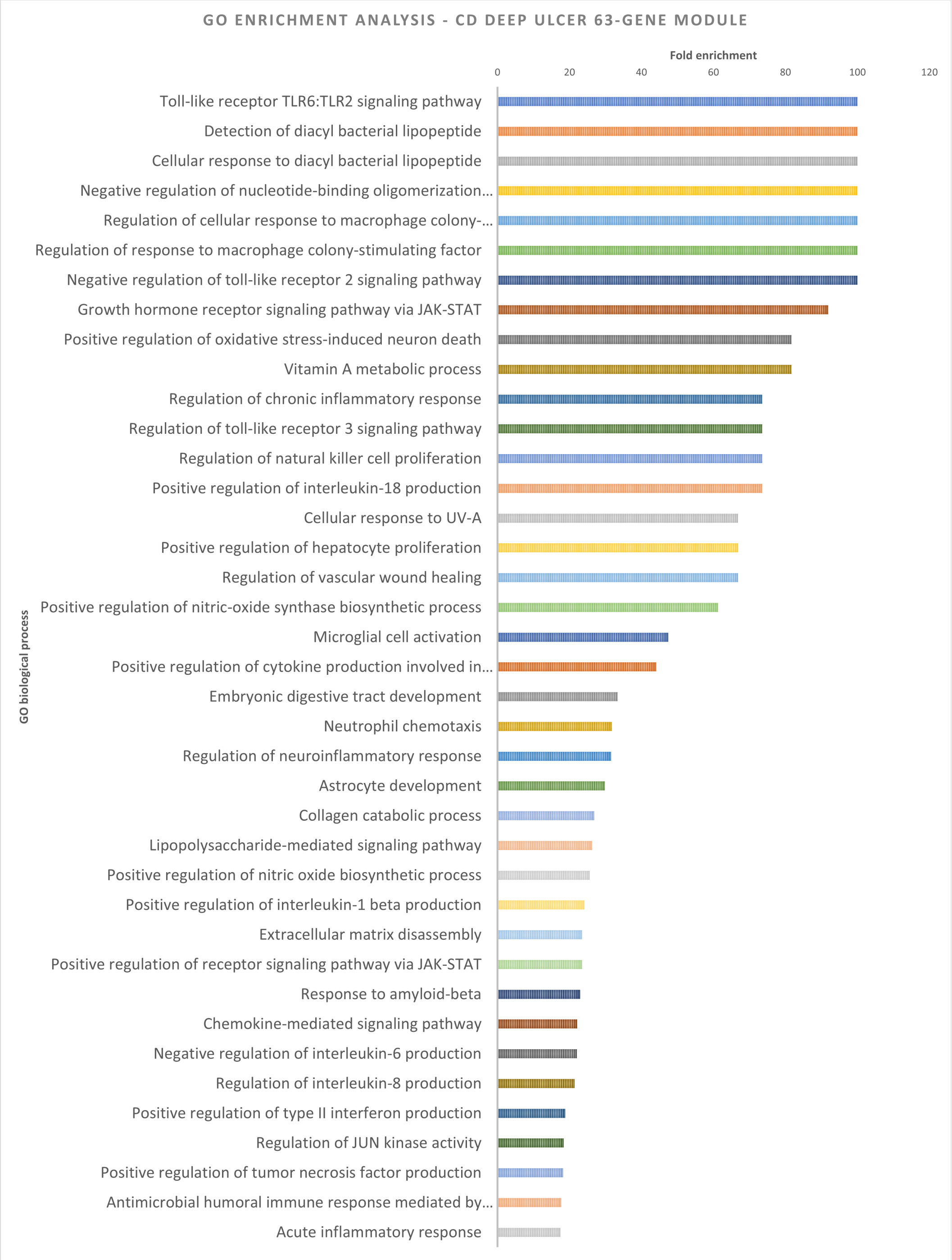


Figure 9: Gene Ontology enrichment analysis. Shown are most enriched biological processes associated with a 63-gene module which was found to be strongly associated with CD with deep ulcer. Results are similar to Figure 8, with the most enriched processes relating to signalling pathways involved in the immune response to pathogens. However, here we also see the detection of slightly different molecules such as bacterial lipopeptides and signalling pathways involving TLR6 (Toll-like receptor 6) and TLR2 (Toll-like receptor 2). These can recognise a wide variety of pathogen-associated molecular patterns (PAMPs) such as lipoproteins and peptidoglycans [47], which extends recognition to Gram-positive bacteria. We also see processes relating to embryonic digestive tract development, extra-cellular matrix disassembly and tumour necrosis factor production, most of which are well-established in IBD [48, 49]. This suggests that the smaller module represents the additional and extended routes to disease symptoms.

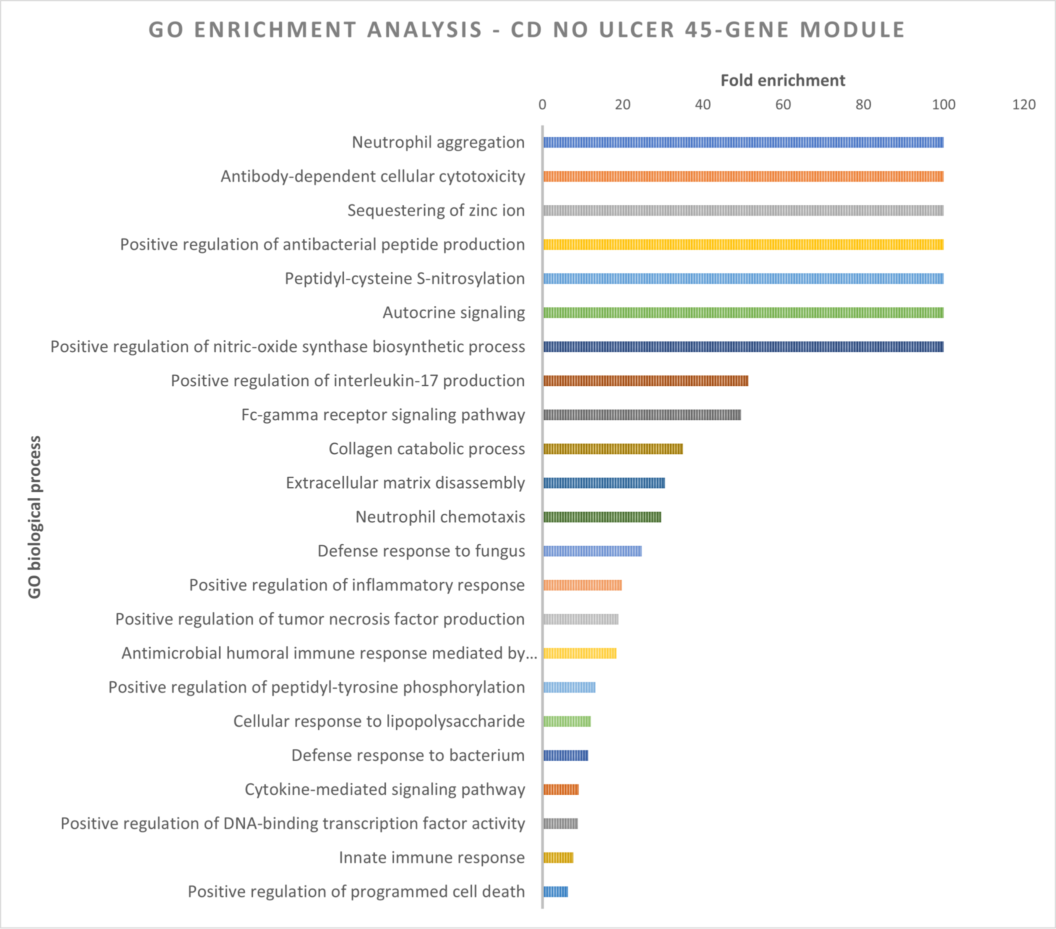
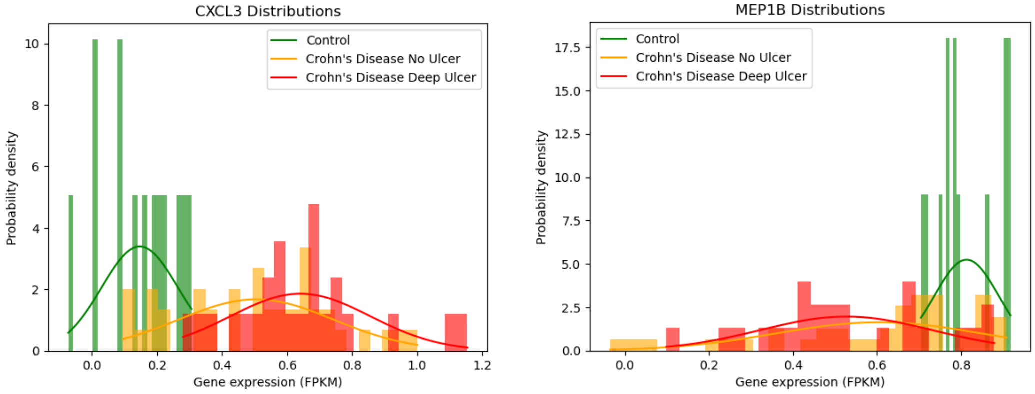


Figure 10: Gene Ontology enrichment analysis. Shown are most enriched biological processes associated with a 45-gene module which was found to be strongly associated with CD without deep ulcer. We obtain similar results to Figures 8 and 9, such as neutrophil aggregation, with additional processes like autocrine signalling, immune response to fungus and use of the fc-gamma receptor signalling pathway. This suggests that the associations of “CD no ulcer” are more wide-ranging than “CD deep ulcer”; for example, fc-gamma receptors can recognise many different types of immunoglobulins [50].



(a) CXCL3 expression

(b) MEP1B expression

Figure 11: Gene expression distributions of CXCL3 (a) and MEP1B (b) across patients with CD deep ulcer, CD no ulcer and control. The gene expression can be approximated by normal distributions. Data from the RISK dataset [21, 22].

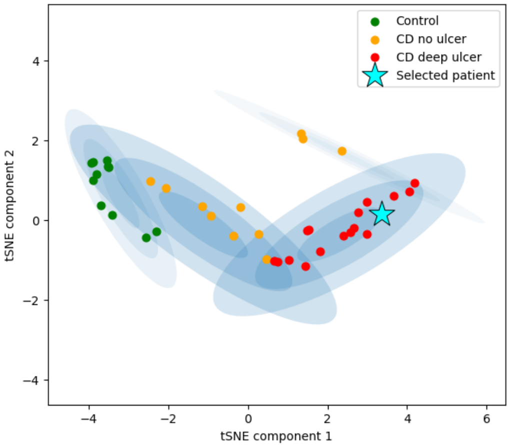
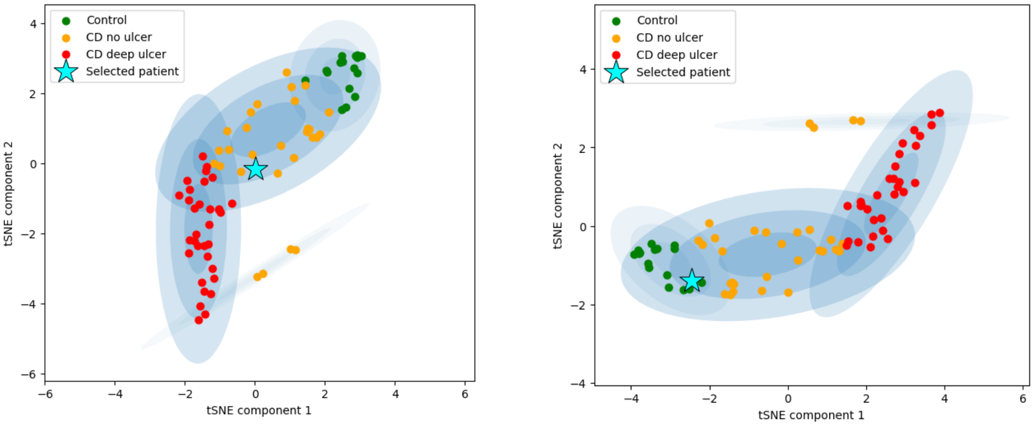


Figure 12: Initial position of Patient 46 (CD deep ulcer) within the clustering model.



(a) 117-gene module alteration

(b) 117-gene and 63-gene module alteration

Figure 13: Visual explanations of the effect of modules on disease in a patient. In comparison to Figure 12, we show the position of Patient 46 within clustering model after modifying the 117-gene module (a) and both 117-gene and 63-gene modules (b), which together were found to account for all positive contribution to CD deep ulcer predictions (Fig. 7). In (a) modifying the 117 genes using the class-contrastive technique results in Patient 46 (with CD deep ulcer [a severe form of the disease], Figure 12) being assigned to the "CD no ulcer" cluster [a less severe form of the disease]. In (b) modifying both the 117-gene and 63-gene modules using the class-contrastive technique results in Patient 46 moving from the CD deep ulcer cluster to the control cluster. This suggests that these modules may be involved in a severe form of CD that leads to deep ulcers.

Table 3: Validation with an independent cohort: Clustering and classification evaluation results for novel classifiers based on Gaussian Mixture Model (GMM), using autoencoder and PCA dimensionality reduction methods. Results shown for binary classification (controls and all CD patients) and multi-class classification (control, CD no ulcer and CD deep ulcer) of disease subtype.

		Binary (control & CD)		**Multi-class (all labels)**	
		Autoencoder	PCA	Autoencoder	PCA
GMM	Accuracy / %	89.2	83.8	64.9	48.6
	F1-Score / %	93.3	89.3	69.9	54.9
	Silh. score	0.366	0.577	0.385	0.255

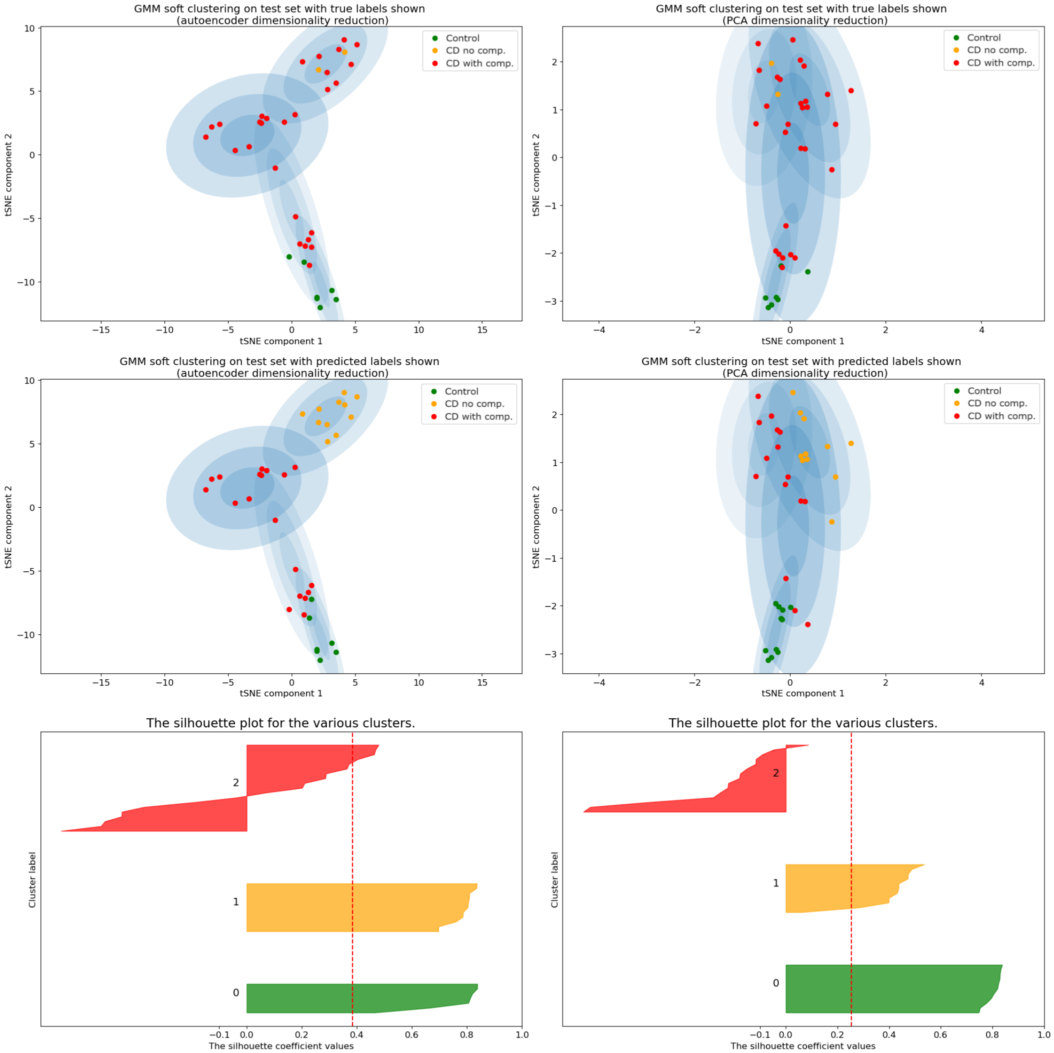


Figure 14: Validation with an independent cohort: Gaussian Mixture Model (GMM) clustering model results after applying dimensionality reduction using autoencoder and tSNE (perplexity=50) (left) and PCA and tSNE (perplexity=150) (right). Deployed on the test set with true labels shown (top third) and predicted labels shown (middle third). Silhouette plots are shown for GMM clusters after applying autoencoder-tSNE (left) and PCA-tSNE (right) methods, with clusters 0, 1 and 2 corresponding to “control”, “CD no complication” and “CD with complication” respectively. These are the disease subtypes and CD with complication is the most severe form of the disease.

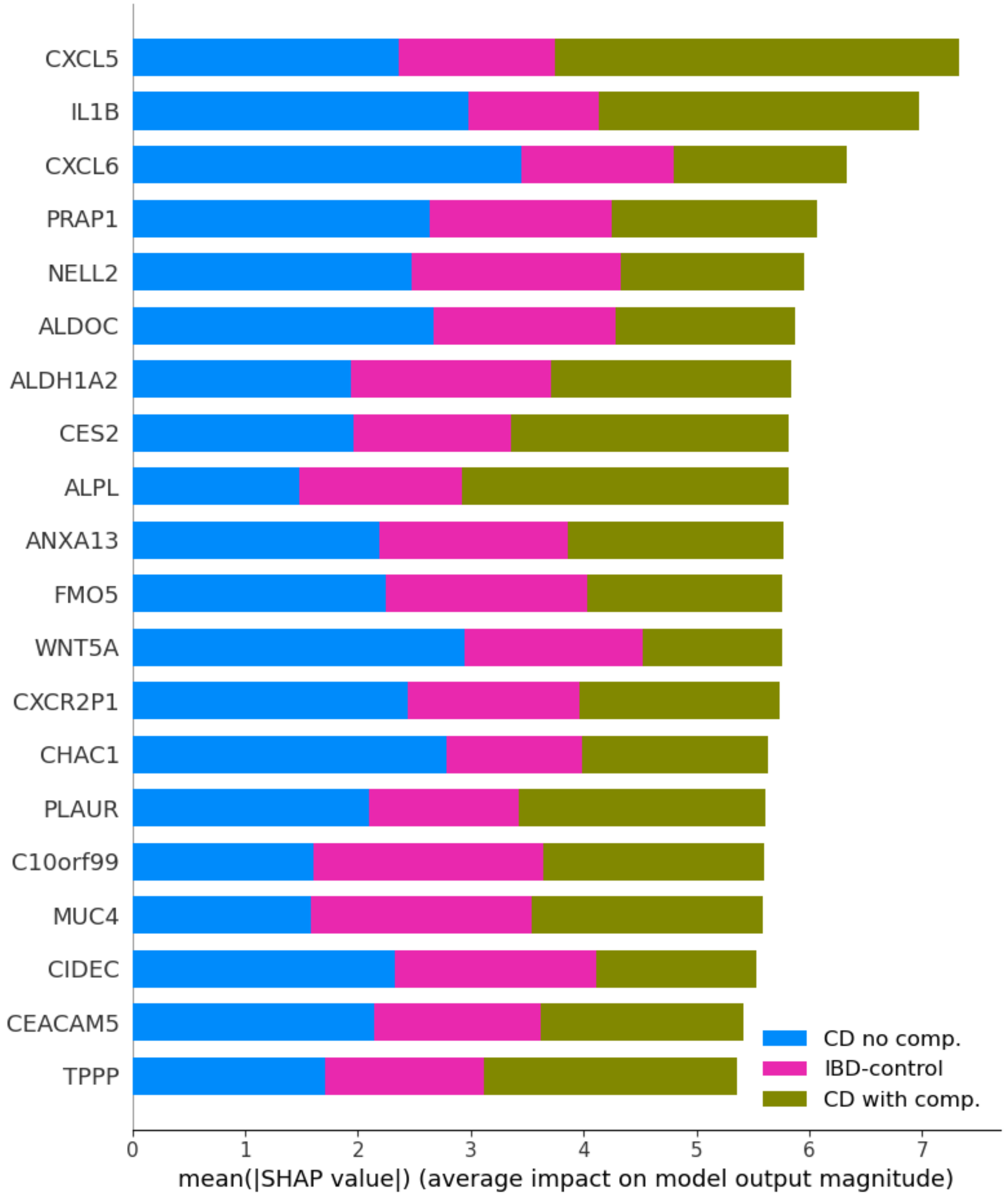


Figure 15: Validation with an independent cohort: Summary plot showing top 20 genes in terms of their average impact on class predictions across all patients. This is for a model which accounts for feature dependence. The blue, pink and green bars depict the magnitude of influence of a gene on the "CD no complication", "control" and "CD with complication" classes respectively. Genes like IL1B, CXCL5 and CXCL6 are within the top 5 and known as susceptibility genes for Crohn's disease.

References

[1] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, January 2009.

[2] Seyed Saeid Seyedian, Forogh Nokhostin, and Mehrdad Dargahi Malamir. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. *Journal of Medicine and Life*, 12(2):113–122, 2019.

[3] Yongxuan Lai, Songyao He, Zhijie Lin, Fan Yang, Qifeng Zhou, and Xiaofang Zhou. An adaptive robust semi-supervised clustering framework using weighted consensus of random k-means ensemble. *IEEE Transactions on Knowledge and Data Engineering*, 33(5):1877–1890, 2021.

[4] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[5] Soumya Banerjee, Pietro Lio, Peter B. Jones, and Rudolf N. Cardinal. A class-contrastive human-interpretable machine learning approach to predict mortality in severe mental illness. *npj Schizophrenia*, 7:60, 12 2021.

[6] Wenkai Han, Yuqi Cheng, Jiayang Chen, Huawen Zhong, Zhihang Hu, Siyuan Chen, Licheng Zong, Liang Hong, Ting-Fung Chan, Irwin King, Xin Gao, and Yu Li. Self-supervised contrastive learning for integrative single cell RNA-seq data analysis. *Briefings in Bioinformatics*, 23(5), 09 2022. bbac377.

[7] Melvyn Yap, Rebecca L. Johnston, Helena Foley, Samual MacDonald, Olga Kondrashova, Khoa A. Tran, Katia Nones, Lambros T. Koufariotis, Cameron Bean, John V. Pearson, Maciej Trzaskowski, and Nicola Waddell. Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Scientific Reports*, 11(1):2641, January 2021.

[8] Jin Hayakawa, Tomohisa Seki, Yoshimasa Kawazoe, and Kazuhiko Ohe. Pathway importance by graph convolutional network and shapley additive explanations in gene expression phenotype of diffuse large b-cell lymphoma. *PLOS ONE*, 17:e0269570, 6 2022.

[9] Yang Yu, Pathum Kossinna, Wenyuan Liao, and Qingrun Zhang. Explainable autoencoder-based representation learning for gene expression data. 12 2021.

[10] M. Pavageau, L. Rebaud, D. Morel, S. Christodoulidis, E. Deutsch, C. Massard, H. Vanacker, and L. Verlingue. DeepOS: pan-cancer prognosis estimation from RNA-sequencing data. preprint, Oncology, July 2021.

[11] Abdul Karim, Zheng Su, Phillip K. West, Matthew Keon, The NYGC ALS Consortium, Jannah Shamsani, Samuel Brennan, Ted Wong, Ognjen Milicevic, Guus Teunisse, Hima Nikafshan Rad, and Abdul Sattar. Molecular classification and interpretation of amyotrophic lateral sclerosis using deep convolution neural networks and shapley values. *Genes*, 12(11), 2021.

[12] Scott Lundberg. Api reference: Core explainers, 2018.

[13] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology*, 2, 2014.

[14] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature machine intelligence*, 2(1):56–67, January 2020.

[15] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, 2021.

[16] Yan Zhang, Zhengkui Lin, Xiaofeng Lin, Xue Zhang, Qian Zhao, and Yeqing Sun. A gene module identification algorithm and its applications to identify gene modules and key genes of hepatocellular carcinoma. *Scientific Reports*, 11:5517, March 2021.

[17] Heewon Park, Koji Maruhashi, Rui Yamaguchi, Seiya Imoto, and Satoru Miyano. Global gene network exploration based on explainable artificial intelligence approach. *PLoS ONE*, 15(11):e0241508, November 2020.

[18] Xiao Ye, Yulin Wu, Jiangsheng Pi, Hong Li, Bo Liu, Yadong Wang, and Junyi Li. Deepgmd: A Graph-Neural-Network-Based Method to Detect Gene Regulator Module. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(6):3366–3373, 2022.

[19] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4:Article17, 2005.

[20] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, December 2008.

[21] Yael Haberman, Timothy L. Tickle, Phillip J. Dexheimer, Mi Ok Kim, Dora Tang, Rebekah Karns, Robert N. Baldassano, Joshua D. Noe, Joel Rosh, James Markowitz, Melvin B. Heyman, Anne M. Griffiths, Wallace V. Crandall, David R. Mack, Susan S. Baker, Curtis Huttenhower, David J. Keljo, Jeffrey S. Hyams, Subra Kugathasan, Thomas D. Walters, Bruce Aronow, Ramnik J. Xavier, Dirk Gevers, and Lee A. Denson. Erratum: Pediatric crohn disease patients exhibit specific ileal transcriptome and microbiome signature (journal of clinical investigation (2014) 124: 8 (3617-3633) doi: 10.1172/jci75436). *Journal of Clinical Investigation*, 125, 2015.

[22] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(Database issue):D991–995, January 2013.

[23] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, November 2012.

[24] Satyam Kumar. Improve your Model Performance with Auto-Encoders, December 2021.

[25] Srivignesh R. Dimensionality Reduction using AutoEncoders in Python, June 2021.

[26] François Chollet et al. Keras. `https://keras.io`, 2015.

[27] Jacob T. Vanderplas. *Python data science handbook: essential tools for working with data*. O'Reilly Media, Inc, Sebastopol, CA, first edition edition, 2016. OCLC: ocn915498936.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[29] Scikit-learn. Selecting the number of clusters with silhouette analysis on kmeans clustering, 2023.

[30] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values, 2020.

[31] Alexander Strehl and Joydeep Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 01 2002.

[32] Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N. Luu, and Tin Nguyen. Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature Communications*, 12(1):1029, February 2021. Number: 1 Publisher: Nature Publishing Group.

[33] Junlin Xu, Jielin Xu, Yajie Meng, Changcheng Lu, Lijun Cai, Xiangxiang Zeng, Ruth Nussinov, and Feixiong Cheng. Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. *Cell Reports Methods*, 3(1):100382, January 2023.

[34] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018.

[35] Subhash Mehto, Kautilya Kumar Jena, Parej Nath, Swati Chauhan, Srinivasa Prasad Kolapalli, Saroj Kumar Das, Pradyumna Kumar Sahoo, Ashish Jain, Gregory A. Taylor, and Santosh Chauhan. The Crohn's Disease Risk Factor IRGM Limits NLRP3 Inflammasome Activation by Impeding Its Assembly and by Mediating Its Selective Autophagy. *Molecular Cell*, 73(3):429–445.e7, February 2019.

[36] Tariq Ahmad, Sara E Marshall, and Derek Jewell. Genetics of inflammatory bowel disease: The role of the HLA complex. *World Journal of Gastroenterology : WJG*, 12(23):3628–3635, June 2006.

[37] Ludwig Werny, Cynthia Colmorgen, and Christoph Becker-Pauly. Regulation of meprin metalloproteases in mucosal homeostasis. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1869(1):119158, January 2022.

[38] Deepak K. Kadayakkara, Pamela L. Beatty, Michael S. Turner, Jelena M. Janjic, Eric T. Ahrens, and Olivera J. Finn. Inflammation Driven by Overexpression of the Hypoglycosylated Abnormal MUC1 Links Inflammatory Bowel Disease (IBD) and Pancreatitis. *Pancreas*, 39(4):510–515, May 2010.

[39] Lucy C Stewart, Andrew S Day, John Pearson, Murray L Barclay, Richard B Gearry, Rebecca L Roberts, and Robert W Bentley. SLC11A1 polymorphisms in inflammatory bowel disease and Mycobacterium avium subspecies paratuberculosis status. *World Journal of Gastroenterology : WJG*, 16(45):5727–5731, December 2010.

[40] Urko M Marigorta, Lee A Denson, Jeffrey S Hyams, Kajari Mondal, Jarod Prince, Thomas D Walters, Anne Griffiths, Joshua D Noe, Wallace V Crandall, Joel R Rosh, David R Mack, Richard Kellermayer, Melvin B Heyman, Susan S Baker, Michael C Stephens, Robert N Baldassano, James F Markowitz, Mi-Ok Kim, Marla C Dubinsky, Judy Cho, Bruce J Aronow, Subra Kugathasan, and Greg Gibson. Transcriptional risk scores link gwas to eqtls and predict complications in crohn's disease. *Nature Genetics*, 49(10):1517–1521, August 2017.

[41] Łukasz Kopiasz, Katarzyna Dziendzikowska, and Joanna Gromadzka-Ostrowska. Colon expression of chemokines and their receptors depending on the stage of colitis and oat beta-glucan dietary intervention—crohn's disease model study. *International Journal of Molecular Sciences*, 23(3):1406, January 2022.

[42] Megan L. Schaller, Madeline L. Sykes, Joy Mecano, Sumeet Solanki, Wesley Huang, Ryan J. Rebernick, Safa Beydoun, Emily Wang, Amara Bugarin-Lapuz, Yatrik M. Shah, and Scott F. Leiser. Fmo5 plays a sex-specific role in goblet cell maturation and mucus barrier formation. April 2024.

[43] Siri Sæterstad, Ann Elisabet Østvik, Elin Synnøve Røyset, Ingunn Bakke, Arne Kristian Sandvik, and Atle van Beelen Granlund. Profound gene expression changes in the epithelial monolayer of active ulcerative colitis and crohn's disease. *PLOS ONE*, 17(3):e0265189, March 2022.

[44] Negar Taheri, Egan L. Choi, Vy Truong Thuy Nguyen, Abhishek Chandra, and Yujiro Hayashi. Wnt signaling in the gastrointestinal tract in health and disease. *Physiologia*, 3(1):86–97, February 2023.

[45] Sebastian Porsdam Mann, Julian Savulescu, and Barbara J. Sahakian. Facilitating the ethical use of health data for the benefit of society: electronic health records, consent and the duty of easy rescue. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083):20160130, December 2016.

[46] Soumya Banerjee, Phil Alsop, Linda Jones, and Rudolf N. Cardinal. Patient and public involvement to build trust in artificial intelligence: A framework, tools, and case studies. *Patterns*, 3(6):100506, June 2022.

[47] Takumi Kawasaki and Taro Kawai. Toll-Like Receptor Signaling Pathways. *Frontiers in Immunology*, 5, 2014.

[48] Alicja Derkacz, Paweł Olczyk, Krystyna Olczyk, and Katarzyna Komosinska-Vassev. The Role of Extracellular Matrix Components in Inflammatory Bowel Diseases. *Journal of Clinical Medicine*, 10(5):1122, March 2021.

[49] Cristiano Pagnini and Fabio Cominelli. Tumor Necrosis Factor's Pathway in Crohn's Disease: Potential for Intervention. *International Journal of Molecular Sciences*, 22(19):10273, September 2021.

[50] Fabian Junker, John Gordon, and Omar Qureshi. Fc Gamma Receptors and Their Role in Antigen Uptake, Presentation, and T Cell Activation. *Frontiers in Immunology*, 11, 2020.