Inflammatory Bowel Diseases

Silvana Trindade, Leandro N. Zanotto, Rolan Alexander Valle Rey
Institute of Computing
State University of Campinas
Campinas, Brazil
silvana@lrc.ic.unicamp.br, 001963@g.unicamp.br, r230204@dac.unicamp.br

***Abstract*—Inflammatory Bowel Diseases (IBDs) describe disorders involving chronic inflammation of the digestive tract. IBDs include Ulcerative Colitis (UC) and Crohn's disease, they are considered as rising worldwide infections. The IBD cases have been expanding in numerous zones of the world, especially in more developed cities, which are more industrialized. For IBD diagnosis, standard diagnostic approaches use clinical exams such as endoscopic imaging, biomarkers, and histopathology. However, proper diagnosis cannot be reached in up to 10% of patients with UC, thus, more efficient strategies for diagnosis is required [13]. This article presents different approaches to data science to address this problem. Using ontology analysis we found that UC is the topic with the most research. The network analysis shows the relevant problems that patients with IBD have in common, and based on this it is possible to create a personalized questionnaire to help doctors in diagnosis. Using Unsupervised Learning (UL) methods, we foud that CD and UC do not form clusters and IBD could be classified into 4 classes.On the other hand, using supervised learning, the ensemble method of soft voting (using Support Vector Machine, Gaussian Processes and k-Nearest Neighbor) achieves a better accuracy compared to Endoscopy and Histology examination.**

***Index Terms*—ontologies, complex networks, machine learning, statistics, symptoms, nutrition, ibd**

I. Introduction

Inflammatory bowel disease (IBD) is a chronic inflammatory condition of unknown etiology that is thought to result from a combination of genetic, immunologic and environmental factors. Table I shows the common symptoms that can be developed by a person with digestive diseases and specific IBDs. These symptoms and specific exams are currently the methodologies for the diagnosis.

TABLE I
Common Symptoms

Digestive Disease	**Crohn's**	**Ulcerative Colitis**
Bleeding, bloating, constipation, diarrhea, heartburn, incontinence, nausea and vomiting, pain in the belly, swallowing problems, and weight gain or loss.	Pain areas in the abdomen, joints, or rectum, bloating, blood in the stool, bowel obstruction, diarrhea, nausea, vomiting, flatulence, fatigue, fever, loss of appetite, anal fissure, cramping, depression, flare, mouth ulcer, slow growth, or weight loss.	Pain in the abdomen, joints, or rectum, pain can be intermittent in the abdomen, bloating, blood in stool, diarrhea, inability to empty bowels, leaking of stool, unusually frequent defecation, urgent need to defecate, anemia, fatigue, fever, or loss of appetite, scarring within the bile ducts or weight loss.

In this work, we propose to use ontologies, networks, machine learning, and statistics analysis to correlate features (exams, habits questions, and literature studies), finding non-trivial relationships and metrics that can help in the IBD treatments. The Unified Modeling Language (UML) is used for software projects. It can be employed for visualization, specification and building the artifacts for complex systems. In this case it is not useful to manage the data we have on this work. This work uses different datasets with different models, for example, on complex networks are created to represent the relation between nodes (people and features). The relational database will also not be useful for this work since it needs a more structured data with known relation between entities. Traditional data analysis requires an understanding of the data and the underlying relationships, whereas our approaches are able to work with unknown data and from the results generate an understanding.

The aim is to study IBDs using correlations between symptoms and habits, helping patients with IBDs improve their quality of life and doctors to help patients. IBDs involve many questions to reach a diagnosis and after that patients need to change habits and discover things that could hinder the treatments.

The rest of the work is organized as follows. Section II describes the ontology analysis. Section III describes the complex network analysis. Section IV describes the machine learning analysis. Section V describes the statistics analysis got from the datasets used in this work. Section VII describes the conclusion of this work.

II. Ontologies

The word *ontology* is used with different meanings in different communities. The most radical difference is perhaps between the philosophical and the computational sense, which emerged in the recent years in the knowledge engineering community, starting from an early informal definition of (computational) ontologies as "explicit specifications of conceptualizations" [12].

This paper looks for ontologies related to the IBDs and some keywords like Food, Diet, Nutrition, Intolerance and Diet from 1991 to 2016. Using pyTag (Automated identification of ontological terms in application area specific literature surveys) [8], we extracted the ontologies from Pubmed citations.

Getting the ontologies according to the Disease and keyword in the literature, some statistics can be used to check the frequency of terms, the dates they are most published or when the researchers started to look for them, for example, ”probiotics”.

A. Dataset

The dataset used is the publications found on Pubmed from 1991 to 2016. They are split into three Diseases, UC, Crohn’s and IBS and also there are five keywords used on each disease search. The Figure 2 presents it.

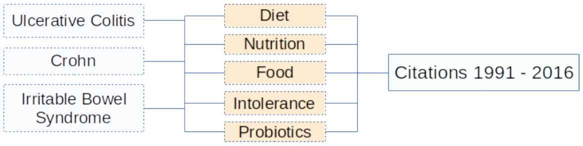


Fig. 1. Citations Search on Pubmed using keyword as filter.

B. Techniques

Figure 2 shows the flow to create the result tables for analysis. The first step is to get the citations from PubMed and use the EndNote [2] to add the Accession Number to export into Bibtex files. PyTag will process all files accessing the internet and Extract 2.0 will check which word is an ontology, annotating and classifying them to create the ontology catalog splitted on organism, disease, biological process, tissue, cellular component, genes/proteins, chemical compound, molecular function, and environment.

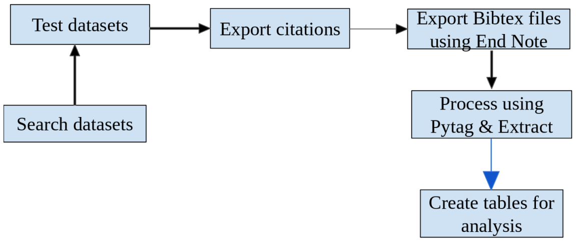
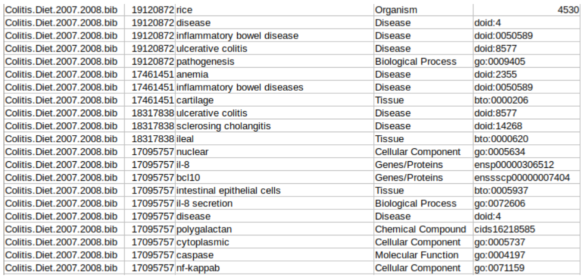

Test datasets
Export citations
Export Bibtex files using End Note
Search datasets
Process using Pytag & Extract
Create tables for analysis

Fig. 2. Ontologies Execution Flow using pyTag and Extract 2.0.

C. Results

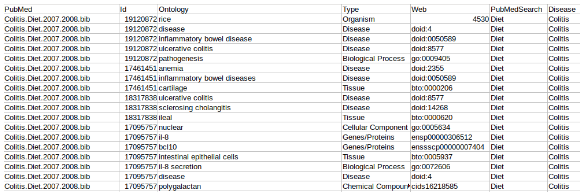
The Figure 3 is the result table processed by pyTag. The first column is the file name, the second is the identifier number of the file, the third is the ontology found on citations, the forth is the category and the last column is the ontology identification. For example doid:4 is the disease ontology found on .

Since this table is not enough four our analysis, two other columns were added. The PubMedSearch, which is the keyword searched with the diseases (e.g. ”Diet”). The second column is the diseases, so they can be useful for the analysis on WEKA.



Colitis.Diet.2007.2008.bib	19120872	rice	Organism	4530
Colitis.Diet.2007.2008.bib	19120872	disease	Disease	doid:4
Colitis.Diet.2007.2008.bib	19120872	inflammatory bowel disease	Disease	doid:0050589
Colitis.Diet.2007.2008.bib	19120872	ulcerative colitis	Disease	doid:8577
Colitis.Diet.2007.2008.bib	19120872	pathogenesis	Biological Process	go:0009405
Colitis.Diet.2007.2008.bib	17461451	anemia	Disease	doid:2355
Colitis.Diet.2007.2008.bib	17461451	inflammatory bowel diseases	Disease	doid:0050589
Colitis.Diet.2007.2008.bib	17461451	cartilage	Tissue	bto:0000206
Colitis.Diet.2007.2008.bib	18317838	ulcerative colitis	Disease	doid:8577
Colitis.Diet.2007.2008.bib	18317838	sclerosing cholangitis	Disease	doid:14268
Colitis.Diet.2007.2008.bib	18317838	ileal	Tissue	bto:0000620
Colitis.Diet.2007.2008.bib	17095757	nuclear	Cellular Component	go:0005634
Colitis.Diet.2007.2008.bib	17095757	il-8	Genes/Proteins	ensp0000306512
Colitis.Diet.2007.2008.bib	17095757	bcl10	Genes/Proteins	enssscp0000007404
Colitis.Diet.2007.2008.bib	17095757	intestinal epithelial cells	Tissue	bto:0005937
Colitis.Diet.2007.2008.bib	17095757	il-8 secretion	Biological Process	go:0072606
Colitis.Diet.2007.2008.bib	17095757	disease	Disease	doid:4
Colitis.Diet.2007.2008.bib	17095757	polygalactan	Chemical Compound	cids16218585
Colitis.Diet.2007.2008.bib	17095757	cytoplasmic	Cellular Component	go:0005737
Colitis.Diet.2007.2008.bib	17095757	caspase	Molecular Function	go:0004197
Colitis.Diet.2007.2008.bib	17095757	nf-kappab	Cellular Component	go:0071159

Fig. 3. PyTag result table.



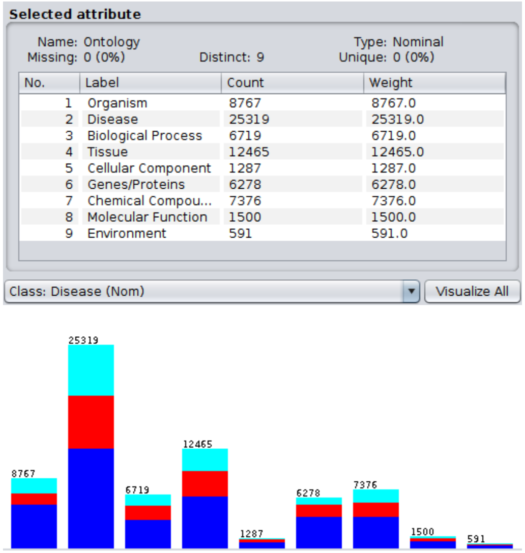
PubMed	Id	Ontology	Type	Web	PubMedSearch	Disease
Colitis.Diet.2007.2008.bib	19120872	rice	Organism	4530	Diet	Colitis
Colitis.Diet.2007.2008.bib	19120872	disease	Disease	doid:4	Diet	Colitis
Colitis.Diet.2007.2008.bib	19120872	inflammatory bowel disease	Disease	doid:0050589	Diet	Colitis
Colitis.Diet.2007.2008.bib	19120872	ulcerative colitis	Disease	doid:8577	Diet	Colitis
Colitis.Diet.2007.2008.bib	19120872	pathogenesis	Biological Process	go:0009405	Diet	Colitis
Colitis.Diet.2007.2008.bib	17461451	anemia	Disease	doid:2355	Diet	Colitis
Colitis.Diet.2007.2008.bib	17461451	inflammatory bowel diseases	Disease	doid:0050589	Diet	Colitis
Colitis.Diet.2007.2008.bib	17461451	cartilage	Tissue	bto:0000206	Diet	Colitis
Colitis.Diet.2007.2008.bib	18317838	ulcerative colitis	Disease	doid:8577	Diet	Colitis
Colitis.Diet.2007.2008.bib	18317838	sclerosing cholangitis	Disease	doid:14268	Diet	Colitis
Colitis.Diet.2007.2008.bib	18317838	ileal	Tissue	bto:0000620	Diet	Colitis
Colitis.Diet.2007.2008.bib	17095757	nuclear	Cellular Component	go:0005634	Diet	Colitis
Colitis.Diet.2007.2008.bib	17095757	il-8	Genes/Proteins	ensp00000306512	Diet	Colitis
Colitis.Diet.2007.2008.bib	17095757	bcl10	Genes/Proteins	enssscp00000007404	Diet	Colitis
Colitis.Diet.2007.2008.bib	17095757	intestinal epithelial cells	Tissue	bto:0005937	Diet	Colitis
Colitis.Diet.2007.2008.bib	17095757	il-8 secretion	Biological Process	go:0072606	Diet	Colitis
Colitis.Diet.2007.2008.bib	17095757	disease	Disease	doid:4	Diet	Colitis
Colitis.Diet.2007.2008.bib	17095757	polygalactan	Chemical Compoun	cids16218585	Diet	Colitis

Fig. 4. PyTag result table with two columns added for analysis on WEKA.

D. Weka Results

WEKA provides a toolbox of learning algorithms, but also a framework which researchers could implement new algorithms without being concerned with supporting infrastructure for data manipulation and scheme evaluation [6]. Therefore it was chosen to analyze the pyTag result table.

Figure 5 shows how many ontologies were found on each keyword. The words related to the Disease are the most found and the ones related to the environment are the ones which have less match.



Selected attribute

Name: Ontology
Missing: 0 (0%)
Distinct: 9
Type: Nominal
Unique: 0 (0%)

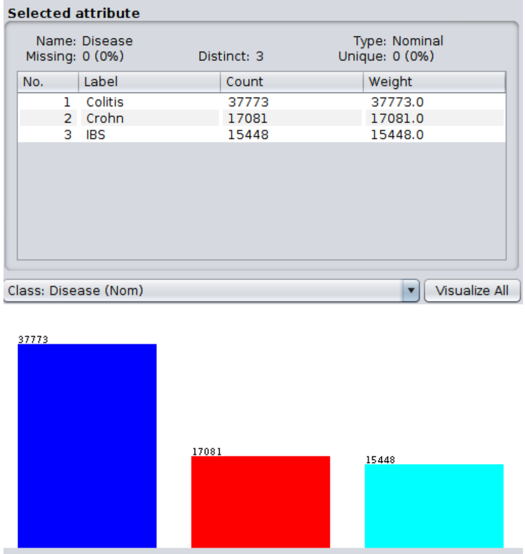
No.	Label	Count	Weight
1	Organism	8767	8767.0
2	Disease	25319	25319.0
3	Biological Process	6719	6719.0
4	Tissue	12465	12465.0
5	Cellular Component	1287	1287.0
6	Genes/Proteins	6278	6278.0
7	Chemical Compou...	7376	7376.0
8	Molecular Function	1500	1500.0
9	Environment	591	591.0

Class: Disease (Nom)

Fig. 5. Ontologies grouped by keyword identified by Extract 2.0

Figure 6 shows how many ontologies were found for each disease. Colitis has the most results compared to other diseases. Thus the other diseases have more research to be done in the next years.

Clustering was done to correlate the pyTag table columns to each other. Two results were selected to show the relation



Selected attribute

Name: Disease Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

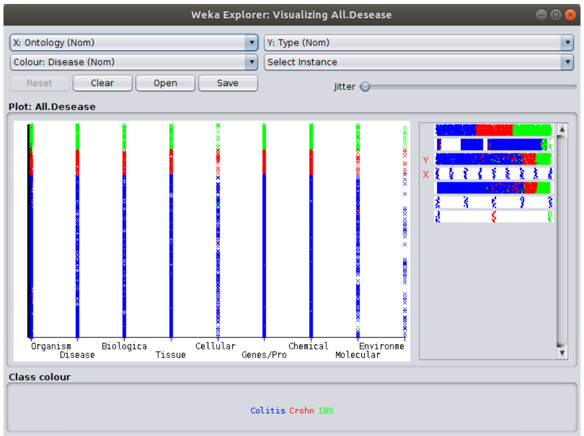
No.	Label	Count	Weight
1	Colitis	37773	37773.0
2	Crohn	17081	17081.0
3	IBS	15448	15448.0

Class: Disease (Nom)

Fig. 6. Ontologies count related to each disease.

between them.

Figure 7 shows the Type column (which are the ontologies) grouped by Extract 2.0 terms. Terms related to Colitis are predominant.



X: Ontology (Nom)

Y: Type (Nom)

Colour: Disease (Nom)

Plot: All.Desease

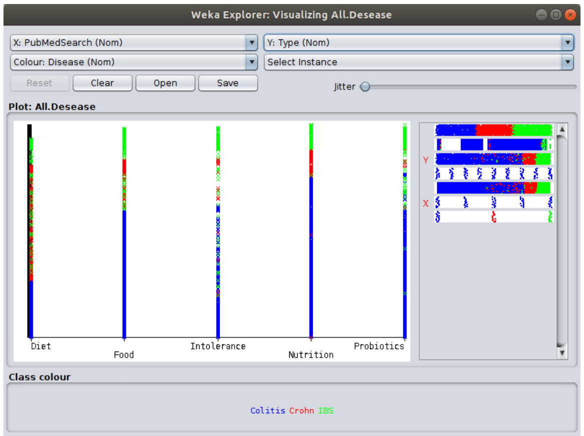
Fig. 7. Ontologies found on each Group.

Figure 8 presents the terms found on citations for each filter. Nutrition has the most terms in blue related to Colitis.

III. Complex Network

The term *complex networks* describes a class of large networks which exhibit the following properties: a) they are sparse, meaning, the number of edges is proportional to the number of nodes; b) they exhibit the small world phenomenon: almost all pairs of vertices are in the same component are within a short distance from each other; c) clustering is present: two nodes of the network that have a common neighbor are somewhat more likely to be connected with each other; d) their degree distribution is scale-free.

In our work [14], our aim using complex network analysis is to explore the habits of people diagnosed with digestive disease. To analyze the patients we use the PageRank, clustering, eigenvector centrality, and degree centrality. For all networks, we verify the existence of communities using the Girvan Newman and k-Clique algorithms. We used the Networkx library and Graphviz to generate the networks.



X: PubMedSearch (Nom)

Y: Type (Nom)

Colour: Disease (Nom)

Plot: All.Desease

Fig. 8. Ontologies found on each publication filter.

A. PageRank

The PageRank (P) measures the prestige of a node based on the prestige of their adjacent nodes. The PageRank value of a node υ is defined as:

$$P(\upsilon) = \frac{q}{n} + (1-q) \times \sum_{\nu \in B_\upsilon} \frac{P(\nu)}{K(\nu)} \qquad (1)$$

where n is the number of nodes, $K(\nu)$ is the outdegree of a node ν, $\nu \in B_\upsilon$ and q is the damping factor (a probability performance a random walk or a random jump).

B. Clustering

We calculate the clustering coefficient of the nodes, using the following equation:

$$c_\upsilon = \frac{1}{deg(\upsilon)(deg(\upsilon)-1))} \sum_{\nu\omega} (\hat{\omega}_{\upsilon\nu}\hat{\omega}_{\upsilon\omega}\hat{\omega}_{\nu\omega})^{1/3}. \qquad (2)$$

The edge weights $\hat{\omega}_{\upsilon\nu}$ are normalized by the maximum weight in the network $\omega_{\upsilon\nu} = \omega_{\upsilon\nu}/max(\omega)$.

C. Centrality Measures

A centrality measure is a function $c : G(n) \longrightarrow R$, where $c_\upsilon(G)$ is the centrality of node υ in the network G.

Eigenvector centrality computes the centrality for a node based on the centrality of its neighbors. The eigenvector centrality for node υ is the υ-th element of the vector x defined by the following equation:

$$Ax = \epsilon x, \qquad (3)$$

where A is the adjacency matrix of the graph G with eigenvalue ϵ. By virtue of the Perron–Frobenius theorem, there is a unique solution x, all of whose entries are positive, if ϵ is the largest eigenvalue of the adjacency matrix A.

The *degree centrality* [4] of a node v in a network G, denoted $d_v(G) = \|\{\nu : G_{v,\nu} \neq 0\}\|$ is the number of edges (a) connecting this node to others (ν) in G. The degree centrality measure is defined by:

$$C_D(v) = \sum_{\nu} a_{v,\nu}. \tag{4}$$

D. Dataset

In the network analysis, we grouped two datasets provided by the Centers for Disease Control and Prevention (CDC). We grouped the data from 2008-2009 and 2009-2010 available in the National Health and Nutrition Examination Survey (Nhanes) [3], comprising 19,142 participants. We used only the following questionnaires: demographic, bowel health, alcohol use, current health status, diet behavior and nutrition, ferritin exams, and consumer behavior. The Sqlite library was used to store the database, and it is possible to look at the network analysis using Jupyter.

The demographic questions comprise information about gender, ethnicity/race, and age. The bowel health questions comprise personal interview data on fecal incontinence and defecating function. The alcohol use questions focus on lifetime and current use of alcohol. The current health status has questions about overall health assessment, quality of life, and illnesses. The consumer behavior questions involve personal interview data on various dietary related consumer behavior topics at a family level. Diet and behavior nutrition provides personal interview data on various dietary behavior and nutrition related topics, such as allergies and junk food consumption.

Considering these questionnaires, only a subset of questions were relevant. We selected only 1,610 patients that have some digestive disease and based on the answers for each questionnaire, a binary profile was constructed.

E. Network Analysis

Based on the database created, we executed some queries and generated a set of networks to represent the habits of patients with digestive diseases, and a network connecting patients that have and do not have the diagnosis based on common features.

Figure 9 shows the network representing the connection of profiles of alcohol use and current health status. In this network, each node is a profile (alcohol use or current health status). The connections between pairs of profiles were established if patients had both profiles, and the weight of edges was the number of patients with them. The measures adopted were the degree centrality (the size of nodes), and eigenvector centrality (color intensity, where the lighter the color, the higher the centrality). The degree centrality measures the number of relationships of a node, where the node with the highest centrality is 0000000 representing the alcohol use. The node 0000000 represents patients with no alcohol problems, and it interacted with more health status profiles than the other alcohol profiles, meaning it is common for people to have these diseases and not consume alcohol. The edge of the network is concentrating almost all health profiles, while the core concentrates the alcohol profiles, this phenomenon occurs because the number of different alcohol profiles is lower than the health status ones, where the health status node with the highest centrality is 1000001, meaning that the profile where the physical health was not good is the most relevant. The node with the highest eigenvector centrality is more relevant than the other nodes (profiles), because its neighbors have high centrality, meaning that this profile has more impact on the network.

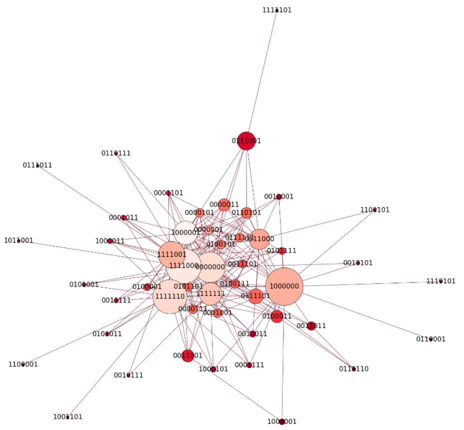


Fig. 9. Network representing alcohol use and current health status profiles.

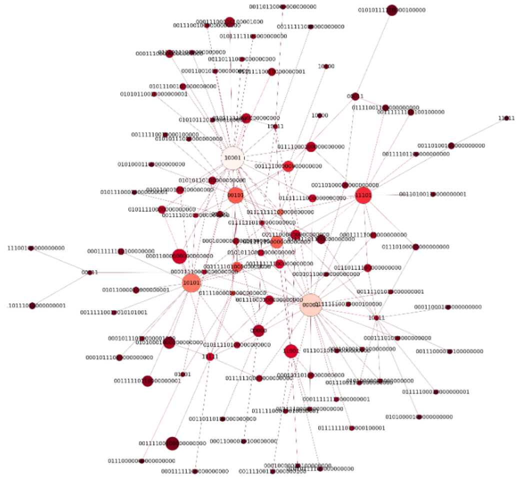


Fig. 10. Network representing nutrition use and consumer behavior profiles.

Figure 10 shows the network that correlates nutrition use and consumer behavior profiles. Each node represents a profile (nutrition or consumer behavior) and an edge is established between two profiles if there are patients with both profiles. The degree centrality and eigenvector centrality measures were used. The degree centrality represented by the node size shows that more patients have the profile 10001, where the relevant information is that these patients consume soft drinks. Some studies concluded that this drink can cause some digestive problems. The eigenvector measure resulted in the same node having higher measure compared to the others, which shows more attention to soft drink consumption. Consumer behavior nodes have higher eigenvector centrality compared to the nutrition nodes centrality values, showing that these.

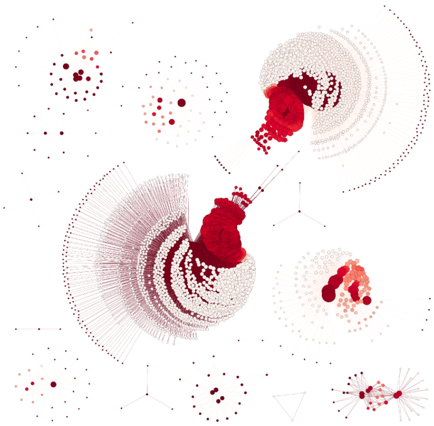


Fig. 11. Network representing the relationship between patients with and without digestive diseases.

Figure 11 shows the network relationship between patients with and without digestive diseases. In this network, each node represents a patient and a connection is established between patients with the same bowel problem. The measures applied were the degree centrality, representing the size of nodes, and the clustering representing the colors. In this network, the highest nodes are nodes representing patients with digestive diseases, the size represents that these patients have more problems in common with others. The network result shows that bowel problems are the most important feature to classify patients with and without digestive diseases.

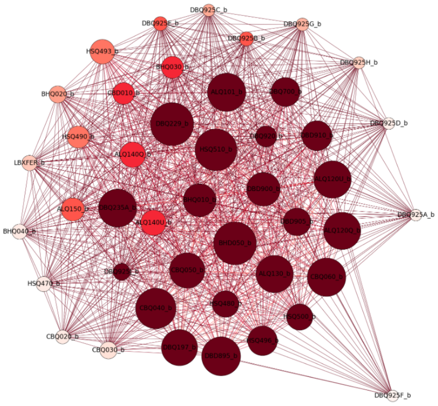


Fig. 12. Network representing the relation between questions answered by patients with digestive diseases.

Figure 12 shows the network representing the relation between questions answered by patients with digestive diseases. In this network, each node is a question (feature) and a connection was established if a patient has answered these questions with abnormal values, for example, the ferritin values of a patient is lower than 20 ng/mL and this patient has a poor diet, then a connection was established between them. The measures applied were the PageRank, representing the size of the nodes, and clustering representing the colors. Considering the relevant features it is possible to create a specific questionnaire to help doctors with digestive diagnosis, and doctor-patient relationship where the diagnosis is already known. The clustering results grouped by the interaction between features (nodes), formed sub-graphs, which can help to establish relations between profiles, for example, the nodes in the edge of the network formed clusters, and the ferritin exam ($LBXFER_b$) formed its own cluster.

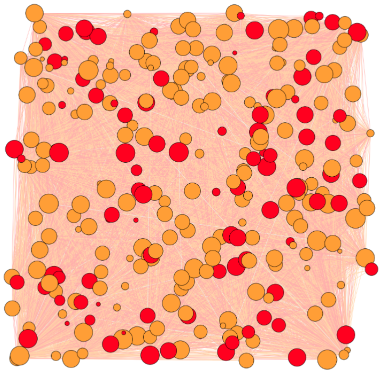


Fig. 13. Network representing symptoms relationship between patient with and without diagnosis with digestive disease.

Figure 13 shows the network representing symptoms relationship between patient with and without diagnosis with digestive disease. In this network, each node is a patient with and without digestive disease, a connection is established between a pair of nodes if one of the following conditions are satisfied: bowel problems; bowel problems and pain; bowel problems, pain, and anxiety. The measure used in this network is degree centrality represented by the size, where the red nodes represent people with digestive disease and orange nodes people without it. The patients with high risk to have digestive diseases are represented by bigger nodes compared to the others. This network can help to identify patients that are not diagnosed with digestive disease but have the potential to have it.

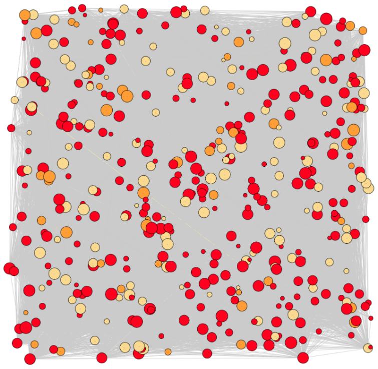


Fig. 14. Network representing the interactions of abnormal exams results from patients diagnosed with digestive diseases.

Figure 14 shows the network representing the interactions of abnormal exams results (features) from patients diagnosed with Ulcerative Colitis (UC), Crohn's disease, or Inflammatory

Bowel Disease Unclassified (IBDU). In this network, each node represents a patient and each connection represents a common abnormal exam result between the nodes. This network is strongly connected, meaning that patients have a high number of common symptoms, thus increasing their degree. The measures used in this network are the degree centrality, representing the size of the nodes, and the colors represent the diseases: UC is orange, Crohn's disease is red, and IBDU is light yellow. Patients with Crohn's disease or IBDU are more connected to the others, meaning they have more abnormal exam results in common with the others.

IV. Machine Learning

In the last decade, there is an ongoing application of Machine Learning (ML) algorithms for the study of diseases. In this work we used unsupervised learning and supervised machine learning algorithms. Unsupervised algorithms are helpful to identify clusters in the data and supervised learning, which helps to make classification based in previous label dataset. We used supervise learning to find a model that deals with the classification task to diagnosis CD or UC.

A. Dataset

We use endoscopic and histological data collected from 287 patients [9]; 178 patients with Crohn's disease (CD), 80 patients with Ulcerative Colitis (UC) and 29 patients with Inflammatory Bowel Disease unclassified (IBDU). Notice that the diagnose of IBDU was label when it not clear label as CD or UC. The remaining 48 patients (CD = 35, UC = 13) were used to validate the model. The Clinical observations (features) were converted into numerical variables $[-1, 0, +1]$ depending on tissue abnormalities. For each patient, abnormal tissues observations were coded as $+1$ and normal as -1. The values of zero were assigned for missing data. Although mouth and perianal locations are not typically biopsied for histology, they were excluded in the previous work [9], and we included just to check their importance.

TABLE II
Features of Clinical Exams

Exams	**Endoscopy**		**Histology**	
Features	Mouth	A-Colon	Mouth.1	A-Colon.1
	Oesophagus	T-Colon	Oesophagus.1	T-Colon.1
	Stomach	D-Colon	Stomach.1	D-Colon.1
	Duodenum	Rectum	Duodenum.1	Rectum.1
	Ileum	Perianal	Ileum.1	Perianal.1

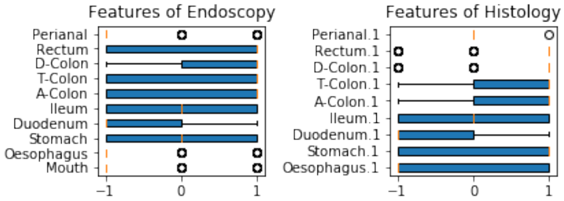


Fig. 15. The boxplot of features for clinical examinations the circles mean outlier and — mean median.

Table II shows the endoscopy and examination histology have the same features. However, they have different distribution as we can see in the Figure 15. These two clinical examinations are well known for their accuracy. The combined model endoscopy and histology get a good performance can be seen in the table with an accuracy of 82.7%, see the table Table III .

TABLE III
Accuracy of Clinical Examinations

Exams	Accuracy %	Precision	Recall	F1-score
Endoscopy (E)	71.0% (0.78)	0.89	0.68	0.75
Histology (H)	76.9% (0. 82)	0.81	0.86	0.83
Combined (E+H)	82.7% (0.87)	0.91	0.83	0.87

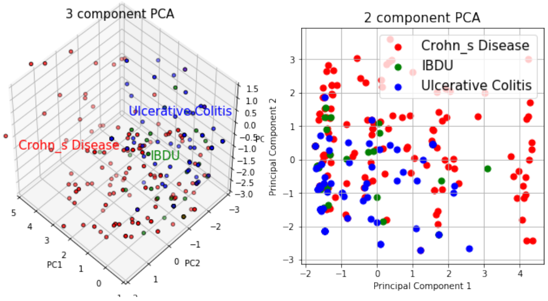
On previous work [10], the authors used Decision Trees, Linear discriminant, Linear SVM (Support Vector Machine), Quadratic SVM, Cubic SVM, Boosted Trees and Bagged Trees in the classification of IBDs. These techniques combined get an 83.3% of accuracy, as can see in the Table IV.

TABLE IV
Accuracy of Current Methods

Current Methods[6]	**Accuracy**
Simple Tree (4 splits)	78.10%
Medium Tree (20 splits)	75.20%
Complex Tree (100 splits)	76.70%
Linear discriminant	81.00%
Linear SVM	80.50%
Quadratic SVM	78.10%
Cubic SVM	73.80%
Boosted Trees	74.80%

B. Unsupervised learning

1) Principal Component Analysis: The PCA algorithm we used to show and helps to figure out if there is any linear feature association between cases (patient) and traits. Figure 16 shows that there is not clear a division between Crohn's disease and Ulcerative Colitis.



(a) (b)

Fig. 16. Principal component analysis, the picture (a) on the left corresponds PCA with 3 components and the picture (b) on right corresponds a PCA with 2 components.

2) Hierarchical Clustering: Figure 17 shows the results of the unsupervised Clustering, where occurs an overlap between Crohn's disease and Ulcerative Colitis also the inflammatory bowel disease unclassified (IBDU) is scattered across different clusters. The classification of CD and UC do not get to form a cluster. It is clear with the Hierarchical Clustering that patients do not group into CD, UC and IBDU classification.

Something to point out is that Hierarchical Clustering show as a different set of four classifications in the diagnosis of IBD. There are some difference and similarities in abnormal features that describe in the Table V.

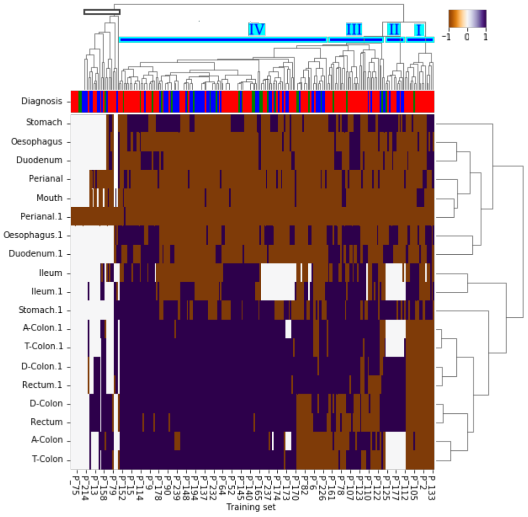


Fig. 17. Hierarchical Clustering with metric hamming and method average.

TABLE V
FEATURE ABNORMALITY DIFFERENCE AMONG FOUR CLASSES

Abnormal	Class I	Class II	Class III	Class IV
Mostly		D-Colon.1		
		Rectum.1		
		Ileum	A-Colon.1	
		Ileum.1	T-Colon.1	
		D-Colon		
		Rectum		
Sometimes	Oesophagus.1		D-Colon	
	Duodenum.1		Rectum	
	Stomach.1		A-Colon	
			T-Colon	

C. Supervised learning

For the training dataset, we considered only the trails where the CD or UC were diagnosed, and for the testing dataset we used the same validation dataset used by [10].

1) Multi-layer Perceptron: Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns from a function, using a dataset to train it. We explored different architectures for multi-layer perceptron. One model with one layer and 50 neurons, another with 3 layers 16 neurons in the input layer, 16 in the hidden layer and 17 in the output layer. In both cases Rectified Linear Unit (ReLU) was used as activation function and Adam as optimizer Adam that is a stochastic gradient-based optimizer.

2) Support Vector Machine: Support Vector Machine (SVM) is used for finding an optimal hyperplane that maximizes the separation margin between classes. SVM's are most used for classification problem. They can also be used for regression, outlier detection and clustering. SVM works great for small datasets, such as our case where we have 287 patients.here we compare different kernels using the library scikit-learn. So we compare the Linear SVC is implementation of SVM , SVM non linear RBF Kernel with parameters of $\gamma = 0.7$ and $C = 1$. As well SVM with kernel polynomial (degree=5 and $C = 1000$).The performance of this different SVM kernel can see in the Table VI

3) Gaussian Processes: Gaussian Processes (GP) are a generic supervised learning method designed to solve regression and probabilistic classification problems by maximizing the log-marginal-likelihood (LML) .

4) K-Nearest Neighbor: The K-Nearest Neighbor (KNN) algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. A set of training samples finds a new point in the near distance, and predict the label from this set. The distance can be the standard Euclidean distance that is most common or other measure.here we chose $K = 4$.

5) Ensemble Vote Classifier: The Ensemble Vote Classifier is multiple classifier system for joining comparative or reasonably good ML classifiers for arrangement through dominant part (Soft voting) or majority casting a ballot win (Hard voting). So we use the Soft Voting Ensemble (SVE) because we want to get the better performance that each classifier gets in some set.

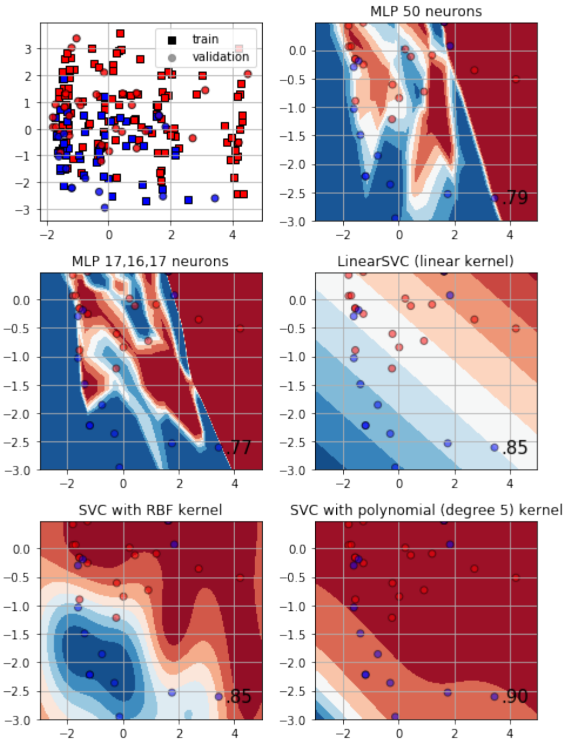


Fig. 18. Decision boundary classifiers MLP and SVM, the accuracy is in down right corner.

In the Soft Voting Ensemble we assigned a weight of (3,2,2) to the classifiers KNN ($K = 4$), GP and SVM $poly^5$ respectively. Notice that we get an overall accuracy of 85% less than the accuracy of SVM $poly^5$ 90%. This happen

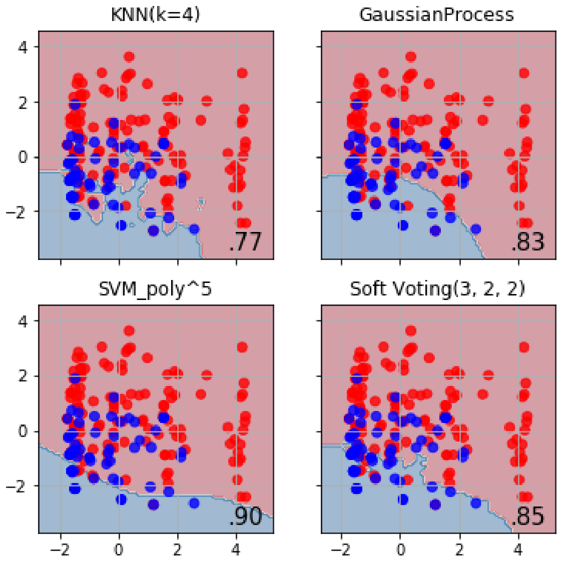


Fig. 19. Decision boundary classifiers: *K*NN, GP, SVM, and Soft Voting, the accuracy is in down right corner.

Because the classifier *K*NN ($K = 4$) has the lower accuracy of 77%, as can see in the Table VII We still use *K*NN ($K = 4$) due to its ability to deal with non linear classification .

D. Results

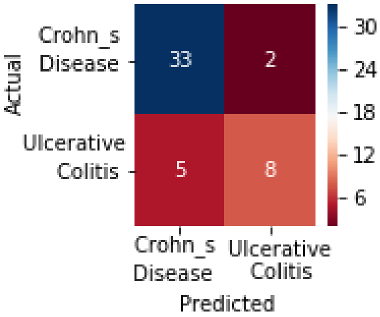


Fig. 20. Confusion matrix of the model soft voting.

TABLE VI
ACCURACY OF DIFFERENT METHODS

Other Methods	Accuracy
MPL(50NN)	79%
MLP(17$NN \rightarrow 16NN \rightarrow 17NN$)	77%
SVC (Linear kernel)	85%
SVC (RBF Kernel)	85%
SVC (polinomial^5 Kernel)	90%

The performance of Soft Voting Classifier was very good with an Accuracy of 85% as can see in the Table VIII and Figure **??** that show the confusion matrix. Although Sensitivity (Recall or True positive rate) 94% was good , the Specificity (True negative rate) 61% was poor, this mean that it more difficult to predict if a patient has UC.

Data manipulation and modeling were performed using The computational tools that we will use were Python (Numpy, Pandas, and Scikit-Learn) [5], Jupyter, and Google Colab (https://colab.research.google.com) that we applied our machine learning study [17]

TABLE VII
ACCURACY OF MODEL SELECTED TO ENSEMBLE A SOFT VOTING

Model Selected	Accuracy
*K*NN	77%
Gaussian Processes	83%
SVC polynomial 5	90%
Soft Voting(3,2,2)	85%

TABLE VIII
REPORT OF CONFUSION MATRIX FOR SOFT VOTING

Soft Voting(3,2,2) Combined	
(KNN=4, Gaussian Process, SVM_poly^5)	
Accuracy %	85.41
Precision	0.87
Recall	0.94
F1-Score	0.90

Machine learning algorithms have proven very helpful, with the unsupervised learning in both cases using principal component analyst and hierarchical clustering shows that there is great complexity in the diagnosis of IBD since there is not clear difference in features in CD and UC. More over it has seem that IBD has to diagnosis into four classes, that could help to develop more personalized and effective prescription in drug treatment as well food diet. In the case of supervised learning the ensemble method of soft voting (SVM, GP, or *K*NN) gets an accuracy of 85%, bigger than accuracy of Endoscopic and histological combined.

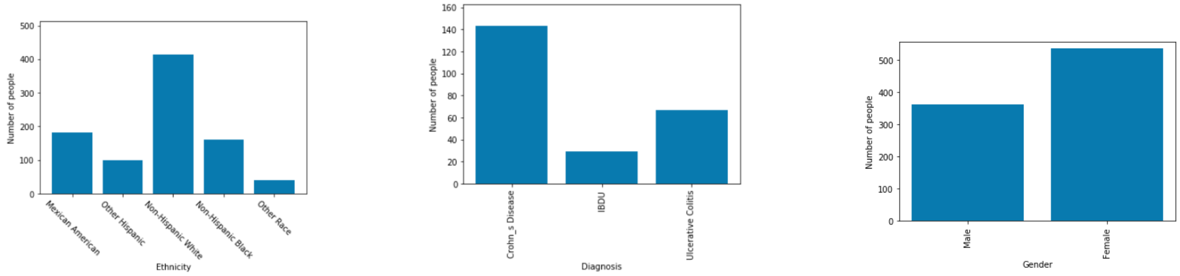
The Machine learning tools have a great potential for development of automatic recommend systems for heath to assist physician in the diagnosis and provide and personalized treatment [1].

V. STATISTICS

In this Section we explore some statistics that can help doctors to investigate patients with IBDs. These statistics were got from datasets described in Section III and Section IV.

Figure 21 shows a set of histograms representing the information about gender, ethnicity/race, and the percentage of people diagnosed with three diseases. The results showed that from the people with digestive disease, 59.7% are female and 40.3% are men. In the second chart, non-Hispanic white patients represent the highest number of people diagnosed with some IBD. The Patients with CD represent the highest percentage compared to the other diseases. Based on these charts we conclude that it is more common for non-Hispanic females to have digestive diseases, CD being the most common one.

In [11], the authors studied the iron deficiency in patients with IBD. Figure 22 shows the relationship between ferritin and consumer behavior, considering only patients with digestive diseases. In the first chart, 38.35% (orange) of patients do not have healthy diet, and 18.77% (orange) have ferritin problems, where the levels are too high or too low. Correlating these information, 53% of patients need to investigate the



(a) (b) (c)

Fig. 21. General information: demographic and percentage of each disease.

cause of this, since they have a healthy diet and still have bad ferritin levels.

The results obtained from our study [16] (Figure 24) shows that 24.86% of patients with digestive diseases have pain, making it hard for usual activities, and from these patients 96.46% have pain and also have bowel problems, which means there is a strong relationship between pain and bowel problems. Another statistic analysis show that 14% of patients have allergy, more specifically: wheat 1.86%, cow's milk 19.88%, eggs 4.97%, fish 3.73%, shellfish 16.77%, corn 1.24%, peanut 6.83%, nuts 3.11%, and 41.61% other allergies. The results are important for doctors to establish a relation where the patient is still having bowel problems while receiving the correct treatment.

VI. Network and Machine Leaning Clusters

Figure 24 shows the relationships between histology and endoscopy exams where the results are abnormal. In the network, nodes are the features (exams) and edges are established if the patient has a pair of abnormal exams, such as, mouth and duodenum exam results are abnormal, then a connection is established between them. The weight of such edge is the number of patients that have the same abnormal pair of results. The network analysis used PageRank and clustering metrics. The results of network indicate that the $Rectum_h$ (histology exams group) feature has more prestige than the others, thus this feature could determine the patient's problem. The feature with low PageRank is mouth exam (endoscopy group), accounting for less than 1% of patients. The network formed four clusters, where the dark red/ruby cluster has more features and these features have nodes with high PageRank. Based on this, a future work could use only these features to train and test machine leaning algorithms to evaluate if the accuracy is improved.

The second figure (Figure 24) on the right, show the Hierarchical Clustering an unsupervised machine learning technique. We found that IBD in cluster I and II the sometimes the features are Oesophagus.1, Duodenum.1 and Stomach.1 were abnormal and the cluster III has sometimes features D-Colon, Rectum, A-Colon and T-Colon. The clustering II, II and IV have mostly abnormal features such as D-Colon.1 and Rectum.1. But the cluster II has abnormal features Ileum, Ileum.1, D-Colon and Rectum. The clusters III and IV have abnormal features A-Colon.1 and T-Colon.1. Notice that the supervised classification of IBD into CD and C are distributed in the cluster I, II, III and IV. This also explain the difficulty for physicians to diagnosis IBD as CD or UC.

VII. Conclusion

In this work, we proposed to analyze digestive diseases using different techniques and datasets which provide different information from patients.

The ontologies show us the main areas which have more research related to each disease since they were found on Pubmed citations. The ontology results can help us get information to classify questions in order to generate the networks and machine learning analysis, and in our work we used it to verify some studies and classify the questions that were relevant to create a subset of questions.

Networks provide us with a set of habits that need to be explored, such as nutrition, mental/physical health and bowel information. These networks are relevant to analyze and map some habits that decrease the quality of life of patients with IBDs. For future work, the relevant features extracted from networks could be used in machine learning analysis using an new extend data set called the 1000IBD project [7].

The machine learning results provide us with an extra vision of these patients, based on their exams, which can be complemented with the results obtained from the network analysis, creating a framework that can help doctors and other health professionals.

The statistics give us a dimension of demographic information, allergies, and degree of pain and bowel problems. Moreover, more analysis considering the methods introduced in this work are available in [15].

References

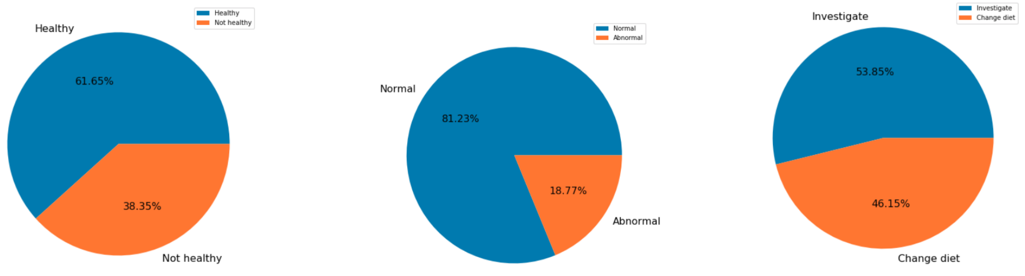
[1] James J. Ashton, Enrico Mossotto, Sarah Ennis, and R. Mark Beattie. Personalising medicine in inflammatory bowel disease—current and future perspectives. *Translational Pediatrics*, 8(1), 2019.

[2] Frances A Brahmi and Carole Gall. Endnote® and reference manager® citation formats compared to "instructions to authors" in top medical journals. *Medical Reference Services Quarterly*, 25(2):49–57, 2006.

[3] Centers for Disease Control and Prevention. National health and nutrition examination survey, 2019.

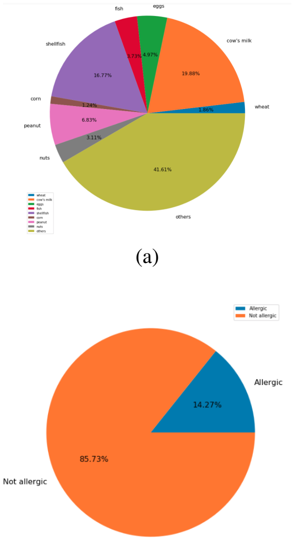
[4] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[5] Aurlien Gron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 1st edition, 2017.



(a) (b) (c)

Fig. 22. Relationship between ferritin and consumer behavior, considering only patients with digestive disease.



(a)

(b)

Fig. 23. Allergies statistics.

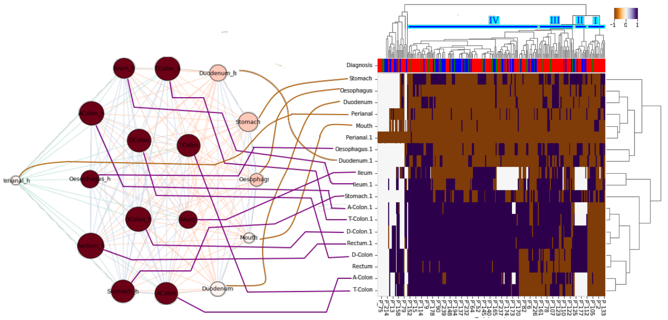


Fig. 24. Network representing the relation in histology and endoscopy exams from patients diagnoses with digestive diseases.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[7] Floris Imhann, K. J. Van der Velde, R. Barbieri, R. Alberts, M. D. Voskuil, A. Vich Vila, V. Collij, L. M. Spekhorst, K. W. J. Van der Sloot, V. Peters, H. M. Van Dullemen, M. C. Visschedijk, E. A. M. Festen, M. A. Swertz, G. Dijkstra, and R. K. Weersma. The 1000ibd project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. *BMC Gastroenterology*, 19(1):5, Jan 2019.

[8] Orges Koci, Michael Logan, Vaios Svolos, Richard K. Russell, Konstantinos Gerasimidis, and Umer Zeeshan Ijaz. An automated identification and analysis of ontological terms in gastrointestinal diseases and nutrition-related literature provides useful insights. *PeerJ*, 6:e5047, jul 2018.

[9] E. Mossotto, J. J. Ashton, T. Coelho, R. M. Beattie, B. D. MacArthur, and S. Ennis. Classification of paediatric inflammatory bowel disease using machine learning. *Scientific Reports*, 7(1), May 2017.

[10] E Mossotto, JJ Ashton, T Coelho, RM Beattie, BD MacArthur, and S Ennis. Classification of paediatric inflammatory bowel disease using machine learning. *Scientific reports*, 7(1):2427, 2017.

[11] Ole Nielsen, Christoffer Soendergaard, Malene Vikner, and Günter Weiss. Rational management of iron-deficiency anaemia in inflammatory bowel disease. *Nutrients*, 10(1):82, 2018.

[12] Steffen Staab and Rudi Studer, editors. *Handbook on Ontologies*. Springer Berlin Heidelberg, 2009.

[13] Gian Eugenio Tontini. Differential diagnosis in inflammatory bowel disease colitis: State of the art and future perspectives. *World Journal of Gastroenterology*, 21(1):21, 2015.

[14] Silvana Trindade, Alexander Valle Rey, and Leandro Zanotto. Complex network analysis. https://github.com/Trindad/digestive-diseases/blob/master/network/network.ipynb, 2019.

[15] Silvana Trindade, Alexander Valle Rey, and Leandro Zanotto. Inflammatory bowel diseases. https://github.com/Trindad/digestive-diseases, 2019.

[16] Silvana Trindade, Alexander Valle Rey, and Leandro Zanotto. Statistics. https://github.com/Trindad/digestive-diseases/blob/master/network/statistics_and_more.ipynb, 2019.

[17] Alexander Valle Rey, Silvana Trindade, and Leandro Zanotto. Machine learning. https://github.com/Trindad/digestive-diseases/tree/master/machine-learning, 2019.