



Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

A new computational phenotyping framework for the clinical characterization of pediatric celiac disease

Giuseppe Albi^a, Valentina Brembilla^c, Erika Lenzi^b, Simone Maffioletti^b,
Emanuele Medolago^c, Chiara Sirtoli^b, Antonio Ferramosca^b, Marco Vincenzo Lenti^{d,e},
Antonio Di Sabatino^{d,e}, Arianna Dagliati^a, Daniele Pala^{b,*}

^a Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy^b Department of Management, Information and Production Engineering, University of Bergamo, Italy^c Department of Engineering and Applied Sciences, University of Bergamo, Italy^d Department of Internal Medicine and Medical Therapeutics, University of Pavia, Pavia, Italy^e First Department of Internal Medicine, San Matteo Hospital Foundation, Pavia, Italy

ARTICLE INFO

Keywords:

Computational phenotyping
Pediatric celiac disease
Topological data analysis
TDA mapper

ABSTRACT

Background and Objective: Pediatric celiac disease (CD) is characterized by marked clinical and immunological heterogeneity that is not fully captured by the current symptom-based Oslo classification. Data-driven computational phenotyping may enable the identification of more homogeneous patient subgroups by integrating multimodal clinical information. This study proposes a topological data analysis (TDA)-based computational framework (pheTDA) to identify and characterize novel sub-phenotypes of pediatric CD from multicentric clinical data.

Methods: A retrospective multicentric cohort of 2,922 pediatric CD patients described by 34 demographic, clinical, serological, genetic, and histological variables was analyzed. Mixed-type data were processed using Gower distance and mapped through a semi-supervised TDA Mapper pipeline, followed by Louvain community detection. The resulting communities were characterized through statistical testing and supervised machine learning. pheTDA was benchmarked against t-SNE followed by DBSCAN and agglomerative hierarchical clustering (AHC), and sensitivity analyses repeated pheTDA using clinical variables only and diagnostic-test variables only.

Results: pheTDA identified seven relatively balanced sub-phenotypes with distinct combinations of symptoms, serological profiles, histological damage, growth delay, and autoimmune comorbidities. Compared with pheTDA, DBSCAN generated a markedly imbalanced nine-cluster partition, whereas the four-cluster AHC solution was dominated by sex. pheTDA showed a clinically broader effect-size profile, with its leading discriminators spanning clinical, serological, genetic, and histological domains. Clinical-only pheTDA communities were strongly aligned with the symptom-based Oslo categories, whereas diagnostic-test-only communities were more fragmented and reflected specific test combinations. Random Forest classifiers achieved a higher mean AUC for pheTDA communities than for Oslo classes (0.94 vs. 0.88).

Conclusions: Multimodal pheTDA provides an exploratory and reproducible framework for identifying clinically interpretable pediatric CD sub-phenotypes. Benchmarking and feature-domain sensitivity analyses support the value of integrating clinical and diagnostic information while emphasizing that the resulting communities are data-dependent computational representations rather than definitive disease classes.

1. Introduction

Celiac disease (CD) is a chronic immune-mediated disorder caused by the ingestion of gluten, that is a protein complex contained mainly in

cereal grains such as common wheat and barley, characterized by an autoimmune response that causes small bowel inflammation in susceptible individuals [1]. The global CD prevalence is around 1%, with some geographical and ethnic variations, and with differences according to

* Corresponding author.

E-mail address: daniele.pala@unibg.it (D. Pala).<https://doi.org/10.1016/j.cmpb.2026.109648>

Received 19 February 2026; Received in revised form 10 September 2026; Accepted 11 September 2026

Available online 17 September 2026

0169-2607/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

sex, with a documented higher prevalence in women (1.5 times the prevalence in men). CD can develop at any age, but its prevalence in children is two times higher than in adults [2]. While the awareness of clinicians and the widespread availability of diagnostic tests are increasing, the worldwide prevalence of CD is increasing, with approximately 95% of celiac patients remaining undiagnosed [3].

The clinical presentation of CD is very wide, and generally labelled using the Oslo definitions [4]. This categorization provides the definition of asymptomatic CD, as individuals who do not present symptoms even in response to direct questioning at initial diagnosis; classical CD, as individuals who presents signs and symptoms of malabsorption; non-classical CD, as individuals with CD who do not present signs and symptoms of malabsorption; subclinical CD, when the CD is below the threshold of clinical detection. However, this broad categorization fails in fully depicting CD phenotypes, or in capturing the risk of complications, as documented by the scientific literature which reports that classical CD patients are more prone to the development of complications with respect to the asymptomatic and the subclinical CD [5], but no significant differences are reported between the other categories. In addition, the clinical manifestation of CD in pediatric patients is heterogeneous and varies by age, e.g. symptoms in infants and toddlers are different than those experienced by older children, and usually occur between 4 and 24 months of age after introducing gluten in their diets, preceded by a possible latent period that could complicate the diagnosis [6].

For pediatric patients, as reported by the European Society Paediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN) Guidelines, the CD diagnosis follows different steps comprising serological tests and the histological evaluation of the intestinal tissue [7]. This former is performed through a biopsy, that is generally evaluated in the clinical practice with two classifications systems: the Marsh classification [8], modified by Oberhuber et al. [9], is the standard in the ESPGHAN guidelines, however, it is criticized by some clinicians for being complex and subjective; and the Corazza and Villanacci approach, which is often preferred since simpler [10]. As evident, the steps performed during the CD diagnosis inevitably comprise the evaluation of different kind of data, and recent works have shown that data-driven methods can be leveraged to redefine the phenotypes of CD, revealing partial overlaps with the Oslo classification and suggesting that simply grouping symptoms is not sufficient to consider the broader spectrum of CD [11].

Computational phenotyping can be defined as a data-driven process in which sub-groups of samples with similar characteristics are identified from a population. The application of computational phenotyping in clinical setting characterized by the use of large repository of structured and unstructured patients data, i.e. Electronic Health Record (EHR), is referred to as *electronic phenotyping* [12]. Halpern et al. [13] conceptualize a learning framework with the underlying hypothesis that patients are characterized by latent variables that cannot be observed directly from the EHR, but for which the analysis of EHR can provide meaningful information. This is possible by leveraging pre-existing domain knowledge in the definition of anchoring observations, called *anchors*, that represent a noisy but meaningful view of the underlying latent variables, which can be used to guide the computational definition of novel phenotype of interest [14].

Topological Data Analysis (TDA) is a branch of mathematics that extracts information and characterizes the data by using quantitative geometric features [15]. One of the two main TDA frameworks, the Mapper algorithm, generates a graph that characterizes and summarizes the topology of the dataset, providing a tool for data visualization and exploration [16]. In the biomedical field, Mapper has found large usage in the task of computational phenotyping [17], specifically, considering identifying subgroups of patients within a population and affected by the same disease or condition of interest, such as Type 2 Diabetes Mellitus (T2DM) [18], aortic stenosis [19] and asthma [20], but with different characteristics. Recently, publicly available tools relying on

TDA Mapper are proposed to specifically address the computational definition of novel phenotypes, starting from an initial clinical definition to anchor the observation to the Mapper graph [21,22].

In this work, we apply the TDA-based pheTDA framework to a large multicentric cohort of pediatric CD patients to identify and characterize multimodal computational sub-phenotypes. The resulting stratification is evaluated through statistical characterization and supervised machine learning, benchmarked against conventional clustering pipelines, and examined through sensitivity analyses based on clinical-only and diagnostic-test-only feature domains. The Oslo classification is retained as a clinical reference rather than as a ground-truth partition.

2. Celiac disease diagnosis according to the ESPGHAN guidelines

Pediatric CD diagnosis integrates serological testing, genetic susceptibility, and histological assessment, but intestinal biopsy is not performed in all patients. This produces a heterogeneous dataset in which unavailable measurements may reflect the diagnostic pathway rather than random missingness. A detailed description of the Marsh-Oberhuber and Corazza-Villanacci classifications, the ESPGHAN diagnostic pathway, and the corresponding flowchart is provided in Supplementary Section 5.1 and Supplementary Fig. 13.

3. Materials and methods

The initial multicentric dataset belongs to a previous study focusing on diagnostic delay in pediatric CD and comprises $n = 3,241$ pediatric patients from $n = 9$ Italian hospitals, diagnosed between 2010 and 2022; the details of the dataset as well as the ethical considerations have already been reported in the previous paper [23]. The data had been completely anonymised after previous publication. Differently, in the current work we only present a sub-analysis addressing the identification of novel pediatric CD sub-groups, starting from an initial CD classification. This is provided by the clinicians with the labels ‘Major’, ‘Minor’ and ‘Silent’, that can be mapped to the Oslo classes ‘Classical’, ‘Non-classical’, and ‘Subclinical’, respectively, in accordance with the existing literature [1]. For each patient different types of features are available, such as symptoms, demographic data and laboratory tests, and immunological, histological and genetic data (see Bianchi et al [23] for the complete list of demographic and clinical variables). We performed data pre-processing following the clinicians suggestions about the variables to include, and missing data imputing according to the CD diagnostic process (Section 3.1). The pheTDA framework was then used to discover seven novel CD sub-phenotypes for which we performed a clinical characterization with TDA enrichment and statistical tests (Section 3.2). Lastly, we evaluated the identified sub-phenotypes clinically with feature selection methods and supervised ML classification models, trained in one-vs rest binary classification tasks as many as the number of the new sub-phenotypes, and comparing their classification performance with the ones obtained considering the initial Oslo classification (Section 3.4). The complete pipeline of the study is reported in Fig. 1. Additional benchmarking against conventional clustering and feature-domain sensitivity analyses are described in Sections 3.3 and 3.5.

3.1. Data pre-processing and dataset consolidation

Growth-related variables were initially evaluated against age- and sex-specific WHO child-growth standards. This assessment was used to derive the Growth Delay variable and to address its missing values before the phenotyping analysis. The complete procedure and the corresponding height and weight boxplots are reported in Supplementary Section 5.2 and Supplementary Fig. 14.

Besides growth delay, other variables related to the patients’ initial cohort presented missing values. Specifically, for the variables representing serological exams and histological classification we pre-process

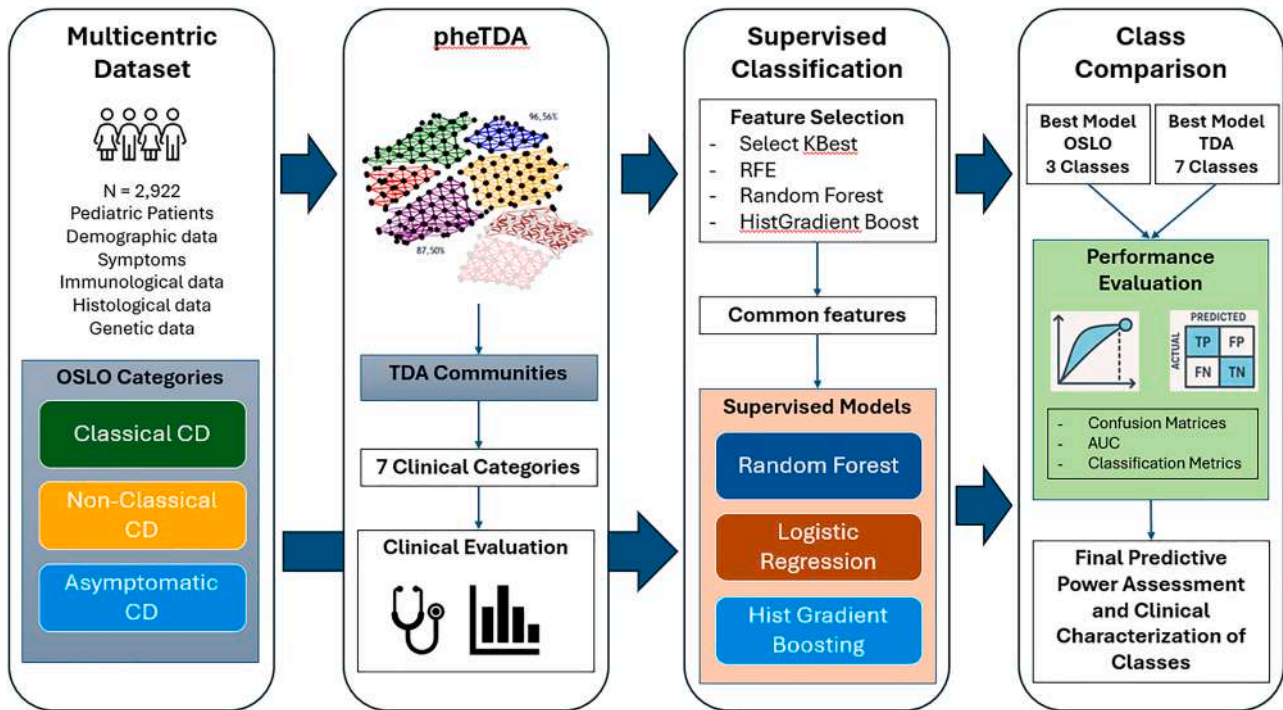


Fig. 1. Complete pipeline of the study: the starting point is a multicentric dataset of pediatric patients for which different variables and the initial Oslo categories are available [23]. Then pheTDA is used to discover communities that are first clinically characterized, and then evaluated by the use of feature selection strategies and ML models, that are trained in as many supervised one-vs-rest binary classification tasks as the number of the new sub-phenotypes. The obtained classification performances are compared with the ones obtained considering the initial Oslo classes.

the variables as follows.

Serological exams. We first encoded the IgA variable with categorical values. Specifically, a new variable called IgA was created, with the following levels: “negative” indicates that the IgA value is $< 20 \text{ mg/dL}$ or not detectable if the patient has a selective deficit of IgA are expressed by a specific variable; “normal” indicates that the IgA values are between 20 mg/dL and ten times the URL; “high” indicates that the IgA values are greater than ten times the URL; “not classified” (NC) if the IgA presents missing value (32.37% of the initial values) and the Corazza histological score is available. Secondly, we encoded the EMA levels in “positive” and “negative”, if the EMA test was positive or negative, respectively; in addition, we added the level “not performed” (NP) if EMA had missing values (30.06% of the initial values) and IgA was “normal” or “negative”, since it is reasonable to assume that the biopsy is directly performed, whereas if IgA was “high”, the level “not available” (NA) was used to impute the EMA variable. Finally, a binary variable for the human leukocyte antigen (HLA) was created by looking at the presence of the D2Q and DQ8 haplotypes, and specifically, if at least one of the two is present, then the HLA variable has value 1, 0 otherwise.

Corazza histological score. In the dataset both the Marsh-Oberhuber and the Corazza classifications were provided for the histological scores. We chose to use the Corazza classification since this is often preferred as it considers a lower number of categories, and previous work reported that its adoption leads to a higher interobserver agreement within pathologists [24]. In addition to the levels already present in the considered dataset, we created two new levels: (i) “not performed” (NP), which represents the case in which missing values are present for the Corazza classification (27.68% of the initial values) and duodenum endoscopy is not performed, or IgA is “high” and simultaneously EMA is “positive”, or HLA is “positive”; and (ii) “not available” (NA), when missing values are present and none of the previous condition is verified.

We discarded the patients for which sex was not registered ($n = 1$), IgA or EMA was not performed ($n = 13$), almost all variables were

missing ($n = 45$), and the initial CD classification was not provided ($n = 260$). We obtained a dataset of $n = 2,922$ pediatric patients, and 34 clinical variables. In Table 1 we report for each of the preprocessed variables, and by grouping patients according to the initial CD classification, the mean and the standard deviation, and the frequency and the percentage, for numerical and categorical variables, respectively, for the dataset given in input to the pheTDA framework. For a complete list of the variables including demographic and other clinical data we refer to Bianchi et al. [23].

Table 1

Patient characteristics pre-processed and presented as number and percentage for categorical variables and mean $\pm$ standard deviation for numerical variables, grouped by patients initial diagnosis. For the complete list of variables see Bianchi et al. [23].

Variable	Major ($n = 1500$)	Minor ($n = 928$)	Silent ($n = 494$)
Growth delay	620 (41.3)	247 (26.6)	39 (7.9)
Human leukocyte antigen (HLA)	1077 (71.8)	737 (79.4)	364 (73.7)
Corazza histology:			
A	34 (2.3)	22 (2.4)	13 (2.6)
B1	490 (32.7)	374 (0.3)	227 (6.0)
B2	515 (34.3)	359 (8.7)	170 (4.4)
Not performed (NP)	418 (27.9)	151 (6.3)	72 (14.6)
Not available (NA)	43 (2.9)	22 (2.4)	12 (2.4)
Immunoglobulin (IgA):			
Negative	120 (8.0)	74 (8.0)	54 (10.9)
Normal	197 (13.1)	117 (2.6)	80 (16.2)
High	779 (51.9)	375 (0.4)	232 (7.0)
Not classified (NC)	404 (26.9)	362 (9.0)	128 (5.9)
Anti-endomysium antibodies (EMA):			
Negative	65 (4.3)	25 (2.7)	19 (3.8)
Positive	1016 (7.7)	522 (6.2)	343 (69.4)
NP	10 (0.7)	18 (1.9)	6 (1.2)
Not classified (NC)	409 (27.3)	363 (9.1)	126 (5.5)

The most prevalent CD classification in the considered patients' population is Major ($n = 1500$ patients of which 62.9% female), followed by Minor ($n = 928$ patients), and Silent ($n = 494$ patients). We note that the Silent category comprises a higher proportion of patients with an EMA positive test and that patients initially classified as Minor have the highest proportion of Corazza's histology classifications B1 and B2 but also of NP (32.7%, 34.3% and 27.9%, respectively), and have the highest proportion of IgA's high and NC levels (51.9% and 21.6%, respectively).

3.2. Computational phenotyping with pheTDA

In this work, the aim is to identify novel sub-phenotypes of CD in pediatric patients. To this end, and we leverage a TDA-based framework called pheTDA comprising two main steps: firstly, the Mapper algorithm is employed in a semi-supervised way, guiding the creation of the topological graph from the input dataset by leveraging the initial phenotypic definition as anchoring, then the second steps of computational phenotyping introduces a new stratification in the population, while characterizing the new subgroups according to the most discriminative features with predictive models.

Following the schematic definition of pheTDA reported in Fig. 2, the framework starts by computing the distance between the n samples of the dataset X in input to create a $n \times n$ distance matrix D , by using a distance function suitable for the types of the dataset's variables. In our case the input dataset consists of both numerical, i.e. age, and categorical variables, and we select the Gower's distance [25]. For two samples i and j having F variables, the similarity between them is defined as:

$$S_{ij} = \frac{\sum_{k=1}^F w_{ijk} s_{ijk}}{\sum_{k=1}^F w_{ijk}}, \quad (1)$$

where w_{ijk} are non-negative weights typically set to 1, and s_{ijk} is the similarity between two data points considering the variable k -th. When dealing with binary or ordinal variables, the similarity is equal to 1 in case the samples have the same value, otherwise the similarity is 0, while when dealing with continuous variables, which have a range of values R_k , the similarity is computed as:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}. \quad (2)$$

The Gower's distance can be computed straightforwardly from the Eq. (1), as $D_{ij} = 1 - S_{ij}$. Note that before computing the distance, continuous variables are scaled in the $[0, 1]$ range.

In the following step, pheTDA applies TDA Mapper on the distance matrix. The algorithm first projects the distances in a lower-dimensional space, here in R^2 , by the means of a *lens* function f , e.g. principal component analysis (PCA), second divide the projection space with an *open cover* U by using a series of overlapping bins $u_1, u_2, \dots, u_r$. Third, it applies a clustering method for each of the subset of samples that in the lower dimensional space fall in correspondence to an intersection of bins. Finally, Mapper create a graph $G = (V, \epsilon)$, where $V = \{v_1, v_2, \dots, v_N\}$ is the set of nodes that are identified as the cluster obtained from the previous step, and $\epsilon = \{\epsilon_1, \epsilon_2, \dots, \epsilon_M\}$ is the set of edges that indicates the intersection between them, i.e. an edge exists between two nodes if they share at least one data point. This algorithm comprises different hyperparameters: *resolution* (r), i.e. the number of bins to use for each dimension of the lower projection space, *gain* (g), i.e. the percentage of overlap between the bins, the clustering method to choose, and eventual hyperparameters related to the lens function. pheTDA guides the choice of these hyperparameters with a two-step grid search. In the first step, the values of the Mapper hyperparameters, namely resolution, gain and cluster method are kept fixed, and the framework selects the lens and its hyperparameters that minimize the graph entropy $H(G)$, defined as:

$$H(G) = \frac{\sum_{i=1}^N H(v_i) |v_i|}{\sum_{i=1}^N |v_i|}, \quad (3)$$

in which node entropy $H(v_i)$ is computed for each node v_i of size $|v_i|$, with respect to the proportion of the initial phenotypic label $P(v_i)$ as:

$$H(v_i) = - \sum_{i=1}^{\phi} P(v_i) * \log_2 P(v_i). \quad (4)$$

In the second step, the Mapper hyperparameters and the clustering methods are selected by visually evaluating statistics computed by pheTDA on the graph created.

After the identification of the best hyperparameters, pheTDA applies a community detection algorithm, i.e. a clustering algorithm that works on a graph, to partition the resulting topological graph and identify communities. In particular, we choose the Louvain algorithm, that works similarly to an agglomerative clustering by joining pairs of nodes: it starts by assigning every node to its own community, and then for each of them it tries to find the maximum positive *modularity* gain by moving each node to all of its neighbouring communities [26]. Modularity is a graph property that measures the relative density of edges inside communities with respect to edges outside communities [27], computed as follows:

$$Q(G) = \frac{1}{2M} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2M} \right) \delta(c_i c_j), \quad (5)$$

where A is the graph adjacency matrix in which an element $A_{ij} = 1$ if existing an edge ϵ_m between two nodes v_i and v_j , otherwise is equal to 0, k_i and k_j are the *degree* of the node v_i and v_j , respectively, expressing the number of connections in which the nodes are involved, $\delta(c_i c_j) = 1$ if v_i belongs to community c_i , v_j belongs to community c_j , and $c_i = c_j$, i.e. the nodes belong to the same community, otherwise $\delta(c_i c_j) = 0$.

Subsequently, each patient is assigned to the community of the nodes in which he/she appears, introducing a new stratification ϕ' in the population cohort. Note that one patient may appear in more than one node in the graph, therefore the assignment to a community may be controversial; pheTDA resolves ties by assigning these patients to the community that appeared most frequently in the nodes which they are associated to, or otherwise to the community belonging to the nodes with largest size.

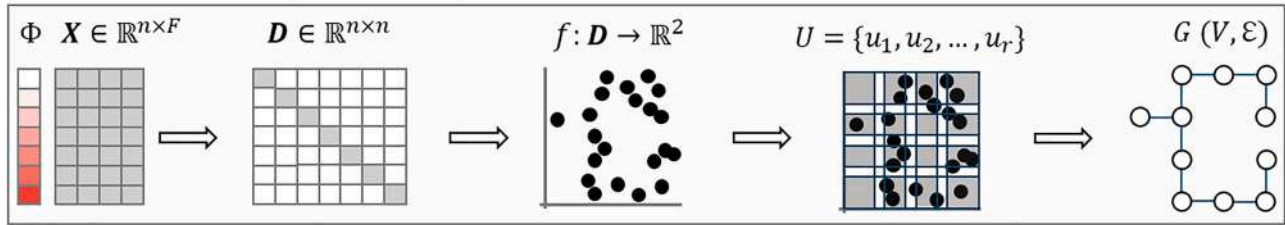
Lastly, predictive modeling is used to characterize the new sub-phenotypes. For each of them, pheTDA trains a machine learning (ML) model - specifically we selected Elastic Net logistic regression [28] - and tunes the hyperparameters with a 5-fold cross-validation in one-vs-rest binary classification tasks, involving the prediction of $P(\phi' = \phi' | X)$, representing the probability to belong to a specific sub-phenotype, given the dataset features. Then, the tool selects the model that achieves the highest Area under the Receiver Operating Characteristic (ROC) Curve (AUC) to extract the most important features $X_{\phi'}$ for each of the new sub-phenotype ϕ' , corresponding to the variables with the highest feature importance. The selected variables are then used by pheTDA to enrich the topological graph, i.e. to colour it according to the distribution of a specific variable. The comprehensive list of lens function available, and the ranges of the hyperparameters evaluated during the grid search, are reported in previous work [21] and in the pheTDA code repository¹.

3.3. Comparison with traditional clustering methods

To compare pheTDA with conventional clustering strategies, DBSCAN and agglomerative hierarchical clustering (AHC) were applied

¹ <https://github.com/Sep905/pheTDA>

Semi-supervised Topological Data Analysis



Computational phenotyping

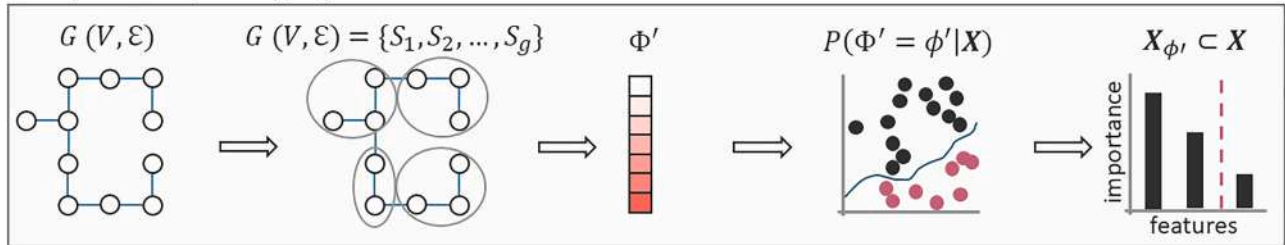


Fig. 2. PheTDA consists of two main components. The first reported on top, is a semi-supervised TDA pipeline that takes as input a tabular patient dataset X with n samples and for each F variable, and a clinical phenotype ϕ . It computes a squared distance matrix D , projects it into $\mathbb{R}^2$ space using a lens function f , covers the projection with overlapping bins, and applies clustering within each bin intersection to create a graph $G = (V, \epsilon)$. The second component, reported in the lower part of the Fig., deals with computational phenotyping, which starts from the topological graph produced by Mapper, and applies a community detection algorithm to partition the graph and to generate a new stratification ϕ' . These subgroups are then characterized by key features extracted from a one-vs-rest predictive model.

to the same preprocessed dataset after t-SNE dimensionality reduction, using the same Gower distance representation. For DBSCAN, the neighborhood parameter was selected with the elbow method and $k = 15$ was retained, yielding nine clusters after exclusion of noise observations. For AHC, the number of clusters (2–10) and linkage criterion (single, complete, or average) were selected by grid search using the Silhouette Coefficient; the optimal solution used complete linkage and four clusters (Silhouette = 0.41). Across all solutions, binary and multilevel categorical variables were compared with Pearson's chi-square test and summarized using bias-corrected Cramér's V , while age was evaluated with the Kruskal–Wallis test and epsilon-squared. P-values were adjusted with the Benjamini–Hochberg procedure. Because the variables used for these comparisons also contributed to cluster derivation, effect sizes were interpreted descriptively rather than as independent validation. Full results are provided in Supplementary Sections 1–3.

3.4. Comparison with the current classification

In order to compare the TDA communities with the current Oslo groups, we applied a supervised learning framework to automatically classify patients according to either classification systems, and we assessed the models' predictive performances for both cases. Several models have been trained in one-vs-rest binary classification, in order to select the best performing ones, after a thorough feature selection process that was performed both to increase the models' performances and to identify the most influential variables for both the Oslo and the TDA classification. In particular, for both classification schemes, i.e. the 3 classes of the Oslo classification and the 7 sub-phenotypes identified with pheTDA, we applied identical ML pipelines, in which numerical features are scaled in the $[0, 1]$ range - to be consistent with the scaling performed when computing the Gower's distance-, and categorical variables are encoded with one-hot encoding, i.e. all the h levels of the variable are kept with the creation of a separate variable for each level, leading to a total of $n = 44$ variables.

Feature selection was performed using four complementary methods implemented in the scikit-learn Python's package [29]:

- **SelectKBest:** a univariate filter-based approach that selects features with the highest statistical relevance measured with the F-statistic;
- **Recursive Feature Elimination (RFE):** a wrapper method that iteratively removes the least important features based on a classifier model performance [30];
- **Random Forest (RF)** [31] features importance, leveraging ensemble-based variable ranking;
- **Histogram-Based Gradient Boosting (HistBoost)** [32] features importance, extracting variable importance as in the RF case.

Each feature selection method was executed twice—once optimizing for overall accuracy and once for the AUC. This process yielded eight feature lists in total. A final consensus feature set was obtained by retaining variables that appeared in at least six of the eight lists, ensuring robust selection of features consistently deemed informative across multiple algorithms and criteria.

Using the selected features, we trained three supervised classification models for each labelling system: RF, Logistic Regression (LR), and HistBoost. Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC. The best-performing algorithm was selected separately for the Oslo-based and TDA-based classifications to ensure a fair comparison. Finally, we compared the predictive performance achieved on the Oslo classes versus the TDA communities. Beyond classification metrics, we also examined the most influential features driving model decisions to identify differences in the underlying clinical determinants of each classification framework. This analysis provided insight into whether the TDA-derived communities captured additional or more nuanced dimensions of disease heterogeneity than the Oslo definitions.

3.5. Sensitivity analysis by feature domain

To assess the sensitivity of the phenotypic structure to the information domain supplied to pheTDA, the complete pipeline was repeated with two restricted feature sets: (i) clinically observable variables, including demographic characteristics, symptoms, and comorbidities; and (ii) diagnostic-test variables, comprising IgA, EMA, HLA, Corazza histology, and immunoglobulin-deficiency indicators. The resulting graphs, communities, and their composition according to the Oslo

categories were compared descriptively with the full multimodal solution. Detailed community profiles are reported in Supplementary Section 4.

4. Results

4.1. pheTDA identifies novel sub-phenotypes of pediatric CD

The lens function and clustering method selected in the pheTDA pipeline were t-SNE [33] and spectral clustering [34], respectively. The optimal values of t-SNE's hyperparameters selected by pheTDA were *perplexity* = 10, *learning rate* = 300, and *number of iterations* = 1500, while the Mapper hyperparameters' values selected were *resolution* = 16 and *gain* = 0.5. This combination allowed us to obtain a connected graph, shown in Fig. 3A in which the overall patient population is represented in $n = 382$ nodes, where each node comprises on average 31 patients, and has a mean node degree of 9.1, i.e. on average, a node is connected to approximately nine other nodes. From the resulting topological graph, the Louvain algorithm identified $\phi' = 7$ communities, whose nodes are coloured in Fig. 3B according to the highest proportion of initial CD classification category. From this Fig. we can observe that the patients classified as Silent CD are more prevalent in communities 3, 5, 6, and 7, situated in the bottom and left part of the graph, and communities 1 and 2 are mainly composed of patients classified as Major. These results can be observed also in the Sankey diagram reported in Fig. 3D, which shows the composition of each community in terms of initial CD classification.

We report the heatmap comparing the Gower's distance matrices computed on the input dataset, and by ordering the patient according to the initial CD classification and with the new stratification ϕ' introduced by pheTDA, respectively, in Fig. 3C. The average pair-wise distance between patients initially classified as Major is 0.209 ± 0.067 , which is higher compared to both the average pair-wise distance of the Minor and Silent groups, with a value of 0.192 ± 0.062 and 0.132 ± 0.05 , respectively. This observation highlights the initial hypothesis that classifying CD patients according exclusively to the clinical manifestation could result in heterogeneous groups. We observe that pheTDA introduces a stratification in which groups contain patients that have a lower average pair-wise distance. In particular, the sub-phenotype 1 ($n = 291$) has a mean pair-wise distance of 0.155 ± 0.059 , the sub-phenotype 2 ($n = 569$) has a mean pair-wise distance of 0.164 ± 0.059 , the sub-phenotype 3 ($n = 457$) has a mean pair-wise distance of 0.139 ± 0.054 , the sub-phenotype 4 ($n = 194$) has a mean pair-wise distance of 0.132 ± 0.051 , the sub-phenotype 5 ($n = 528$) has a mean pair-wise distance of 0.152 ± 0.054 , the sub-phenotype 6 ($n = 392$) has a mean pair-wise distance of 0.141 ± 0.06 , and the sub-phenotype 7 ($n = 491$) has a mean pair-wise distance of 0.163 ± 0.061 . We also note from the heatmap that the new stratification introduced by pheTDA leads to a lower inter-cluster distance, considering that TDA Mapper is able to discover related groups providing better insights on how individual samples relate to the whole population.

Enrichment plots are reported in Fig. 4, and show the distribution of the most important variables, extracted by the trained models, and in

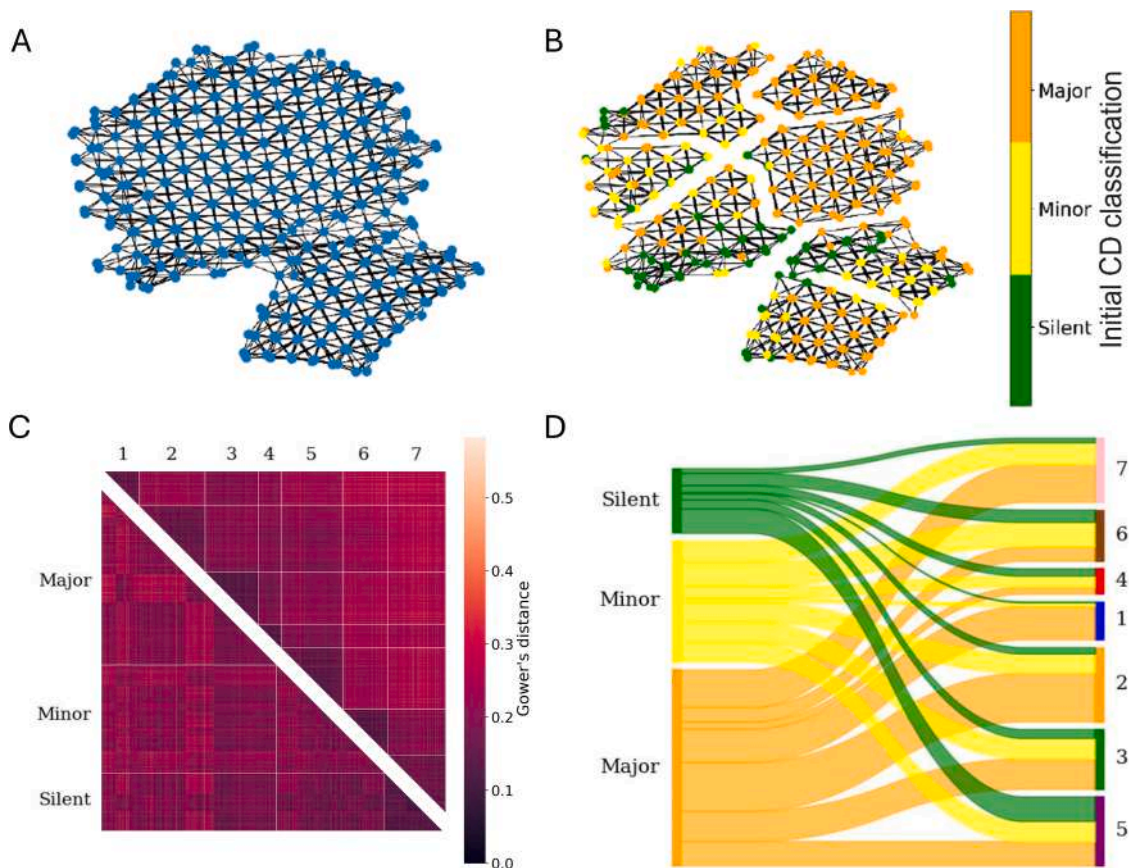


Fig. 3. Topological representation of the input dataset obtained with pheTDA. (A) The pheTDA framework selects *t-SNE* as lens function, the Mapper hyperparameters value and spectral clustering to obtain a topological representation of the input dataset as a graph. (B) Topological graph in which the identified $n = 7$ communities are drawn disconnected, and nodes are coloured according to the highest proportion of initial CD classification category. (C) Heatmap showing Gower's distance matrices between pair of patients and computed from the input dataset, in which patients are ordered by initial CD classification in the bottom triangular matrix, and by the new stratification introduced by pheTDA in the upper triangular matrix. (D) Sankey diagram showing the composition of the new stratification, i.e. the identified CD sub-phenotypes, according to the initial CD classification.

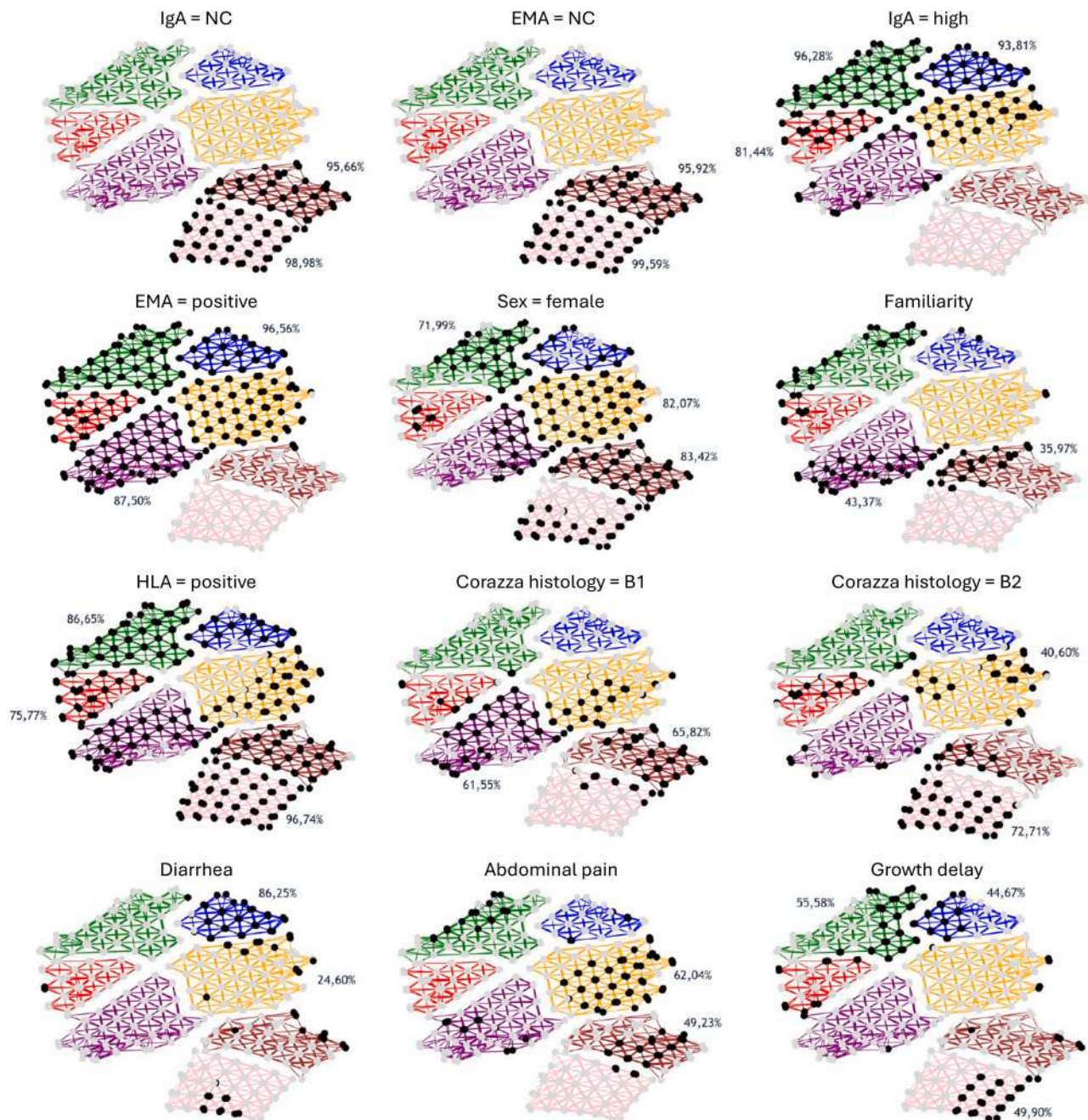


Fig. 4. Enrichment plots showing the distribution of selected variables in the nodes of the Mapper graph. We report the variables extracted from the optimal logistic regression model that are selected as the most important for each sub-phenotypes, and shared across them. For each of these variables, we colour the graph's node by black, if the variable is present in the node, otherwise we colour the node by grey. Edges are coloured according to the sub-phenotypes colours, as reported in the Sankey diagram.

common between the new sub-phenotypes. In the topological graphs reported, nodes are coloured in black if the patients that belong to them present the value of the feature of interest, in grey otherwise. Differently, edges are coloured according to the communities' colours reported by the Sankey diagram (Fig. 5D). Additionally, each enrichment plot reports in the title the name and the value of the variables of interest. Considering the IgA and EMA serological tests, we observe that these have a strong impact on the definition of the new stratification, since they split the patients into two well-distinct regions, spatially positioned on the top and bottom portions of the plot. Indeed, the sub-phenotypes 6 and 7 are prevalently composed of patients with both values of IgA and EMA that is equal to "NC" (6 has percentage of IgA = NC and EMA = NC of 95.66 % and 95.92%, respectively, while 7 has

percentage of IgA = NC and EMA = NC of 98.98 % and 99.59%, respectively). Sub-phenotypes 1, 3 and 4 are prevalently composed of patients with IgA = "high" (93.81%, 96.28% and 81.44%, respectively), and considering also the sub-phenotypes 2 and 5, these are mainly made by patients with EMA = "positive". Differently, patients with the variable HLA = "positive" are present in all the sub-phenotypes, but more prevalent in sub-phenotype 3 (86.65 %), 4 (75.77%) and 7 (96.74%).

When assessing the enrichment plots by focusing on other variables, we observe that these are less impactful in the global definition of the patients' groups, but their combination locally defines the new sub-phenotypes. For example, female subjects are spread on the entire network while characterizing the sub-phenotypes 2, 3 and 6 (82.07%, 71.99% and 83.42%, respectively); the prevalence of familiarity is

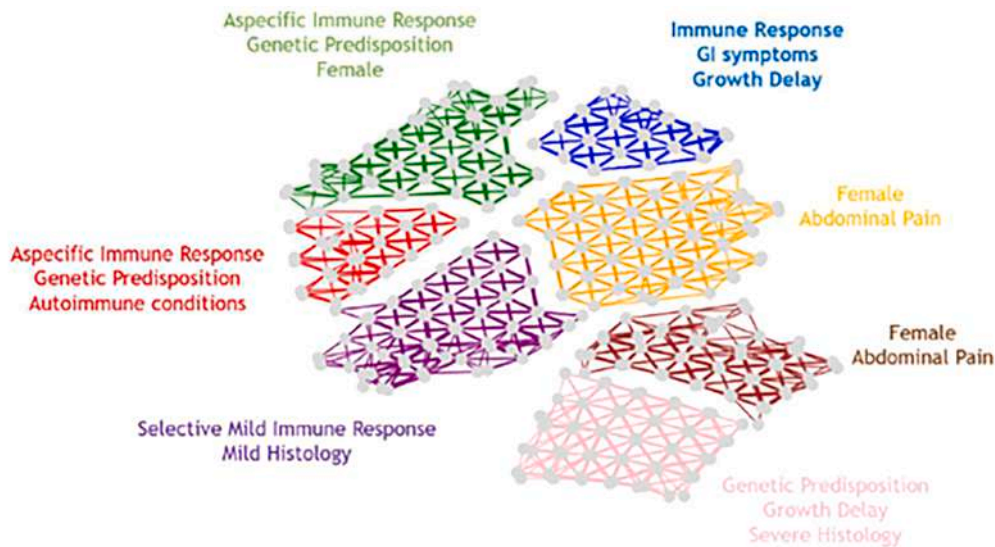


Fig. 5. Enrichment plot with summarized clinical perspective of the sub-phenotypes. Similar to an enrichment plot, we report the topological graph where edges are coloured by the colour assigned to the specific sub-phenotype. For each of them, the clinical considerations that can be taken by looking at the most important variables are reported.

higher in sub-phenotypes 5 and 6 (43.37% and 35.97%, respectively); GI symptoms characterize the sub-phenotype 1, where the 86.25% experiences diarrhea, and the subphenotype 2 has the highest percentage of patients with experienced abdominal pain; growth delay is more frequent in sub-phenotypes 1 (44.67%), 3 (55.58%), and 7 (49.90%). Considering the histological evaluation, patients with Corazza histological class equals to B1 are more concentrated in the sub-phenotypes 5 and 6, with a proportion of 61.55% and 65.82%, respectively, while also present in a lower proportion in 2, 4 and 7; patients with Corazza histological class equals to B2 are more present in the sub-phenotypes 7 (72.71%) and 2 (40.60%). Note that pheTDA is able to discover communities consistently with what suggested by the referring physician, related to Corazza histological class equals to “NP”: these patients are present in the sub-phenotypes 1 and 3, for which the histological evaluation is not performed, since the diagnosis is confirmed by the results of laboratory and genetic testing, i.e. HLA is “positive” and IgA is “high” for the sub-phenotype 3, and EMA is “positive” and IgA is “high” for the sub-phenotype 1.

Lastly, we summarize these findings in a single plot in Fig. 5, showing the characterization of all the sub-phenotypes according to the previous variables, and adding a clinical perspective with the aim of providing a more practical tool for the physicians on case one need of identify a possible sub-phenotypes based on patients characteristics.

For the sub-phenotype 1, in addition to EMA and HLA, several GI symptoms (diarrhea, weight loss, GI upper) are present. A possible interpretation may be that these symptoms can predispose to malabsorption, potentially leading to growth delay. These patients also show notable immune response, as IgA = “High” and EMA = “positive” are present. Sub-phenotypes 2 and 6 are characterized by the presence of female patients and abdominal pain; interestingly, the latter is defined by EMA = “NC” and IgA = “NC”, therefore the two phenotypes may be separated by different levels of immune response or, more probably, they could be merged. Moreover, given the absence of severe symptomatology, the sub-phenotype 6 may correspond to a less severe pathophysiology. The sub-phenotype 2 is characterized by patients with relevant immune response, genetic predisposition and female sex. No particular symptoms are identifiable, but growth delay is frequently present as well as microcytic anaemia. Sub-phenotype 4 is similar to 3, but it is composed mainly by male patients. The presence of other autoimmune diseases is notable, as type 1 diabetes, dermatitis herpetiformis and skin/tissue autoimmune diseases are present. Sub-phenotype

5 is characterized by “normal” or “negative” IgA levels, with “positive” as result for EMA test, therefore a selective immune response is present. Furthermore, histological damage is denoted by the presence of B1 and B2 levels of Corazza classification, with higher proportion of B1 type. Therefore, this phenotype manifests an evident histological alteration. The sub-phenotype 7 is characterized by severe histological damage; more in particular, it distinguishes it from the 5 since this shows a minority of Corazza B2 cases. These severe intestinal damages can likely lead to growth delay, that is another feature that characterizes sub-phenotype 7, where also a genetic predisposition is observed.

4.2. Comparison with traditional clustering methods

DBSCAN identified nine clusters after excluding 24 observations as noise. The solution was markedly imbalanced: cluster 0 included 1,549 patients (53.5%) and cluster 3 included 681 (23.5%), while each remaining cluster represented no more than 7.0% of the analyzed cohort. Its largest effects involved familiarity (Cramér’s $V = 0.569$), EMA (0.557), IgA (0.550), HLA (0.536), and sex (0.424). AHC produced four more balanced clusters (13.1%–32.3% of the cohort), but separation was strongly dominated by sex (Cramér’s $V = 0.802$), followed by EMA (0.559), IgA (0.551), HLA (0.343), and abdominal pain (0.331).

In contrast, the seven pheTDA communities ranged from 6.6% to 19.5% of the cohort. Their five leading discriminators were IgA (Cramér’s $V = 0.641$), diarrhea (0.581), EMA (0.562), HLA (0.417), and Corazza histology (0.417), spanning serological, clinical, genetic, and histological domains. The Oslo categories were primarily differentiated by clinical manifestations, with diarrhea showing the largest effect (0.391). Thus, although the between-group tests are descriptive rather than independent validation, pheTDA yielded a relatively balanced partition with a broader multimodal clinical profile than the two conventional pipelines. Fig. 6 summarizes the strongest effect sizes; complete distributions, heatmaps, and tests are provided in Supplementary Sections 1–3.

4.3. Sensitivity analysis by feature domain

pheTDA identified 10 communities when restricted to clinical variables and 14 communities when restricted to diagnostic-test variables. In the clinical-only analysis, most communities were dominated by a single Oslo category, and the optimized graph entropy was 0.989,

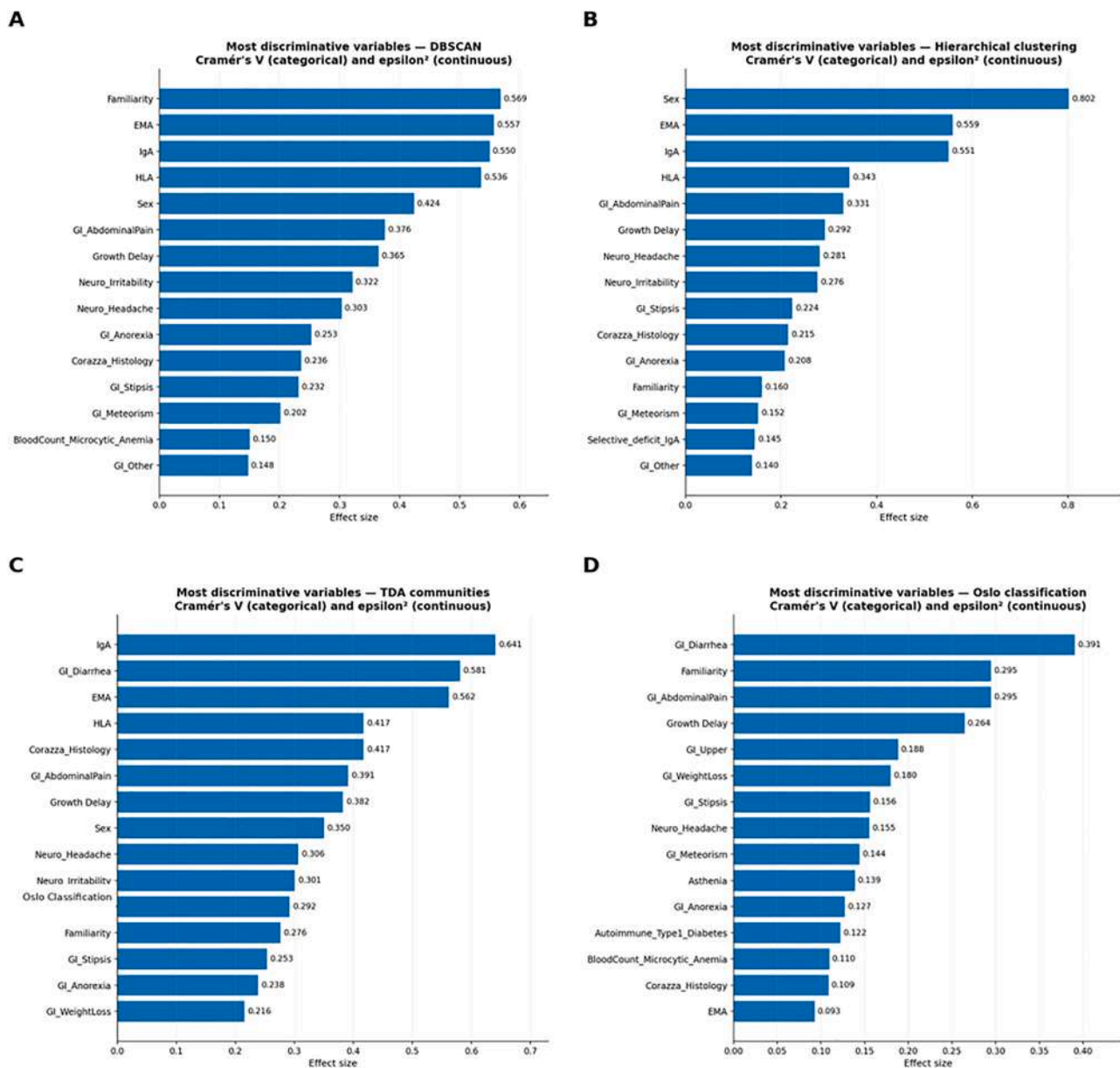


Fig. 6. Variables with the largest effect sizes across the four partitions: (A) DBSCAN, (B) agglomerative hierarchical clustering, (C) pheTDA communities, and (D) Oslo classification. Cramér's V is reported for categorical variables and epsilon-squared for age. These descriptive effect sizes quantify between-group separation within each solution; they do not constitute independent validation because the same variables contributed to cluster construction.

consistent with the visual separation of Major, Minor, and Silent presentations in Fig. 7A. The diagnostic-test-only solution had higher entropy (1.374) and showed substantial mixing of Oslo categories (Fig. 7B); its communities were chiefly defined by combinations of IgA, EMA, HLA, and histological states. These findings indicate that clinical variables alone tend to recover the symptom-oriented organization represented by Oslo, whereas diagnostic variables describe a different and more fragmented structure. The full multimodal analysis integrates these complementary dimensions. Detailed community compositions are reported in Supplementary Tables 5 and 6.

4.4. Comparison between Oslo classification and TDA with supervised ML

For both the original Oslo classes and the TDA communities, RF is the ML model that reported the best mean classification performances, as shown in Table 2. Looking at all the classification metrics reported in Table 2, precision, recall, F1-score and accuracy appear slightly lower in

general for the models applied to the TDA communities with respect to those applied to the Oslo groups, however it should be noted that there are seven TDA communities and only three Oslo classes, and increasing the number of classes increases the chance of misclassification in each single class leading to a natural decrease of these classification metrics. AUC however, which is the most robust metric when increasing the number of classes, is higher for TDA communities, suggesting that TDA communities are generally slightly easier to detect using a classification approach, and therefore groups patients in more clinically characterized categories.

Fig. 8 shows the confusion matrices and ROC curves for the RF model applied to both classification systems. Regarding the Oslo classification, it can be noticed that the asymptomatic group, although less numerous than the other two, is the most accurately predicted, denoting that the clinical distinction between classical and non-classical is fainter. Looking at the TDA communities classification, the most misclassified community is community 4, which is often labelled as 5 or 3, probably due to

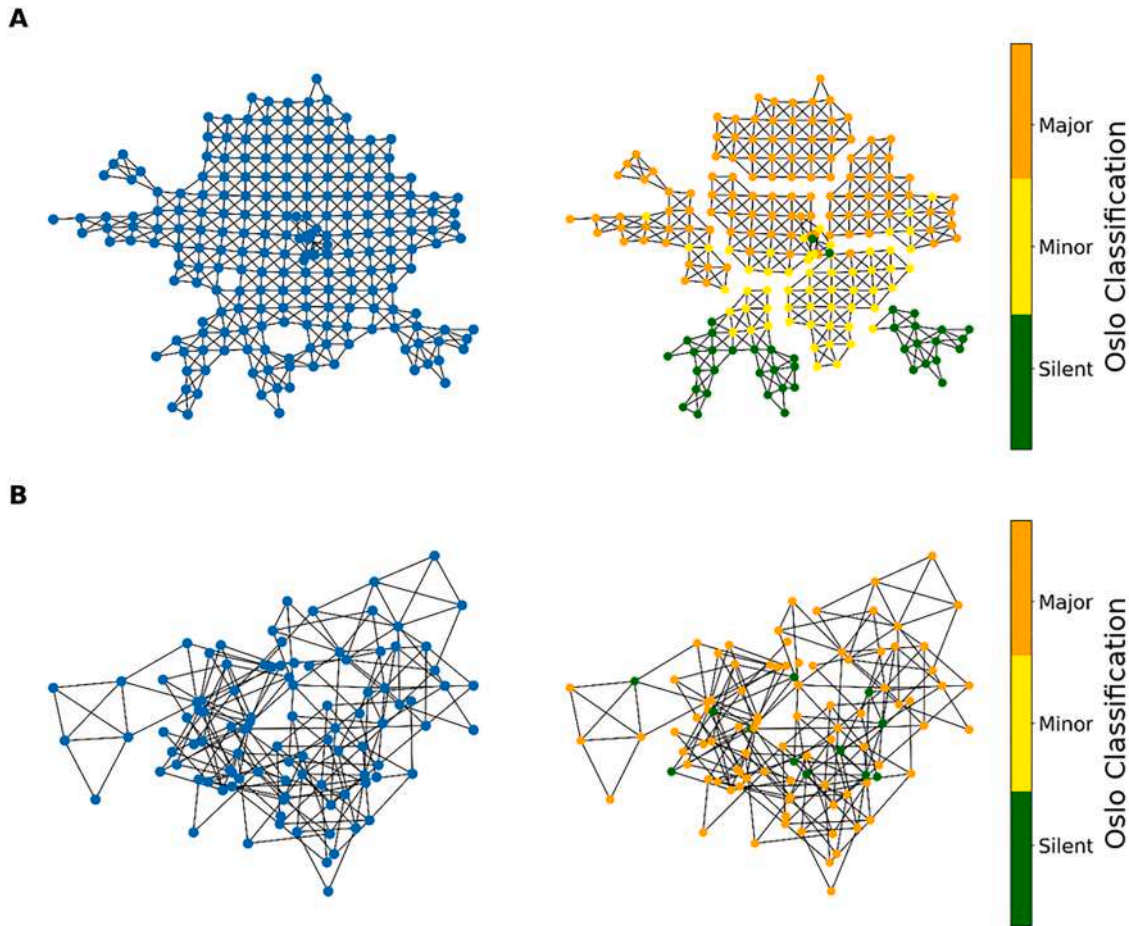


Fig. 7. Feature-domain sensitivity analysis. pheTDA graphs generated using (A) clinical variables only and (B) diagnostic-test variables only. For each analysis, the uncolored graph is shown on the left and the same graph color-coded by the Oslo classification on the right.

Table 2

Performances for the three classification models trained on both the Oslo classes and TDA communities in one-vs-rest binary classification tasks. For each model the mean values of precision, recall, F1-score, accuracy and Area Under the receiver operating characteristic Curve (AUC) are reported, by averaging their values on all the binary tasks. The best values are reported in bold.

Classification	Model	Precision	Recall	F1-score	Accuracy	AUC
Oslo	RF	0.76	0.76	0.76	0.76	0.88
	LR	0.67	0.66	0.65	0.66	0.84
	HistBoost	0.72	0.71	0.71	0.72	0.86
pheTDA	RF	0.71	0.70	0.71	0.72	0.94
	LR	0.57	0.54	0.55	0.54	0.87
	HistBoost	0.66	0.66	0.65	0.66	0.92

the similarities that these groups share in terms of presence of autoimmune comorbidities or types of immune response.

4.5. Histology distribution

Although sometimes not collected in children, gut histology remains the main indicator of the presence of active CD, as the clinical presentation and the possible development of complications depends on the level of malabsorption. The exclusion of the histological examination from some patients led to a high number of missing data in the original dataset used to train the supervised classifiers, however, statistical tests show that each TDA community presents notable differences in their typical histological presentation. Table 3 shows the percentage of each Corazza histology category for each community; pairwise Chi-square

tests with the correction of Bonferroni were applied to all pairs of communities obtaining P-values lower than 0.001. Although this result is not unexpected, since the histological information was used in the creation of TDA communities, it shows that the same communities are easy to predict by a ML algorithm even without including this information, and therefore the level of intestinal damage could be estimated even in absence of a biopsy. Particularly, community 7, which is associated with the lowest classification error, is the one that contains the highest percentage of high histological damage, i.e. Corazza B2.

5. Discussion and conclusion

Although very common and widely studied, numerous mechanisms characterizing the etiology and clinical presentation of CD are still unknown. Plus, the heterogeneity of symptoms and the high prevalence of asymptomatic or subclinical forms make CD an extremely under-diagnosed disease, with an estimated 95% of patients still not discovered at this time [3]. For these reasons, clinical characterization of CD is extremely important, as understanding the possible disease phenotypes is a first step towards the creation of precise diagnostic and prognostic strategies that consider the individual risk of complications and disease trajectory. In this context, data-driven approaches can be a useful tool to find patterns in such heterogeneous data. Although classical clustering methods have been shown to be effective in past studies on adult datasets [11], pediatric CD presents a much more complicated case, as it can present a higher number and diversity of symptoms, and requires more sophisticated sub-phenotyping approaches. In fact, the current standard categorization of CD is still made of the Oslo definitions, which are based

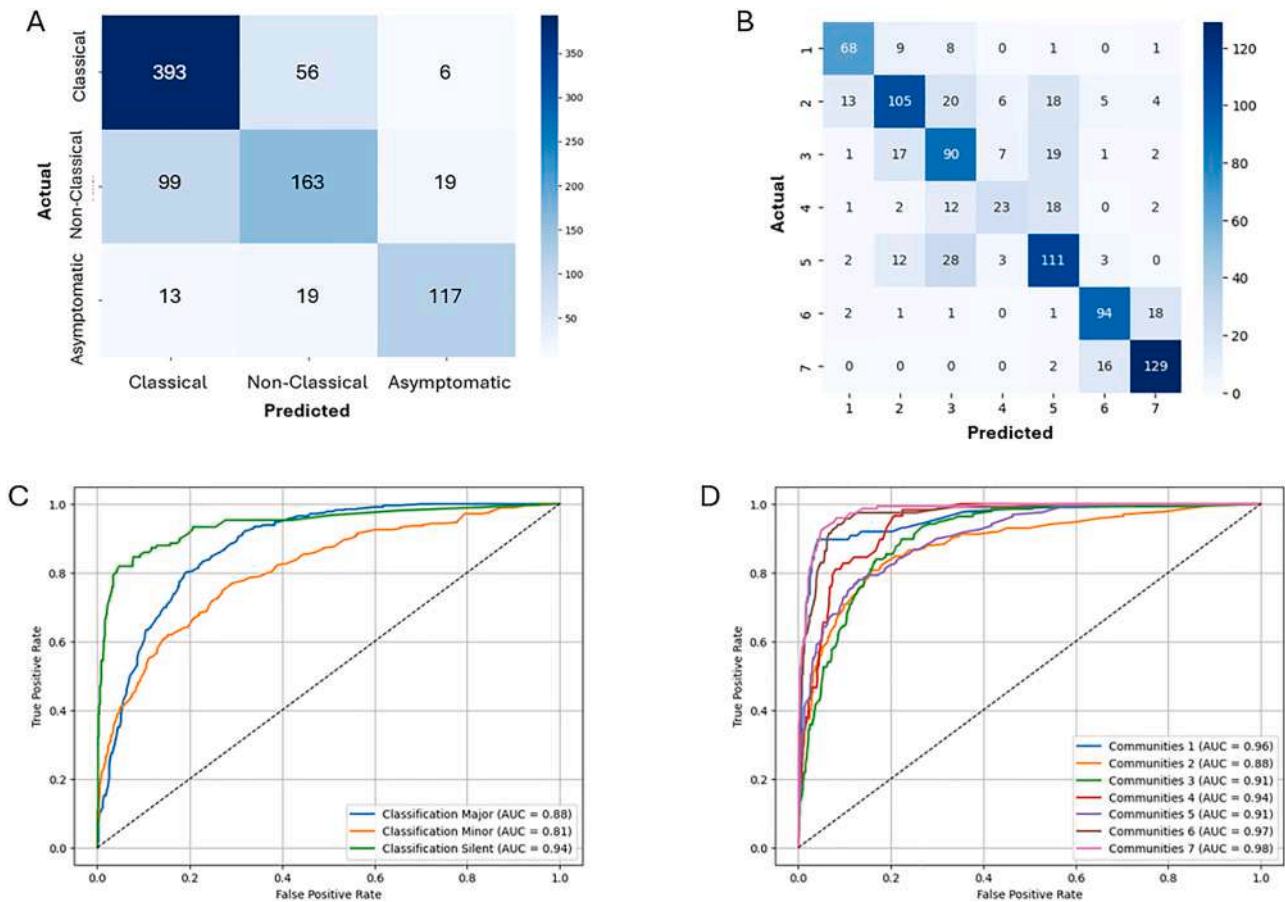


Fig. 8. Classification results on the Oslo classes and TDA Communities. (A) Confusion matrix obtained on the Oslo classes. (B) Confusion matrix obtained on the TDA communities. (C) ROC curves obtained on the Oslo classes with AUC values for each initial class reported in the legend. (D) ROC curves obtained on the TDA communities with AUC values for each community reported in the legend.

Table 3
Number of observations and related percentage of each Corazza histological category for each pheTDA community.

Corazza histology	pheTDA community						
	1	2	3	4	5	6	7
Normal	3 (1.03%)	-	6 (1.31%)	-	-	1 (0.26%)	-
A	2 (0.69%)	11 (1.93%)	5 (1.09%)	11 (5.67%)	26 (4.92%)	8 (2.05%)	6 (1.23%)
B1	32 (11%)	249 (43.76%)	34 (7.44%)	65 (33.51%)	325 (61.55%)	256 (65.64%)	127 (26.02%)
B2	50 (17.18%)	231 (40.60%)	74 (16.19%)	97 (50%)	114 (21.59%)	121 (31.03%)	355 (72.75%)
Not Performed	204 (70.10%)	78 (13.71%)	338 (73.96%)	32 (10.82%)	63 (11.93%)	4 (1.03%)	-

mainly on symptoms and are defined in the same way for both adult and pediatric patients, in spite of their clearly different clinical patterns which are confirmed also by the different diagnostic criteria.

The study presented in this paper represents a first attempt to create a more precise clinical characterization of pediatric CD, identifying recognizable sub-phenotypes that integrate the inherent complexity and multimodality of CD-related data. Our approach is based on a graph representation learning method that leverages topological data analysis called pheTDA, which has shown good performances in past applications on small datasets. In particular, our approach finds seven subgroups, all characterized by different prevailing clinical traits. The higher performance demonstrated by independent supervised classifiers in identifying these subgroups with respect to the original Oslo classes suggests that the identified subgroups capture patterns that are more

readily separable from the available clinical variables than the original Oslo classes . In particular, all these phenotypes appear to be highly related to intestinal damage, as demonstrated by the final statistical tests. These findings further suggest that non-histological clinical information contains signals associated with the identified phenotypic patterns, although prospective validation would be required before considering their use for the classification of new patients. .

The benchmarking analysis further clarifies the contribution of pheTDA. DBSCAN produced a highly imbalanced solution, whereas AHC was largely separated by sex; in contrast, pheTDA generated more balanced communities whose leading discriminators represented all major data domains. The sensitivity analysis also showed that using clinical manifestations alone tended to reconstruct an Oslo-like organization, while diagnostic-test variables alone generated narrower test-

pattern groups. Together, these findings support multimodal integration as the defining contribution of the proposed phenotyping strategy rather than similarity to the Oslo categories, which were used as a clinical reference rather than as a ground-truth partition.

Arguably, this study presents some limitations. For example, the high missingness rate of the data required intensive imputation based mostly on previous knowledge (e.g. the creation of a “not performed” variable for histology). Missing values, in fact, were not uniformly treated as randomly missing observations. Where clinically justified by the pediatric CD diagnostic pathway, missingness was encoded using explicit categories (e.g., “not performed” or “not available”), thereby preserving information regarding whether a diagnostic examination was omitted according to the clinical workflow. These clinically informed encodings should therefore be distinguished from conventional statistical imputation of an unobserved value. This procedure however, although important to maximize the clinical information contained in the data, could lead to an increase in overfitting, as artificial variables are used for clustering. This effect might be made even worse by the underlying dependences between different variables. Plus, as with other unsupervised and topology-based approaches, the representation generated by Mapper depends on the preprocessing strategy, the distance metric, the lens function, and the Mapper hyperparameters. Therefore, the communities identified by pheTDA should be interpreted as data-driven structures resulting from the adopted analytical configuration rather than as unique or intrinsic disease classes. Nevertheless, the supervised reclassification procedure shows that the found communities are easy to identify by an automated machine learning algorithm, with higher performances compared to the current classes in spite of the probable presence of some overfitting.

Overall, approaches such as pheTDA can be promising tools to find underlying sub-phenotypes of complex, multifactorial diseases such as CD, allowing to properly find groups of patients that share clinically relevant characteristics that are difficult to identify by the common clinical practice or standard clustering algorithms. Future studies should assess the reproducibility of the identified phenotypes in independent and prospectively collected pediatric CD cohorts and investigate whether additional molecular or longitudinal data can further refine the resulting patient stratification.

Ethics information

The research reported in this paper does not require an ethics statement, as no subjects were directly involved, but an old and completely anonymized dataset was used.

Code availability

The Python code to implement TDA is available at this link: https://github.com/Sep905/pheTDA/tree/main/pediatric_celiac_disease. For ethical reasons, we are not allowed to share the original dataset, however, to facilitate code use, we created a small fac-simile that can be found inside the same folder.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 5.5 in order to check grammar/spelling and improve the style of some limited parts of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRedit authorship contribution statement

Giuseppe Albi: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Valentina**

Brembilla: Methodology, Formal analysis, Data curation. **Erika Lenzi:** Methodology, Formal analysis, Data curation. **Simone Maffioletti:** Methodology, Formal analysis, Data curation. **Emanuele Medolago:** Methodology, Formal analysis, Data curation. **Chiara Sirtoli:** Methodology, Formal analysis, Data curation. **Antonio Ferramosca:** Writing – review & editing, Validation, Supervision. **Marco Vincenzo Lenti:** Resources, Data curation. **Antonio Di Sabatino:** Resources, Project administration, Funding acquisition, Data curation. **Arianna Dagliati:** Methodology, Investigation, Conceptualization. **Daniele Pala:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis.

Declaration of competing interest

The Authors have no conflicts of interest to disclose.

Acknowledgements

This work was funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) – project n. PNC00000003 – AdvanCed Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors’ views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them - CUP B53C22006700001.

We thank all the participants to the previous study who contributed to the data collection: Paola Ilaria Bianchi, Clarissa Petrucci, Giulia Gambini, Nicola Aronico, Matteo Varallo, Carlo Maria Rossi, Elena Pozzi, Elena Groppali, Francesca Siccario, Giulia Franchino, Gian Vincenzo Zuccotti, Grazia Di Leo, Chiara Zanchi, Fernanda Cristofori, Ruggiero Francavilla, Marina Aloï, Giulia Gagliostro, Monica Montuori, Sara Romaggioli, Caterina Strisciuglio, Marco Crocco, Noemi Zampatti, Angela Calvi, Renata Auricchio, Costantino De Giacomo, Silvia Maria Elena Caimmi, Carolina Carraro, Annamaria Staiano, Sabrina Cenni, Mauro Congia, Enrico Schirru, Francesca Ferretti, Carolina Ciacci, Nicoletta Vecchione, Mario Andrea Latorre, Semela Resuli, Giusy Cinzia Moltisanti, Giulia Maria Abruzzese, Andrea Quadrelli, Simone Saglio, Pietro Canu, Damiano Ruggeri, Annalisa De Silvestri, Catherine Klersy, Gian Luigi Marseglia, and Gino Roberto Corazza.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.cmpb.2026.109648](https://doi.org/10.1016/j.cmpb.2026.109648).

References

- [1] A.D. Sabatino, G.R. Corazza, Coeliac disease, *Lancet* 373 (2009) 1480–1493, [https://doi.org/10.1016/S0140-6736\(09\)60254-3](https://doi.org/10.1016/S0140-6736(09)60254-3).
- [2] Y. Sahin, Celiac disease in children: a review of the literature, *World J. Clin. Pediatr.* 10 (2021) 53–71, <https://doi.org/10.5409/wjcp.v10.i4.53>.
- [3] G.K. Makharia, A. Chauhan, P. Singh, V. Ahuja, Review article: epidemiology of coeliac disease, *Aliment. Pharmacol. Ther.* 56 (2022) S3–S17, <https://doi.org/10.1111/apt.16787>.
- [4] J.F. Ludvigsson, D.A. Leffler, J.C. Bai, F. Biagi, A. Fasano, P.H.R. Green, M. Hadjivassiliou, K. Kaukinen, C.P. Kelly, J.N. Leonard, K.E.A. Lundin, J. A. Murray, D.S. Sanders, M.M. Walker, F. Zingone, C. Ciacci, The Oslo definitions for coeliac disease and related terms, *Gut* 62 (2013) 43–52, <https://doi.org/10.1136/gutjnl-2011-301346>.
- [5] F. Biagi, A. Schieppatti, G. Maiorano, G. Fraternale, S. Agazzi, F. Zingone, C. Ciacci, U. Volta, G. Caio, R. Tortora, C. Klersy, G.R. Corazza, Risk of complications in coeliac patients depends on age at diagnosis and type of clinical presentation, *Digest. Liver Disease* 50 (2018) 549–552, <https://doi.org/10.1016/j.dld.2017.12.001>.
- [6] C. Gallegos, R. Merkel, Current evidence in the diagnosis and treatment of children with celiac disease, *Gastroenterol. Nurs.* 42 (2019) 41, <https://doi.org/10.1097/SGA.0000000000000365>.

- [7] S. Husby, S. Koletzko, I. Korponay-Szabó, K. Kurppa, M.L. Mearin, C. Ribes-Koninckx, R. Shamir, R. Troncone, R. Auricchio, G. Castillejo, R. Christensen, J. Dolinsek, P. Gillett, A. Hróbjartsson, T. Kolta, M. Maki, S.M. Nielsen, A. Popp, K. Størdal, K. Werkstetter, M. Wessels, European society paediatric gastroenterology, hepatology and nutrition guidelines for diagnosing coeliac disease 2020, *J. Pediatr. Gastroenterol. Nutr.* 70 (2020) 141–156, <https://doi.org/10.1097/MPG.0000000000002497>.
- [8] M.N. Marsh, **Gluten, major histocompatibility complex, and the small intestine. a molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue')**, *Gastroenterology* 102 (1992) 330–354.
- [9] G. Oberhuber, G. Granditsch, H. Vogelsang, The histopathology of coeliac disease: time for a standardized report scheme for pathologists, *Eur. J. Gastroenterol. Hepatol.* 11 (1999) 1185–1194, <https://doi.org/10.1097/00042737-199910000-00019>.
- [10] G.R. Corazza, V. Villanacci, Coeliac disease, *J. Clin. Pathol.* 58 (2005) 573–574, <https://doi.org/10.1136/jcp.2004.023978>.
- [11] M.V. Lenti, D. Pala, N. Aronico, A. Dagliati, P.I. Bianchi, G. Santacroce, S. Quaglini, G. Lanzola, G.R. Corazza, A.D. Sabatino, F.D. Pasquale, A.L. Bello, D. Bagordo, C. C. D'Agate, M. Neri, U. Volta, M.G. Mumolo, M. Astegiano, A.S. Calabrò, F. Zingone, G. Latella, A.D. Sario, A. Carroccio, C. Ciacci, F. Luzzza, C. Bagnato, M. C. Fantini, L. Elli, G. Cammarota, A. Gasbarrini, P. Portincasa, M.A. Latorre, C. Petrucci, C. Quatraccioni, C. Iannelli, N. Vecchione, C.M. Rossi, G. Broglio, G. Ianiro, I. Marsilio, S. Bibbò, B. Marinoni, D. Tomaselli, L. Abenavoli, E. Lynch, A. Carrieri, P. Mansueto, G. Alunno, C. Rossi, F. Onnis, K. Efthymakis, N. Cesaro, M. Vernerio, F.B. Svizzero, F.P. Semeraro, A. Vanoli, Latent class analysis identifies novel coeliac disease subgroups with distinctive clinical features: a multicentric study, *Eur. J. Intern. Med.* 139 (2025), <https://doi.org/10.1016/j.ejim.2025.05.020>.
- [12] J.M. Banda, M. Seneviratne, T. Hernandez-Boussard, N.H. Shah, Advances in electronic phenotyping: from rule-based definitions to machine learning models, *Annu. Rev. Biomed. Data Sci.* 1 (2018) 53–68, <https://doi.org/10.1146/annurev-biodatasci-080917-013315>.
- [13] Y. Halpern, Y. Choi, S. Horng, D. Sontag, Using anchors to estimate clinical state without labeled data, *AMIA Annu. Symp. Proc.* 2014 (2014) 606–615.
- [14] Y. Halpern, S. Horng, Y. Choi, D. Sontag, Electronic medical record phenotyping using the anchor and learn framework, *J. Am. Med. Inform. Assoc.* 23 (2016) 731–740, <https://doi.org/10.1093/jamia/ocw011>.
- [15] P.Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, G. Carlsson, Extracting insights from the shape of complex data using topology, *Sci. Rep.* 3 (2013) 1236, <https://doi.org/10.1038/srep01236>.
- [16] G. Singh, F. Memoli, G. Carlsson, Topological methods for the analysis of high dimensional data sets and 3D object recognition, *The Eurograph. Associat.* (2007), <https://doi.org/10.2312/SPBG/SPBG07/091-100>.
- [17] M. Nicolau, A.J. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, *Proc. Natl. Acad. Sci. USA* 108 (2011) 7265–7270, <https://doi.org/10.1073/pnas.1102826108>.
- [18] L. Li, W.-Y. Cheng, B.S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, J.T. Dudley, Identification of type 2 diabetes subgroups through topological analysis of patient similarity, *Sci. Transl. Med.* 7 (2015) 311ra174, <https://doi.org/10.1126/scitranslmed.aaa9364>.
- [19] G. Casacclang-Verzosa, S. Shrestha, M.J. Khalil, J.S. Cho, M. Tokodi, S. Balla, M. Alkhouli, V. Badhwar, J. Narula, J.D. Miller, P.P. Sengupta, Network tomography for understanding phenotypic presentations in aortic stenosis, *JACC: Cardiovas. Imaging* 12 (2019) 236–248, <https://doi.org/10.1016/j.jcmg.2018.11.025>.
- [20] T.S.C. Hinks, X. Zhou, K.J. Staples, B.D. Dimitrov, A. Manta, T. Petrossian, P. Y. Lum, C.G. Smith, J.A. Ward, P.H. Howarth, A.F. Walls, S.D. Gadola, R. Djukanović, Innate and adaptive T cells in asthmatic patients: relationship to severity and disease mechanisms, *J. Allergy Clin. Immunol.* 136 (2015) 323–333, <https://doi.org/10.1016/j.jaci.2015.01.014>.
- [21] G. Albi, A. Gerbasi, M. Chiesa, G.I. Colombo, R. Bellazzi, A. Dagliati, A Topological Data Analysis Framework for Computational Phenotyping, in: J.M. Juarez, M. Marcos, G. Stiglic, A. Tucker (Eds.), *Artificial Intelligence in Medicine*, Springer Nature Switzerland, Cham, 2023, pp. 323–327, https://doi.org/10.1007/978-3-031-34344-5_38.
- [22] M. Chiesa, A. Dagliati, A. Gerbasi, G. Albi, L. Ballarini, L. Piacentini, PIUMA: Phenotypes Identification Using Mapper from topological data Analysis, <http://bioconductor.org/packages/PIUMA/>, 2025.
- [23] P.I. Bianchi, M.V. Lenti, C. Petrucci, G. Gambini, N. Aronico, M. Varallo, C. M. Rossi, E. Pozzi, E. Groppali, F. Siccardi, G. Franchino, G.V. Zuccotti, G. Di Leo, C. Zanchi, F. Cristofori, R. Francavilla, M. Aloï, G. Gagliostro, M. Montuori, S. Romaggioli, C. Strisciuglio, M. Crocco, N. Zampatti, A. Calvi, R. Auricchio, C. De Giacomo, S.M.E. Caimmi, C. Carraro, A. Staiano, S. Cenni, M. Congia, E. Schirru, F. Ferretti, C. Ciacci, N. Vecchione, M.A. Latorre, S. Resuli, G.C. Moltisanti, G. M. Abruzzese, A. Quadrelli, S. Saglio, P. Canu, D. Ruggeri, A. De Silvestri, C. Klersy, G.L. Marsegli, G.R. Corazza, A. Di Sabatino, Diagnostic delay of celiac disease in childhood, *JAMA Netw. Open.* 7 (2024) e245671, <https://doi.org/10.1001/jamanetworkopen.2024.5671>.
- [24] G.R. Corazza, V. Villanacci, C. Zambelli, M. Milione, O. Luinetti, C. Vindigni, C. Chioda, L. Albarello, D. Bartolini, F. Donato, Comparison of the interobserver reproducibility with different histologic criteria used in celiac disease, *Clin. Gastroenterol. Hepatol.* 5 (2007) 838–843, <https://doi.org/10.1016/j.cgh.2007.03.019>.
- [25] J.C. Gower, A general coefficient of similarity and some of its properties, *Biometrics* 27 (1971) 857–871, <https://doi.org/10.2307/2528823>.
- [26] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech.* 2008 (2008) P10008, <https://doi.org/10.1088/1742-5468/2008/10/P10008>.
- [27] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, *Phys. Rev. E* 69 (2004) 026113, <https://doi.org/10.1103/PhysRevE.69.026113>.
- [28] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 67 (2005) 301–320, <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.
- [29] 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, É. Duchesnay, Scikit-learn: machine learning in python, *J. Mach. Learn. Res.* 12 (2011) 2825–2830.
- [30] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, *Mach. Learn.* 46 (2002) 389–422, <https://doi.org/10.1023/A:1012487302797>.
- [31] L. Breiman, Random Forests, *Mach. Learn.* 45 (2001) 5–32, <https://doi.org/10.1023/A:1010933404324>.
- [32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (accessed October 26, 2025).
- [33] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (2008) 2579–2605.
- [34] U. von Luxburg, A tutorial on spectral clustering, *Stat. Comput.* 17 (2007) 395–416, <https://doi.org/10.1007/s11222-007-9033-z>.