Computer Methods and Programs in Biomedicine 230 (2023) 107320

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Automated analysis of small intestinal lamina propria to distinguish normal, Celiac Disease, and Non-Celiac Duodenitis biopsy images

Oliver Faust[a], Simona De Michele[b], Joel EW Koh[c], V Jahmunah[c], Oh Shu Lih[c], Aditya P Kamath[d], Prabal Datta Barua[e,f,g], Edward J. Ciaccio[h], Suzanne K. Lewis[h], Peter H. Green[h], Govind Bhagat[b,h], U. Rajendra Acharya[i,c,j,*]

[a] Anglia Ruskin University Cambridge Campus, UK
[b] Department of Pathology and Cell Biology, Columbia University Irving Medical Center, USA
[c] Department of Computer Engineering, Ngee Ann Polytechnic, Singapore, Singapore
[d] Brown University, Providence, RI, USA
[e] Cogninet Australia, Sydney, NSW 2010, Australia
[f] School of Management & Enterprise, University of Southern Queensland, Australia
[g] Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia
[h] Department of Medicine, Celiac Disease Center, Columbia University Irving Medical Center, USA
[i] School of Science and Technology, Singapore University of Social Sciences, 463 Clementi Road, 599494, Singapore
[j] Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan

ARTICLE INFO

Article history:
Received 14 August 2022
Revised 16 December 2022
Accepted 18 December 2022

Keywords:
Celiac Disease
Computer-aided diagnosis
Biopsy
Lamina propria
Inflammation
Explainable artificial intelligence

ABSTRACT

Background and objective: Celiac Disease (CD) is characterized by gluten intolerance in genetically predisposed individuals. High disease prevalence, absence of a cure, and low diagnosis rates make this disease a public health problem. The diagnosis of CD predominantly relies on recognizing characteristic mucosal alterations of the small intestine, such as villous atrophy, crypt hyperplasia, and intraepithelial lymphocytosis. However, these changes are not entirely specific to CD and overlap with Non-Celiac Duodenitis (NCD) due to various etiologies. We investigated whether Artificial Intelligence (AI) models could assist in distinguishing normal, CD, and NCD (and unaffected individuals) based on the characteristics of small intestinal lamina propria (LP).

Methods: Our method was developed using a dataset comprising high magnification biopsy images of the duodenal LP compartment of CD patients with different clinical stages of CD, those with NCD, and individuals lacking an intestinal inflammatory disorder (controls). A pre-processing step was used to standardize and enhance the acquired images.

Results: For the normal controls versus CD use case, a Support Vector Machine (SVM) achieved an Accuracy (ACC) of 98.53%. For a second use case, we investigated the ability of the classification algorithm to differentiate between normal controls and NCD. In this use case, the SVM algorithm with linear kernel outperformed all the tested classifiers by achieving 98.55% ACC.

Conclusions: To the best of our knowledge, this is the first study that documents automated differentiation between normal, NCD, and CD biopsy images. These findings are a stepping stone toward automated biopsy image analysis that can significantly benefit patients and healthcare providers.

1. Introduction

Celiac Disease (CD) is an immune-mediated enteropathy caused by a maladaptive intestinal immune response toward gluten molecules in susceptible individuals [1]. The estimated prevalence of CD is 1.4% based on serologic tests and 0.7% based on biopsy test results [2]. In high-risk populations, the prevalence rate increases to 4.7% [3], and areas of high prevalence are sometimes located in developing countries [4,5]. Meta analysis on published studies shows that there was an annual increase in CD prevalence of 7.5% [6]. CD can also cause extra-intestinal disorders, e.g., osteoporosis [7], and mineral and vitamin deficiencies [8]. The symptoms and secondary effects have a detrimental impact on the quality of life of patients and result in high healthcare costs [9]. This negative

* Corresponding author at: School of Science and Technology, Singapore University of Social Sciences, 463 Clementi Road, 599494, Singapore.
E-mail address: aru@np.edu.sg (U.R. Acharya).

https://doi.org/10.1016/j.cmpb.2022.107320

Acronyms

ACC	accuracy
ADASYN	adaptive synthetic sampling
AHE	adaptive histogram equalization
AI	artificial intelligence
CD	Celiac Disease
CLAHE	contrast limited adaptive histogram equalization
DT	decision tree
ECG	electrocardiogram
FN	false negative
FP	false positive
HOG	histogram of gradient
HP	Helicobacter Pylori
IBD	inflammatory bowel disease
KNN	K-nearest neighbor
LP	lamina propria
ML	machine learning
NCD	Non-Celiac Duodenitis
PHOG	Pyramid Histogram of Gradient
PPV	Positive Predictive Value
RBF	Radial Basis Function
SD	Standard Deviation
SEN	sensitivity
SHAP	SHapley Additive exPlanations
SPE	specificity
SVM	Support Vector Machine
TN	true negative
TP	true positive

impact, coupled with the high prevalence of this disease, results in a significant public health problem. The prerequisite for addressing this problem is to develop diagnosis support methods that can establish and track CD cases in a cost-effective manner. These methods must be safe, reliable, and functional [10]. Having such methods will enable the detection of more CD cases with the same resources and facilitate earlier treatment. Hence, it is essential to create innovative CD detection methods which are cost-effective and readily integrate with standard diagnostic pathways.

CD can cause extra-intestinal disorders, e.g., osteoporosis [7], and mineral and vitamin deficiencies [8]. The symptoms and secondary effects have a detrimental impact on the quality of life of patients and result in high healthcare costs [9]. This negative impact, coupled with the high prevalence of this disease, results in a significant public health problem. The prerequisite for addressing this problem is to develop diagnosis support methods that can establish and track CD cases in a cost-effective manner. These methods must be safe, reliable, and functional [10]. Having such methods will enable the detection of more CD cases with the same resources and facilitate earlier treatment. Hence, it is essential to create innovative CD detection methods which are cost-effective and readily integrate with standard diagnostic pathways.

Current diagnostic paradigms incorporate objective methods for detecting CD [11]. These methods often rely on manual or semi-automated optical or histological image analysis. Manual analysis is time-consuming because a diagnosis can only be established by analyzing all the available evidence. In this case, the evidence must be discovered by scanning images of the small intestine, the longest organ of the human digestive system [12–14]. Optical images are acquired via traditional or video capsule endoscopy, which are analyzed by gastroenterologists to detect and grade CD-associated mucosal changes [15]. Analysis of duodenal biopsy specimens is currently the internationally accepted gold standard for CD diagnosis in adults [13,16]. In this procedure, biopsy samples, obtained via an endoscope, are microscopically examined by pathologists to determine the presence of increased intraepithelial lymphocytes, crypt hyperplasia, and villous atrophy, and the changes are semiquantitatively graded, often using the Marsh scoring system or a modified scheme introduced by Oberhuber et al. [17,18]. Experimental studies show that Machine Learning (ML) methods, which are cost effective, can be used for histopathologic feature analysis in a variety of tissues for diagnostic and prognostic purposes [19]. However, only a few studies have applied such computational approaches to CD detection [20–25]. These approaches focused on discerning normal from CD biopsy images or differentiating clinical CD stages based on differences in mucosal architecture. Non-Celiac Duodenitis (NCD) does not feature as a distinct data class; currently, this failure mandates the need for an expert to manually rule out non-celiac-related inflammation prior to any computer-aided diagnostic support procedure. The additional human analysis step can increase the subjectivity of the diagnosis.

The histologic changes of CD, including alterations in the epithelial and Lamina Propria (LP) compartments, result from deregulated innate and adaptive immune responses induced by gluten. Likewise, immune and inflammatory mechanisms also underlie the mucosal abnormalities in NCD, which differ depending on the etiology. Changes in LP constituents, including inflammatory cells, are recognized in CD and NCD, with some differences having been reported between CD and certain NCD entities [26–30]. However, to date, computational, image-based analysis of the small intestinal LP in CD and NCD (and comparison with non-inflamed LP) has not been performed to determine if this approach has any diagnostic utility.

This paper, proposes a novel ML algorithm for the automated analysis of small intestinal biopsy imagery, focusing on LP inflammatory cells in normal duodenum, CD, and NCD. The ability of Artificial Intelligence (AI) models to distinguish CD from normal controls and NCD based on the LP cellular composition appears promising and might suggest differences in the types, density, and distribution of inflammatory or stromal cells in different small intestinal inflammatory diseases. The ability to discern between normal, NCD, and CD by analyzing the LP changes will be a significant step towards creating a fully automated small intestinal biopsy analysis system that can assist pathologists whenever the villous architecture cannot be reliably assessed. Additionally, interpretation of the best classifier performance can be done by determining the best feature(s) that can influence the classifier to yield highest classification results via AI methods. Also, using morphometric analysis of small intestinal mucosa, prior studies highlighted differences in the LP volume and cell type of untreated celiac mucosa compared to normal controls. However, automated, computational analysis of differences in the LP cellular contexture of CD and NCD or normal biopsies has not been performed. In the current study, we extend our work by assessing the NCD class of biopsy images and determining whether the LP cellular/inflammatory milieu can help discern different etiologies of intestinal inflammatory disorders and discriminate between inflamed and uninflamed mucosa. Incorporating the NCD class in an automated detection system for CD is important to ensure an accurate diagnosis of CD, as some of the histopathologic features of CD and NCD may overlap. Our current methodology can differentiate normal versus CD and NCD biopsy images. Hence. the main goal of this study is to investigate seven use cases of different combinations (varying LP cellular compositions based on differences in the types, density, and distribution of inflammatory or stromal cells) of binary classification schemes to determine the classification which obtains the highest accuracy with our proposed classifiers.

The next section provides background on the Marsh scoring system used to classify the biopsy images. Section 3 describes the individual methods used during the study design. Section 4 doc-

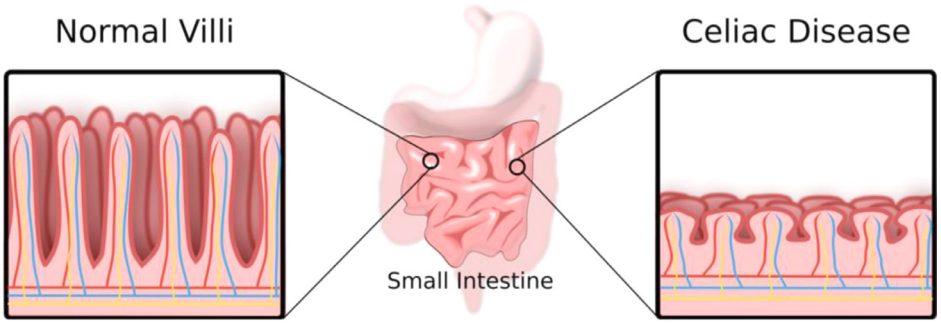

Fig. 1. Small intestine showing normal villi and crypts (left) and celiac-disease-associated villous atrophy and crypt hyperplasia (right).

Fig. 2. Flowchart documenting the sequence of algorithms used to process and analyze the image data.

uments the classification performance of the ML model for each of the binary classification problems. The results of our efforts are documented as system performance measures in Section 5. We discuss our work in relation to previous studies in Section 6. This section also contains limitations and future work. The final section of the paper provides conclusions of our work.

2. Background

The mucosal architecture of the small intestine differs in normal versus pathologic conditions. Fig. 1 shows normal small intestinal villi versus those from an untreated CD patient with villous atrophy. In CD, the villi tend to be shortened and blunted compared to healthy villi, and the crypts are hyperplastic.

The classic histopathologic changes of CD in the small bowel are categorized by the "Marsh classification" [31], which was introduced in 1992, and subsequently modified by Oberhuber et al. in 1999 into six stages, the previous stage 3 being split into three substages [18,32].

- Marsh 0: normal mucosa
- Marsh I: increased number of intraepithelial lymphocytes, usually exceeding 20 per 100 enterocytes
- Marsh II: hyperplasia of the crypts of Lieberkuhn with preservation of villous architecture.
- Marsh III: villous atrophy accompanied by crypt hyperplasia.
  - IIIa: partial villous atrophy,
  - IIIb: subtotal villous atrophy,
  - IIIc: total villous atrophy.

3. Methods

This section outlines the methods used to design a ML model that can discriminate biopsy images of the LP compartment from controls, NCD patients, and CD patients across the Marsh spectrum I, II, IIIa, IIIb, and IIIc. During the design, we adopted a traditional feature engineering approach. The design was structured into pre-processing, feature extraction, and classification. The flow chart in Fig. 2, documents the data processing steps, which establish the system's functionality. The following sections describe the algorithms involved in these processing steps.

3.1. Data acquisition

The images utilized in this study were acquired from duodenal biopsies of 31 controls without any small intestinal disorder, 45 celiac patients, and 20 with NCD, including non-specific duodenitis, Inflammatory Bowel Disease (IBD), Helicobacter Pylori (HP) infection and autoimmune enteropathy, diagnosed at Columbia University Irving Medical Center in New York. From these 96 cases, a total of 284 digital images of biopsies were acquired with a slide scanner (Leica Aperio AT2, Buffalo Grove, IL). Our goal was to include as many images of the lamina propria with diverse inflammatory and non-inflammatory cells as possible, selected randomly, while minimizing the presence of epithelium in the same image. Representative photomicrographs of the LP compartments were obtained at 40× magnification with the Aperio Image scope v12.4.0.7018. The biopsies, from which the photomicrographs were obtained, were classified by two pathologists as: normal, NCD, Marsh I, Marsh II, Marsh IIIa, Marsh IIIb, or Marsh IIIc. Table 1 details information concerning the acquired data. Fig. 3a depicts a representative example of a biopsy image taken from the normal or control set. Fig. 3b depicts an example of an image from the NCD set. Fig. 3c–3g show examples for Marsh I, Marsh II, Marsh IIIa, Marsh IIIb, and Marsh IIIc, respectively. Once the images were obtained and labeled, they were individually pre-processed to enhance and standardize image quality.

3.2. Pre-processing

The Matlab software was used for the machine learning technique employed in this study. The acquired biopsy digital photomicrographs were stored as grayscale images and as three additional color images from the red, blue, and green channels. The images were pre-processed using the Contrast Limited Adaptive Histogram Equalization (CLAHE) method [33]. CLAHE is an adaptable contrast improvement technique based upon Adaptive Histogram Equalization (AHE) [34], wherein a histogram is computed for the

Table 1
Number of subjects and number of images per type.

	Control	NCD	Marsh I	Marsh II	Marsh IIIa	Marsh IIIb	Marsh IIIc
Subjects	31	20	7	6	10	10	12
Images	91	58	21	18	30	30	36

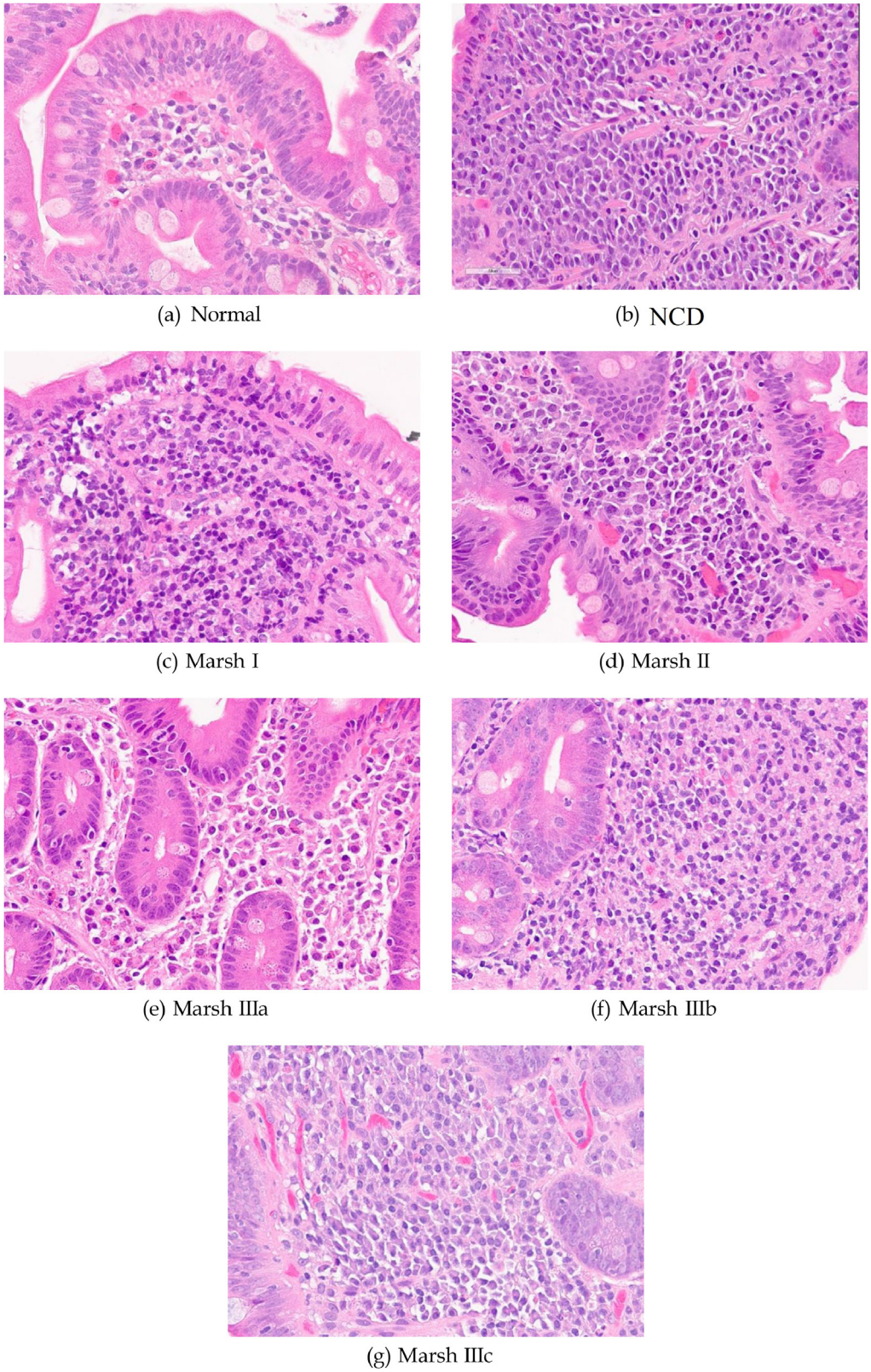

Fig. 3. Example figures for the seven image classes. The Marsh scoring for each histologic slide is labeled.

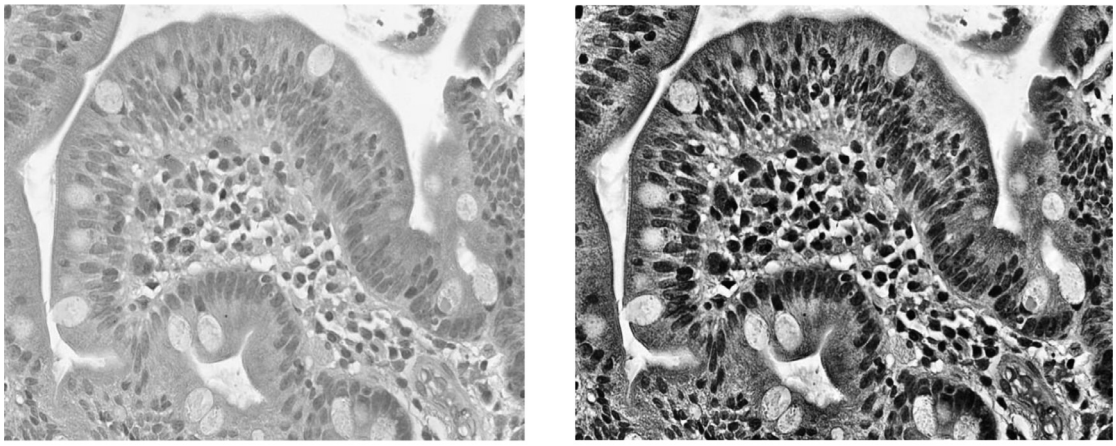

Fig. 4. Original (4a) and CLAHE processed (4b) images from the normal control class (normals). CLAHE implements histogram equalization for enhancing the local contrast of an image.

associated region of a pixel [33]. This contrast boosting technique enhances the difference between intensities of pixels that are located in close proximity [35]. In addition, CLAHE is an enhancement of the AHE, wherein the enhancement computation is altered by adding a user-specified maximum to the height of the local histogram and to the maximum contrast enhancement factor [33]. The CLAHE parameters used are described as follows; contrast enhancement limit: 0.01, number of histogram bins used to build contrast enhancing transformation: 256, desired histogram shape: uniform, distribution shape: uniform, distribution parameter: 0.4.Fig. 4 depicts the original and CLAHE processed images of the normal class. Subsequently, the images were re-scaled to a standard size of 878 $\times$ 1252. The pre-processing method used in this study is specific to the image data used in this study.

3.3. Feature extraction from pre-processed images

The Histogram of Gradient (HOG) technique works by counting the occurrence of gradient orientation in an image [36]. Thus, the local appearance of objects is described using the distribution of edge directions determined by gradient assessment. In Pyramid Histogram of Gradient (PHOG), the spatial layout of the image is preserved by dividing the image into sub-regions at multiple resolutions and applying the HOG descriptor to each sub-region [36]. The number of sub-regions depends on the image size and resolution. The Canny edge detector is applied, and the histogram of orientation gradients is then calculated for all bins in each level. The histograms are then combined to form the PHOG representation of the input image.

The distance in a coordinate system between two different images is known as the PHOG distance [37]. This distance is computed by partitioning the original image into four equal-sized regions and then estimating the HOG feature for each region. The next step is to obtain the HOG region to be calculated. The previous four regions are further reduced to four additional sub-regions to calculate the same. l is the level at which the division and calculation steps are done. The pyramid sub-regions consist of $4l$ units while maintaining the global image at a value of $l = 0$. Therefore, the total HOG is equal to the summation of the $4l$ sub-regions, considering all previous regions [37]. k represents the equidistant intervals, which signifies the binned orientation that has been normalized, and represents the corresponding HOG. Hence, in this study, $k = 16$ bins, an angle of 360°, and 3 levels of the pyramid were used to compute PHOG.

Table 2
Performed binary classification schemes.

Number	Combination
1	Normal versus CD with Marsh I+II
2	Normal versus CD without Marsh I+II
3	Normal versus NCD
4	Normal versus NCD + CD with Marsh I+II
5	Normal versus NCD + CD without Marsh I+II
6	NCD versus CD with Marsh I+II
7	NCD versus CD without Marsh I+II

3.4. Feature selection and classification

The first feature selection step was to concatenate all the PHOG features extracted from the same image to form a feature list. Next, we grouped the feature lists according to the label of the image from which the feature list entries were extracted. For example, 91 feature lists were grouped to form the normal (control) set. As a result, we established seven sets of feature lists: Control, NCD, Marsh I, Marsh II, Marsh IIIa, Marsh IIIb, and Marsh IIIc. After that, we were able to establish binary classification problems that reflect the use case scenarios. Binary classification refers to the fact that a classification algorithm is tasked with deciding which one of two possible groups a particular feature vector belongs to. Table 2 describes the data arrangements for the binary classification problems, which reflect the seven use case scenarios. For example, in the third classification problem, we ask the classification algorithm to differentiate between control and NCD. That reflects the use case scenario of deciding whether a biopsy image is normal or shows NCD symptoms. Student's t-test [38] was used to guide the feature selection process. To be specific, each binary classification problem involves two sets of feature lists, and Student's t-test was used to calculate p- and t-values from all the listed features. As such, both p- and t-value measure the ability of a feature to discriminate between the classes.

Higher t-values indicate better performance. Therefore, feature ranking was accomplished by arranging the feature sets in descending order with respect to the t-value measure. Table A.12 in the Appendix reports the mean and Standard Deviation (SD) as well as p- and t-values of extracted features for the normal (control) and NCD classes. The table entries were also arranged in descending order based on the t-value measure. To limit the space requirement for that table, we display only features with a t-value that is greater than 6. Similar tables were created for all the seven

binary classification problems. That allowed us to establish which features work best for a particular classification problem. The best features were combined to form feature vectors which can be used to train and test nine classification algorithms, such as the Decision Tree (DT) [39], K-Nearest Neighbor (KNN) [40], Support Vector Machine (SVM) [41] (with Radial Basis Function (RBF) and 1st, 2nd, as well as 3rd polynomial kernels), and ensemble classifiers [42] (AdaBoost, bagged tree, and subspace). Each result was established according to the rules of k-fold cross-validation [43] ($k = 3$), repeated ten times to obtain the most accurate results.

Before we could start to train and test the classification algorithms, we had to address the problem of imbalanced data. Data imbalance refers to one class-specific feature set having significantly more entries. For example, we have created a binary classification problem for the third use case scenario to differentiate between normal and NCD cases. The normal feature data was extracted from 91 images, whereas the NCD data came from only 58 images. Hence, the data was biased towards normal. To illustrate that bias, suppose a dumb classifier that labels every feature as normal. Such an impractical setup would be correct $91/(91 + 58) \times 100 = 61\%$ of the time. In a practical setting, the classification model would over-report normal cases. This over-reporting of the class with more training data also exists for models that were established through training and testing regimes like 3-fold cross-validation. Therefore, it is vital to reduce the training bias by balancing the data. For this study, we have used the Adaptive Synthetic Sampling (ADASYN) [44] approach, which uses a weighted distribution for the minority classes to balance data. ADASYN works by generating more synthetic data for the minority class data that are more challenging to learn by the model as compared to those that are easier to learn. Hence, learning is enhanced wherein the biased brought about by the imbalance is reduced and the classification decision boundary is altered by shifting toward the more challenging data [44]. ADASYN was employed per fold during the 3-fold validation of models. During testing, we established the performance measures of Accuracy (ACC), Positive Predictive Value (PPV), Sensitivity (SEN), and Specificity (SPE).

4. Results

The classification algorithms achieved performance results for the seven use cases. The use cases are best described by the data arrangement shown in Table 2. The nine classification algorithms produced (predicted) labels for all the feature vectors in the test fold. These labels were compared with the ground truth by establishing:

- True Positive (TP): The number of correctly identified positive cases (ground truth = predicted label = positive).
- True Negative (TN): The number of correctly identified negative cases (ground truth = predicted label = negative).
- False Positive (FP): The number of incorrectly identified positive cases (ground truth ≠ predicted label = positive).
- False Negative (FN): The number of incorrectly identified negative cases (ground truth ≠ predicted label = negative).

For use cases 1–5, the normal set was used as negative ground truth. For use cases 6 and 7 the NCD set was used as negative ground truth. All sets that were not negative were treated as positive ground truth. Based on these fundamental measures, we utilized the following three performance measures:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \times 100, \quad \mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100, \quad \mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \times 100 \tag{1}$$

Table 3
Performance results of the best classification algorithm for the individual binary classification problems.

Number	Classifier	ACC (%)	SEN (%)	SPE (%)
1	Subspace	74.4	75.4	73.0
2	Subspace	76.1	82.0	70.2
3	SVM (1st order polynomial kernel)	98.5	97.7	99.0
4	Bagging	75.9	79.8	67.5
5	Bagging	76.4	80.1	70.2
6	Subspace	98.2	99.5	95.3
7	Subspace	97.1	99.0	93.6

+ CD without Marsh I+II). For this problem, the Bagging classifier obtained 76.4% ACC, which was the highest performance.

Table 3 lists the classifier that achieved the highest ACC for a given binary classification problem. The table shows that the SVM (1st order polynomial kernel) yielded the highest ACC of 98.5%, for the third binary classification problem (Number 3), which reflects the use case of differentiating normal biopsy images from those showing NCD. We reason that the polynomial kernel enables the conversion of the original data space into a new one with a higher dimension, resulting in enhanced separability [45], thus yielding satisfactory classification results. On the other hand, the classifiers achieved the least performance for the binary classification problem (Number) 4 (Normal versus NCD

Tables 4–10 provide the detailed performance results from each of the nine classification algorithms for all the seven binary classification problems. To be specific, Table 4 details the performance results for the binary classification problem reflecting the use case of differentiating normal biopsy images from those showing CD with Marsh I+II, etc. In each of these seven tables, we have highlighted the classification algorithm which achieved the highest accuracy (row with gray background in each table).

5. Explainable AI using Shapley analysis

The success of a particular classification system depends upon trust. A breach of trust will erode confidence in a particular system [46]. This is particularly relevant to medical decision support systems [47] because such systems require human experts to rely on machine learning.

Hence, despite the good classification results achieved by AI models, they remain underused as physicians are not able to comprehend the basis on which these models make predictions regarding patients' health. To counter this, explainable artificial intelligence techniques are being explored fervently and are hence emanating in the healthcare domain presently [48,49]. Jahmunah et al. [48] used the gradient-weighted class activation mapping visualization technique to show the different locations on the Electrocardiogram (ECG) signals that were influential in the prediction of myocardial infarction. Loh et al. [49] conducted a systematic review and discussed various explainable AI techniques used in healthcare. SHapley Additive exPlanations (SHAP) methods have recently been presented in some studies to aid in interpreting ML models, irrespective of their complexity level. The SHAP method is efficacious in allowing the detection and ordering of features that determine compound classification and activity forecasting in any ML model [50]. In a recent study, Ibrahim et al. [51] performed Shapley analysis on a decision tree-based ML model to gain insights into the model's prediction capabilities and to pinpoint features that influenced the model most in decision-making for the detection of acute myocardial infarction. Shapley values are helpful in exposing the contribution of each feature to each prediction [51]. In our study, we applied the Shapley model to the output (top 10 discriminatory PHOG features) of the best performing SVM (1st order polynomial kernel) classifier using the Python program.

Table 4
Performance results of various classifiers for the binary classification scheme Number 7 (NCD vs. CD - Marsh I+II).

Classifier	TP	TN	FP	FN	ACC (%)	PPV (%)	SEN (%)	SPE (%)	F1-score
DT	929	542	38	31	95.52	96.24	96.77	93.38	0.96
KNN 5	943	552	28	17	97.07	97.19	98.21	95.10	0.98
SVM RBF	937	552	28	23	96.69	97.23	97.58	95.08	0.97
SVM Poly1	940	553	27	20	96.93	97.32	97.91	95.24	0.98
SVM Poly2	935	549	31	25	96.36	96.91	97.35	94.61	0.97
SVM Poly3	933	554	26	27	96.56	97.40	97.17	95.47	0.97
AdaBoost M1	902	539	41	58	93.56	96.27	94.05	92.38	0.97
Bagging	932	546	34	28	95.98	96.65	97.07	94.05	0.97
Subspace	951	544	36	9	97.06	96.49	99.04	93.60	0.98

Table 5
Performance results of various classifiers for the binary classification scheme Number 6 (NCD vs. CD + Marsh I+II).

Classifier	TP	TN	FP	FN	ACC (%)	PPV (%)	SEN (%)	SPE (%)	F1-score
DT	1327	547	33	23	97.10	97.65	98.30	94.35	0.98
KNN 5	1334	554	26	16	97.82	98.13	98.81	95.43	0.98
SVM RBF	1328	549	31	22	97.25	97.76	98.37	94.59	0.98
SVM Poly1	1333	558	22	17	97.98	98.41	98.74	96.17	0.99
SVM Poly2	1326	553	27	24	97.35	98.04	98.22	95.24	0.98
SVM Poly3	1328	556	24	22	97.61	98.27	98.37	95.81	0.98
AdaBoost M1	1196	552	28	154	90.54	97.79	88.59	95.20	0.98
Bagging	1324	552	28	26	97.20	98.00	98.07	95.17	0.98
Subspace	1343	553	27	7	98.23	98.07	99.48	95.29	0.99

Table 6
Performance results of various classifiers for the binary classification scheme Number 5 (Control vs. NCD+CD wo. Marsh I+II).

Classifier	TP	TN	FP	FN	ACC (%)	PPV (%)	SEN (%)	SPE (%)	F1-score
DT	1144	607	303	396	71.46	79.57	74.24	66.88	0.77
KNN 5	1068	673	237	472	71.04	82.13	69.32	74.00	0.75
SVM RBF	1197	588	322	343	72.85	79.21	77.71	64.76	0.78
SVM Poly1	1134	699	211	406	74.79	84.57	73.61	76.89	0.78
SVM Poly2	1150	621	289	390	72.27	80.34	74.64	68.42	0.77
SVM Poly3	1145	592	318	395	70.87	78.60	74.30	65.17	0.76
AdaBoost M1	1174	655	255	366	74.66	82.80	76.19	72.16	0.79
Bagging	1234	637	273	306	76.38	82.15	80.10	70.21	0.81
Subspace	1053	726	184	487	72.62	85.81	68.36	79.93	0.76

Table 7
Performance results of various classifiers for the binary classification scheme Number 4 (Control vs. (NCD+CD + Marsh I+II)).

Classifier	TP	TN	FP	FN	ACC (%)	PPV (%)	SEN (%)	SPE (%)	F1-score
DT	1456	558	352	474	70.90	80.60	75.43	61.29	0.78
KNN 5	1292	657	253	638	68.63	83.80	66.94	72.25	0.74
SVM RBF	1489	546	364	441	71.67	80.48	77.14	60.10	0.79
SVM Poly1	1367	684	226	563	72.21	85.99	70.81	75.15	0.77
SVM Poly2	1405	625	285	525	71.47	83.31	72.76	68.65	0.77
SVM Poly3	1408	587	323	522	70.26	81.58	72.94	64.68	0.77
AdaBoost M1	1454	637	273	476	73.62	84.40	75.33	70.00	0.79
Bagging	1541	614	296	389	75.88	84.04	79.83	67.53	0.82
Subspace	1317	714	196	613	71.50	87.46	68.22	78.58	0.76

Table 8
Performance results of various classifiers for the binary classification scheme Number 3 (Control vs. NCD).

Classifier	TP	TN	FP	FN	ACC (%)	PPV (%)	SEN (%)	SPE (%)	F1-score
DT	569	893	17	11	98.13	97.18	98.07	98.16	0.98
KNN 5	561	894	16	19	97.66	97.38	96.75	98.25	0.97
SVM RBF	562	898	12	18	98.00	98.08	96.92	98.68	0.97
SVM Poly1	567	901	9	13	98.53	98.55	97.73	99.01	0.98
SVM Poly2	564	896	14	16	98.00	97.78	97.21	98.46	0.97
SVM Poly3	565	892	18	15	97.79	97.14	97.40	98.02	0.97
AdaBoost M1	418	374	536	162	53.24	48.21	69.14	39.55	0.63
Bagging	568	893	17	12	98.06	97.18	97.89	98.16	0.97
Subspace	561	902	8	19	98.20	98.64	96.71	99.12	0.98

Table 9
Performance results of various classifiers for the binary classification scheme Number 2 (Control vs. CD wo. Marsh I+II).

Classifier	TP	TN	FP	FN	ACC (%)	PPV (%)	SEN (%)	SPE (%)	F1-score
DT	714	631	279	246	71.92	72.42	74.47	69.43	0.73
KNN 5	769	606	304	191	73.54	72.08	80.17	66.64	0.76
SVM RBF	733	612	298	227	71.93	71.61	76.46	67.41	0.74
SVM Poly1	790	627	283	170	75.77	74.18	82.38	68.96	0.78
SVM Poly2	765	604	306	195	73.23	71.71	79.87	66.37	0.75
SVM Poly3	713	598	312	247	70.10	70.16	74.44	65.75	0.71
AdaBoost M1	752	622	288	208	73.48	72.73	78.37	68.42	0.75
Bagging	745	641	269	215	74.11	73.89	77.67	70.50	0.75
Subspace	785	637	273	175	76.05	74.86	81.95	70.15	0.78

Table 10
Performance results of various classifiers for the binary classification scheme Number 1 (Control vs. CD w. Marsh I+II).

Classifier	TP	TN	FP	FN	ACC (%)	PPV (%)	SEN (%)	SPE (%)	F1-score
DT	947	594	316	403	68.17	75.30	70.15	65.39	0.72
KNN 5	924	634	276	426	68.92	77.41	68.44	69.62	0.72
SVM RBF	1009	596	314	341	71.00	76.46	74.74	65.55	0.75
SVM Poly1	1011	652	258	339	73.55	79.95	74.89	71.69	0.77
SVM Poly2	979	611	299	371	70.34	76.91	72.52	67.13	0.74
SVM Poly3	978	586	324	372	69.21	75.33	72.44	64.46	0.74
AdaBoost M1	1004	619	291	346	71.82	77.76	74.37	68.13	0.76
Bagging	1037	620	290	313	73.30	78.49	76.81	68.25	0.77
Subspace	1018	664	246	332	74.41	80.88	75.41	72.97	0.77

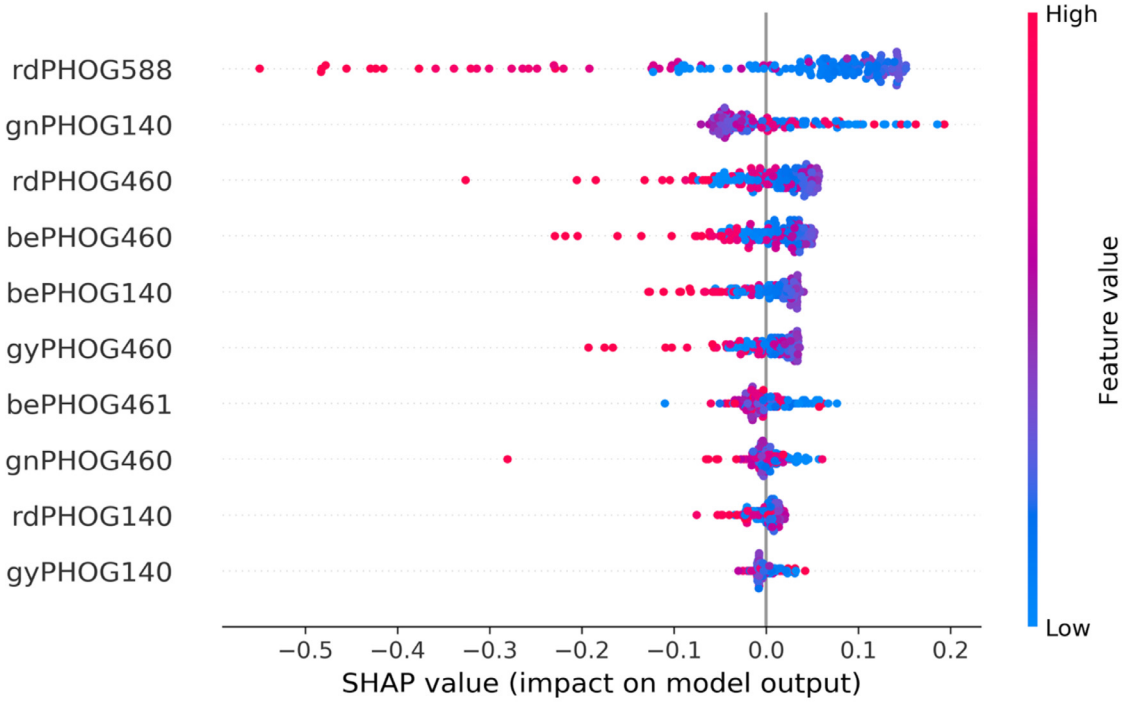

Fig. 5. Results of Shapley analysis performed on SVM (1st order polynomial kernel) classifier.

Fig. 5 shows the results of the Shapely analysis performed on the classifier. The analysis shows how the 'SHAP' values influence the feature values (high or low values). From the figure., it can be concluded that the features at the top signify high feature importance, as these contribute more to the model's prediction than those at the bottom. The color representation can determine the value of each feature. For example, features 'rdPHOG588' to 'gyPHOG460' comprise more red dots as compared to the blue or purple dots, while the rest of the features from 'bePHOG461' to 'gyPHOG140' contain fewer red dots. This observation attests that features 'rdPHOG588' to 'gyPHOG460' have greater relevance in impacting the model's output than features 'rdPHOG588' till 'gyPHOG460'. This could possibly be due to the extraction of features 'rdPHOG588' till 'gyPHOG460' from highly discriminatory celiac images for the different channels.

6. Discussion

The satisfactory ability of our AI model to distinguish CD from normal controls and NCD based on the LP compartment, shown in

the Results section tables, appears promising. Its ability to differentiate classes suggests, in part, differences in the types, density, and distribution of inflammatory or stromal cells in different small intestinal inflammatory diseases. This assumption is supported by transcriptional analyses of small intestinal mucosa in CD and allied diseases, which have revealed changes in LP innate and adaptive immune cells and stromal cells, as well as extracellular matrix remodeling enzymes, in comparison to normal mucosa [52–56]. The ability to discern between normal, CD, and NCD by analyzing the LP changes can therefore be a significant step towards creating a fully automated small intestinal biopsy analysis system that will assist pathologists whenever the villous architecture cannot be reliably assessed. Additionally, the best classifier performance can be interpreted by determining the most important feature(s) that can influence the classifier to yield the highest classification results via AI methodology.

In summary, several stages of analysis were implemented for the completion of our study. The following list details the contribution of our work to the knowledge of automated CD detection by analyzing small intestinal biopsy images:

- We have analyzed various disease combinations, as expressed by seven biopsy image classes using high-magnification digital images (40×).
- To the best of our knowledge, this is the first work that discerns normal from inflamed duodenal biopsies based on AI analysis of the LP compartment.
- This is one of the earliest studies to establish features that best influence the top-performing classifier to yield the highest classification ACC.
- We ran a separate analysis with and without CD images classified as Marsh I+II.
- We developed an accurate ML model using a new method with PHOG features.
- We generated an accurate and robust high-performance model.
- Three-fold cross-validation was repeated ten times to obtain the results.
- We have used ADASYN to extend and balance the data for each fold. This makes the approach even more robust.

6.1. Current quandaries in Celiac Disease diagnosis

CD diagnosis continues to be a public health issue, and the disease remains undiagnosed in the majority of affected persons [57–59]. Our previous study and others have demonstrated a marked variability in intestinal biopsy interpretation in the diagnosis of CD that results in both the under- and over-diagnosis of CD [60–62]. Prior studies have established that computational methods could be useful for automation. Hence, more accurate CD detection using the available resources is vital to improving public health. Statistical assessment of extracted features has shown that it is possible to automate the classification of CD related biopsy images. Innovations introduced by Sali et al. [21], Syed et al. [22], and Wei et al. [23] suggest that AI algorithms can be useful for decision support in CD diagnosis. This has led to quantifying decision support quality through objective measures, such as classification ACC. Providing medical decision support through automated biopsy image analysis, as developed in our current study, is therefore a new and possibly important avenue to address the problem. Using morphometric analysis of small intestinal mucosa, prior studies highlighted differences in the LP volume and cell type of untreated celiac mucosa compared to normal controls [27]. However, automated, computational analysis of differences in the LP cellular contexture of CD and NCD or normal biopsies has not been performed. In the current study, we extend our work by assessing the NCD class of biopsy images and determining whether the LP cellular/inflammatory milieu can help discern different etiologies of intestinal inflammatory disorders and discriminate between inflamed and uninflamed mucosa. Incorporating the NCD class in an automated detection system for CD is important to ensure an accurate diagnosis of CD, as some of the histopathologic features of CD and NCD may overlap. Our current methodology is capable of differentiating normal versus CD and NCD biopsy images.

Future advances in automated detection and classification should carefully consider the findings of prior studies. Future advances in automated detection and classification should carefully consider the findings of prior studies. For example, using morphometric analysis of small intestinal mucosa, prior studies have highlighted differences in the LP volume and cell type of untreated celiac mucosa as compared to normal controls [27]. However, automated, computational analysis of differences in the LP compartments of CD and NCD or normal biopsies have not yet been performed. In the current study, we extended our work by assessing the NCD class of biopsy images and determining whether the differences in the LP constituents can help discern different etiologies of intestinal inflammatory disorders and discriminate between inflamed and uninflamed mucosa. Incorporating the NCD class in an automated detection system for CD is important to ensure an accurate diagnosis of CD, as some of the histopathologic features of CD and NCD overlap. In addition, our current methodology can differentiate normal versus CD and NCD biopsy images. We achieved this functionality by extracting novel PHOG features from the LP compartment of the biopsy images. Hence, evaluation of the LP compartment is valuable for further design and implementation of a fully automated CD and NCD detection system. Such a system could have great potential to accelerate the diagnostic process and improve patient outcomes. Table 11 summarizes selected studies on the automated detection of CD based on biopsy images, on which our current study builds.

6.2. Limitations and strength of the study

This study has some limitations. Obtaining biopsy samples is an invasive [63] and resource intensive procedure. It requires specialized medical facilities equipped with endoscopic instrumentation guided by human experts. Both human expertise and specialized facilities are limited resources where demand outstrips supply, resulting in elevated cost. We accepted these limiting factors since biopsy images are considered the gold standard for CD diagnostics. They contain salient information that is unique as compared with data acquired via other modalities such as endoscopy or video capsule endoscopy.

The biopsy images used for our study were selected manually based on the decision of two pathologists. Although the images used for analysis attempted to capture only the LP inflammatory and stromal cells, the inclusion of crypt, and in some cases villous epithelium was unavoidable. Hence, the possibility of incorporating certain architectural features or epithelial abnormalities by the classifier algorithms cannot be entirely excluded. Furthermore, only a limited number of cases from a few NCD entities were assessed in the current study. Therefore, to improve the performance, we will require more data from a wider range of patients/diseases and more pathologists to classify and select the images and reduce bias.

As part of the pre-processing step, the biopsy images were resized for standardization. This may possibly lead to loss of information and bias during feature extraction, hence affecting the classification accuracy. Thus, this poses another limitation to our study method.

Another data-related limitation comes from the small number of biopsy images that were used to train and test the ML model. Due to the small data pool, we could not test our model using

Table 11
Summary of selected research work on automated CD detection, based on biopsy image analysis.

Author, Year	No images	Method	Classes	Results
Ciaccio et al., 2008 [20]	31	Edge detection & linear relationships	Marsh IIIa, IIIb, IIIc	Statistical Assessment
Sali et al., 2019 [21]	162 biopsyimages from 34 patients	Stain normalization & Deep learning	Marsh I, IIIa, IIIb, and IIIc	ACC between 89.54% and 90.61%
Syed et al., 2019 [22]	102	Deep learning		ACC of 93.4%
Wei et al., 2019 [23]	1230 slides from 1048 patients	Deep learning	Binary	ACC between 83.30% and 92.2%
Kowsari et al., 2020 [24]	461 from 150 patients	Deep learning	normal, CD, Enteropathy	F1-score = 89.66%
Koh et al.,2021 [25]	91 biopsy images 77 CD and 14 normal	Feature engineering and ML	Binary	ACC between 82.92% and 88.89%
This work	91 biopsy images from 31 subjects	PHOG features SVM DT KNN AdaBoost Bagging Subspace	Seven binary problems based on seven image sets	Set 1: 74.41% ACC Set 2: 76.05% ACC Set 3: 98.53% ACC Set 4: 75.88% ACC Set 5: 76.38% ACC Set 6: 98.23% ACC Set 7: 97.06% ACC

an independent dataset. Having only 284 images was the reason for choosing a feature-based CD detection method. Specifically, incorporating PHOG based feature extraction enabled us to exercise tight control of the design process. We used feature ranking and classification-based feature selection during the design phase. However, the design process required some subjective decision-making, such as determining which features to extract and what ML classifier to use. Furthermore, the feature ranking was based on a linear methodology that explored the statistical but not the decision support relevance of the extracted features. In general, feature engineering is an exercise in data reduction that leads to information reduction as well. Unfortunately, that information extraction results from subjective design decisions that can limit the useful information content extracted from the available data [64]. Furthermore, the information contained in the 284 images analyzed might be insufficient to extract adequate knowledge on how to construct a generalized CD detection system.

Despite these limitations, the study has also novelties and strengths. The approach used in our study is beneficial as we achieved the highest classification accuracy of 98.5% with the SVM (1^{st} order polynomial kernel) for the classification of normal versus NCD classes. This attests that our proposed technique is exemplary in discerning normal from NCD classes. Furthermore, this study is novel as it is the earliest to have discerned between normal, NCD, and CD classes by investigating the lamina propria composition.

6.3. Future work direction

We plan to improve data quality by sourcing additional biopsy images representing diverse small intestinal inflammatory diseases from other gastroenterology centers and include additional pathologists to review them. Moreover, the increased data will enable us to address subjectivity issues inherent in feature-based AI modeling by using a deep learning approach. A larger data pool would also enable us to further test our model with an independent image set, as part of our future work. Additionally, since a deep learning algorithm extracts knowledge directly from labeled data without the drawback imposed by additional feature engineering, we expect that with more image data, deep learning results will be more transferable as compared to traditional feature engineering results.

Once standardized, our proposed methodology may be useful to assist in reducing inter- and intra-observer interpretations of biopsy images. In addition, there are several application scenarios for automated detection of CD for which our work can act as a springboard. For example, when expertise is in short supply, such a technique can assist in selecting patients for screening.

7. Conclusion

From a public health perspective, automated CD detection via image analysis technology and AI can enable testing more patients with the same resources. This can lead to a more rapid and early-stage CD detection capability. In addition, detecting subtle changes in the density and distribution of particular types of inflammatory cells in the LP in CD biopsies may improve the detection of CD and assist pathologists in the diagnosis whenever the biopsy orientation does not allow reliable assessment of the villous architecture. Furthermore, through this study, clinicians can gain knowledge on the right type of PHOG featuring to be extracted from normal versus celiac imagery for improved classification accuracy. Finally, the results we obtained in the classification of images of the LP in CD vs. NCD suggest differences in the types and distribution of inflammatory (or stromal) cells in different small intestinal inflammatory diseases, which can be explored by novel spatial transcriptomic and single-cell analytic technologies to identify diagnostic and disease severity-related biomarkers.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2022.107320.

Appendix A. Feature list

Table A.12
Range (Mean ± SD) of some clinical PHOG features (p-value < 0.05 and f-value > 6) for normal and NCD classes.

	Control		NCD			
Feature	Mean	SD	Mean	SD	p-value	t-value
rdPHOG460	1.42E-04	8.28E-05	5.86E-04	1.47E-04	0	20.9596
bePHOG460	1.57E-04	9.14E-05	7.28E-04	2.08E-04	0	19.7576
gyPHOG460	1.48E-04	8.89E-05	5.17E-04	1.50E-04	0	16.9364
rdPHOG140	6.20E-04	2.46E-04	1.59E-03	3.93E-04	0	16.7133
bePHOG140	6.80E-04	2.90E-04	1.94E-03	5.99E-04	0	14.9583
gnPHOG460	1.53E-04	9.31E-05	4.52E-04	1.42E-04	0	14.2104
gyPHOG140	6.34E-04	2.59E-04	1.47E-03	4.27E-04	0	13.4525
rdPHOG588	1.61E-04	7.04E-05	6.05E-04	2.50E-04	0	13.1887
bePHOG461	1.57E-04	9.55E-05	4.08E-04	1.37E-04	0	12.1936
gnPHOG140	6.40E-04	2.65E-04	1.33E-03	3.87E-04	0	11.9549
bePHOG588	1.77E-04	8.29E-05	7.93E-04	3.90E-04	0	11.8805
bePHOG44	2.95E-03	7.86E-04	4.76E-03	9.92E-04	0	11.7403
gyPHOG588	1.64E-04	7.57E-05	5.67E-04	2.64E-04	0	11.3191
rdPHOG44	2.68E-03	6.54E-04	4.09E-03	8.10E-04	0	11.1228
gnPHOG588	1.66E-04	7.80E-05	4.98E-04	2.32E-04	0	10.5158
rdPHOG453	1.73E-04	9.98E-05	3.48E-04	9.85E-05	0	10.4927
bePHOG589	1.75E-04	8.40E-05	4.43E-04	1.94E-04	0	9.9399
bePHOG453	1.72E-04	1.03E-04	3.82E-04	1.38E-04	0	9.9392
bePHOG141	6.67E-04	2.83E-04	1.24E-03	3.84E-04	0	9.7823
gyPHOG461	1.63E-04	1.06E-04	3.47E-04	1.18E-04	0	9.6677
gyPHOG44	2.71E-03	6.94E-04	4.03E-03	8.79E-04	0	9.6302
rdPHOG461	1.55E-04	9.25E-05	3.19E-04	1.07E-04	0	9.536
gyPHOG589	1.80E-04	8.50E-05	3.83E-04	1.49E-04	0	9.4625
rdPHOG589	1.73E-04	7.28E-05	3.59E-04	1.41E-04	0	9.2958
bePHOG12	1.17E-02	2.59E-03	1.53E-02	2.16E-03	0	9.2431
gnPHOG461	1.67E-04	1.06E-04	3.28E-04	1.06E-04	0	9.0512
rdPHOG581	1.99E-04	7.95E-05	4.16E-04	1.74E-04	0	8.9301
gnPHOG589	1.81E-04	8.77E-05	3.58E-04	1.39E-04	0	8.681
gnPHOG44	2.71E-03	7.02E-04	3.85E-03	8.33E-04	0	8.6247
bePHOG581	1.98E-04	9.32E-05	4.80E-04	2.42E-04	0	8.5018
rdPHOG133	7.68E-04	2.81E-04	1.21E-03	3.33E-04	0	8.3989
gyPHOG141	6.83E-04	2.83E-04	1.12E-03	3.37E-04	0	8.2884
rdPHOG141	6.63E-04	2.54E-04	1.08E-03	3.20E-04	0	8.2809
bePHOG133	7.67E-04	3.10E-04	1.30E-03	4.51E-04	0	7.958
gyPHOG453	1.70E-04	1.05E-04	2.96E-04	8.69E-05	0	7.9312
bePHOG45	2.94E-03	7.49E-04	3.91E-03	7.19E-04	0	7.9171
rdPHOG12	1.08E-02	2.21E-03	1.35E-02	1.92E-03	0	7.8099
gyPHOG12	1.09E-02	2.33E-03	1.37E-02	2.01E-03	0	7.783
bePHOG15	2.11E-02	2.90E-03	1.79E-02	2.15E-03	0	7.6825
bePHOG452	1.40E-04	8.48E-05	2.70E-04	1.10E-04	0	7.6475
gnPHOG141	6.89E-04	2.88E-04	1.08E-03	3.18E-04	0	7.5892
bePHOG580	1.59E-04	7.53E-05	3.27E-04	1.58E-04	0	7.5784
gnPHOG12	1.08E-02	2.27E-03	1.34E-02	1.91E-03	0	7.5211
gyPHOG581	1.94E-04	8.97E-05	3.59E-04	1.54E-04	0	7.3859
gnPHOG453	1.75E-04	1.11E-04	2.81E-04	7.51E-05	0	6.9925
gnPHOG581	1.94E-04	9.30E-05	3.38E-04	1.39E-04	0	6.9453
gyPHOG45	2.93E-03	7.23E-04	3.77E-03	7.60E-04	0	6.762
bePHOG13	1.16E-02	2.59E-03	1.40E-02	1.86E-03	0	6.7303
bePHOG132	6.15E-04	2.45E-04	9.56E-04	3.34E-04	0	6.7099
bePHOG37	3.39E-03	7.65E-04	4.30E-03	8.55E-04	0	6.6271
rdPHOG45	2.91E-03	6.54E-04	3.69E-03	7.27E-04	0	6.6124
gnPHOG45	2.91E-03	7.29E-04	3.70E-03	7.38E-04	0	6.3843
bePHOG36	2.73E-03	5.60E-04	3.39E-03	6.66E-04	0	6.2586
gyPHOG133	7.46E-04	2.95E-04	1.07E-03	3.23E-04	0	6.2473
gyPHOG580	1.66E-04	7.91E-05	2.65E-04	1.05E-04	0	6.1913
bePHOG336	1.01E-03	3.80E-04	1.40E-03	3.66E-04	0	6.171
gnPHOG580	1.66E-04	8.24E-05	2.60E-04	9.92E-05	0	6.0232
gnPHOG13	1.16E-02	2.52E-03	1.38E-02	1.96E-03	0	6.0082

References

[1] P.H. Green, C. Cellier, C. disease, N. Engl. J. Med. 357 (2007) 1731–1743.
[2] P. Singh, A. Arora, T.A. Strand, D.A. Leffler, C. Catassi, P.H. Green, C.P. Kelly, V. Ahuja, G.K. Makharia, Global prevalence of Celiac Disease: systematic review and meta-analysis, Clin. Gastroenterol. Hepatol. 16 (2018) 823–836.
[3] M. Abu-Zekry, D. Kryszak, M. Diab, C. Catassi, A. Fasano, Prevalence of Celiac Disease in egyptian children disputes the east–west agriculture-dependent spread of the disease, J. Pediatr. Gastroenterol. Nutr. 47 (2008) 136–140.
[4] C. Catassi, I.-M. Ratsch, L. Gandolfi, R. Pratesi, E. Fabiani, R. El Asmar, M. Frijia, I. Bearzi, L. Vizzoni, Why is coeliac disease endemic in the people of the sahara? Lancet N. Am. Ed. 354 (1999) 647–648.
[5] S. Mahadov, P.H. Green, Celiac Disease: a challenge for all physicians, Gastroenterol. Hepatol. 7 (2011) 554.
[6] J.A. King, J. Jeong, F.E. Underwood, J. Quan, N. Panaccione, J.W. Windsor, S. Coward, J. deBruyn, P.E. Ronksley, A.A. Shaheen, et al., Incidence of Celiac Disease is increasing over time: a systematic review and meta-analysis, Off. J. Am. Coll. Gastroenterol. ACG 115 (2020) 507–525.
[7] T. Kemppainen, H. Kr¨oger, E. Janatuinen, I. Arnala, V.-M. Kosma, P. Pikkarainen, R. Julkunen, J. Jurvelin, E. Alhava, M. Uusitupa, Osteoporosis in adult patients with Celiac Disease, Bone 24 (1999) 249–255.
[8] E. Topal, F. Catal, N.Y. Acar, H. Ermistekin, M.S. Sinanog˘lu, H. Karabiber, M.A. Selimoglu, Vitamin and mineral deficiency in children newly diagnosed with Celiac Disease, Turk. J. Med. Sci. 45 (2015) 833–836.

[9] K. Long, A. Rubio-Tapia, A. Wagie, L. Melton Iii, B. Lahr, C. Van Dyke, J. Murray, The economics of coeliac disease: a population-based study, Aliment. Pharmacol. Ther. 32 (2010) 261–269.

[10] O. Faust, U.R. Acharya, T. Tamura, Formal design methods for reliable computer-aided diagnosis: a review, IEEE Rev. Biomed. Eng. 5 (2012) 15–28.

[11] A. Rostom, J.A. Murray, M.F. Kagnoff, American gastroenterological association (aga) institute technical review on the diagnosis and management of Celiac Disease, Gastroenterology 131 (2006) 1981–2002.

[12] A. Fasano, I. Berti, T. Gerarduzzi, T. Not, R.B. Colletti, S. Drago, Y. Elitsur, P.H. Green, S. Guandalini, I.D. Hill, Prevalence of Celiac Disease in at-risk and not-at-risk groups in the united states: a large multicenter study, Arch. Intern. Med. 163 (2003) 286–292.

[13] P.H. Green, The role of endoscopy in the diagnosis of Celiac Disease, Gastroenterol. Hepatol. 10 (2014) 522.

[14] I. Parzanese, D. Qehajaj, F. Patrinicola, M. Aralica, M. Chiriva-Internati, S. Stifter, L. Elli, F. Grizzi, Celiac Disease: from pathophysiology to treatment, World J. Gastrointest. Pathophysiol. 8 (2017) 27.

[15] G. Wimmer, A. Vécsei, A. Uhl, CNN transfer learning for the automated diagnosis of Celiac Disease, in: Proceedings of the Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, 2016, pp. 1–6.

[16] M. John, Eisenberg center for clinical decisions and communications science. management of postpartum hemorrhage: current state of the evidence. comparative effectiveness review summary guides for clinicians, in: Comparative Effectiveness Review Summary Guides for Clinicians, Agency for Healthcare Research and Quality, Rockville, MD, 2016, pp. 1–20.

[17] M. Marsh, Grains of truth: evolutionary changes in small intestinal mucosa in response to environmental antigen challenge, Gut 31 (1990) 111.

[18] 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.

[19] M. Salvi, U.R. Acharya, F. Molinari, K.M. Meiburger, The impact of pre-and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis, Comput. Biol. Med. 128 (2021) 104129.

[20] E.J. Ciaccio, G. Bhagat, A.J. Naiyer, L. Hernandez, P.H. Green, Quantitative assessment of the degree of villous atrophy in patients with coeliac disease, J. Clin. Pathol. 61 (2008) 1089–1093.

[21] R. Sali, L. Ehsan, K. Kowsari, M. Khan, C.A. Moskaluk, S. Syed, D.E. Brown, Celiacnet: Celiac Disease severity diagnosis on duodenal histopathological images using deep residual networks, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 962–967.

[22] S. Syed, M. Al-Boni, M.N. Khan, K. Sadiq, N.T. Iqbal, C.A. Moskaluk, P. Kelly, B. Amadi, S.A. Ali, S.R. Moore, Assessment of machine learning detection of environmental enteropathy and Celiac Disease in children, JAMA Netw. open 2 (2019) e195822.

[23] J.W. Wei, J.W. Wei, C.R. Jackson, B. Ren, A.A. Suriawinata, S. Hassanpour, Automated detection of Celiac Disease on duodenal biopsy slides: A deep learning approach, J. Pathol. Inform. 10 (2019).

[24] K. Kowsari, R. Sali, L. Ehsan, W. Adorno, A. Ali, S. Moore, B. Amadi, P. Kelly, S. Syed, D. Brown, Hmic: Hierarchical medical image classification, a deep learning approach, Information 11 (2020) 318.

[25] J.E.W. Koh, S. De Michele, V.K. Sudarshan, V. Jahmunah, E.J. Ciaccio, C.P. Ooi, R. Gururajan, R. Gururajan, S.L. Oh, S.K. Lewis, Automated interpretation of biopsy images for the detection of Celiac Disease using a machine learning approach, Comput. Methods Progr. Biomed. 203 (2021) 106010.

[26] M.N. Marsh, P.T. Crowe, 5 morphology of the mucosal lesion in gluten sensitivity, Bailliere's Clin. Gastroenterol. 9 (1995) 273–293.

[27] I. Dhesi, M. Marsh, C. Kelly, P. Crowe, Morphometric analysis of small intestinal mucosa, Virchows Arch. A 403 (1984) 173–180.

[28] C.J. Moran, O.K. Kolman, G.J. Russell, I.S. Brown, M. Mino-Kenudson, Neutrophilic infiltration in gluten-sensitive enteropathy is neither uncommon nor insignificant: assessment of duodenal biopsies from 267 pediatric and adult patients, Am. J. Surg. Pathol. 36 (2012) 1339–1345.

[29] I. Brown, M. Bettington, C. Rosty, The role of histopathology in the diagnosis and management of coeliac disease and other malabsorptive conditions, Histopathology 78 (2021) 88–105.

[30] N. Harpaz, A.D. Polydorides, Upper gastrointestinal manifestations of inflammatory bowel disease, Surg. Pathol. Clin. 13 (2020) 413–430.

[31] 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.

[32] G. Corazza, V. Villanacci, C. disease, J. Clin. Pathol. 58 (2005) 573–574.

[33] E.D. Pisano, S. Zong, B.M. Hemminger, M. DeLuca, R.E. Johnston, K. Muller, M.P. Braeuning, S.M. Pizer, Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms, J. Digit. Imaging 11 (1998) 193.

[34] S.M. Pizer, J.B. Zimmerman, E.V. Staab, Adaptive grey level assignment in ct scan display, J. Comput. Assist. Tomogr. 8 (1984) 300–305.

[35] B.B. Singh, S. Patel, Efficient medical image enhancement using clahe enhancement and wavelet fusion, Int. J. Comput. Appl. 167 (2017) 0975–8887.

[36] S.M. Khaligh-Razavi, What you need to know about the state-of-the-art computational models of object-vision: a tour through the models, arXiv preprint arXiv:1407.2776 (2014).

[37] S.A. Amirshahi, M. Koch, J. Denzler, C. Redies, Phog analysis of self-similarity in aesthetic images, in: Human Vision and Electronic Imaging XVII, volume 8291, International Society for Optics and Photonics, 2012, p. 82911J.

[38] T.K. Kim, T test as a parametric statistic, Korean journal of anesthesiology 68 (2015) 540.

[39] L. Rokach, O.Z. Maimon, Data mining with decision trees: theory and applications, World Sci. 69 (2007).

[40] Z. Zhang, Introduction to machine learning: k-nearest neighbors, Ann. Transl. Med. 4 (2016).

[41] K.S. Durgesh, B. Lekha, Data classification using support vector machine, J. Theor. Appl. Inf. Technol. 12 (2010) 1–7.

[42] L. Rokach, Ensemble methods for classifiers, in: Data Mining and Knowledge Discovery Hand-Book, Springer, 2005, pp. 957–980.

[43] B. Daniel, Cross-validation, Encyclopedia of Bioinformatics and Computational Biology 1 (2018) 542–545.

[44] H. He, Y. Bai, E.A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: Proceedings of the IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)„ IEEE, 2008, pp. 1322–1328.

[45] M. Achirul Nanda, K. Boro Seminar, D. Nandika, A. Maddu, A comparison study of kernel functions in the support vector machine and its application for termite detection, Information 9 (2018) 5.

[46] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: a review of machine learning interpretability methods, Entropy 23 (2020) 18.

[47] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable ai systems for the medical domain?, arXiv preprint arXiv:1712.09923 (2017).

[48] V. Jahmunah, E. Ng, R.-S. Tan, S.L. Oh, U.R. Acharya, Explainable detection of myocardial infarction using deep learning models with grad-cam technique on ecg signals, Comput. Biol. Med. 146 (2022) 105550.

[49] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022), Comput. Methods Progr. Biomed. (2022) 107161.

[50] R. Rodr´ıguez-Pérez, J. Bajorath, Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions, J. Comput. Aided Mol. Des. 34 (2020) 1013–1026.

[51] L. Ibrahim, M. Mesinovic, K.W. Yang, M.A. Eid, Explainable prediction of acute myocardial infarction using machine learning and shapley values, IEEE Access 8 (2020) 210410–210417.

[52] M.M. Leonard, Y. Bai, G. Serena, K.P. Nickerson, S. Camhi, C. Sturgeon, S. Yan, M.R. Fiorentino, A. Katz, B. Nath, et al., Rna sequencing of intestinal mucosa reveals novel pathways functionally linked to Celiac Disease pathogenesis, PLoS One 14 (2019) e0215132.

[53] B. Diosdado, M. Wapenaar, L. Franke, K. Duran, M. Goerres, M. Hadithi, J. Crusius, J. Meijer, D. Duggan, C. Mulder, et al., A microarray screen for novel candidate genes in coeliac disease pathogenesis, Gut 53 (2004) 944–951.

[54] B. Diosdado, H. Van Bakel, E. Strengman, L. Franke, E. Van Oort, C.J. Mulder, C. Wijmenga, M.C. Wapenaar, Neutrophil recruitment and barrier impairment in Celiac Disease: a genomic study, Clin. Gastroenterol. Hepatol. 5 (2007) 574–581.

[55] E. Bigaeva, W.T. Uniken Venema, R.K. Weersma, E.A. Festen, Understanding human gut diseases at single-cell resolution, Hum. Mol. Genet. 29 (2020) R51–R58.

[56] R. Ciccocioppo, A. Di Sabatino, M. Bauer, D.N. Della Riccia, F. Bizzini, F. Biagi, M.G. Cifone, G.R. Corazza, D. Schuppan, Matrix metalloproteinase pattern in celiac duodenal mucosa, Lab. Invest. 85 (2005) 397–407.

[57] J. West, R. Logan, P. Hill, A. Lloyd, S. Lewis, R. Hubbard, R. Reader, G. Holmes, K. Khaw, Seroprevalence, correlates, and characteristics of undetected coeliac disease in england, Gut 52 (2003) 960–965.

[58] K. Rostami, C. Mulder, J. Werre, F. Van Beukelen, J. Kerckhaert, J. Crusius, A. Pena, F. Willekens, J. Meijer, High prevalence of Celiac Disease in apparently healthy blood donors suggests a high prevalence of undiagnosed Celiac Disease in the dutch population, Scand. J. Gastroenterol. 34 (1999) 276–279.

[59] A. Rubio-Tapia, J.A. Murray, Classification and management of refractory coeliac disease, Gut 59 (2010) 547–557.

[60] C. Arguelles-Grande, C.A. Tennyson, S.K. Lewis, P.H. Green, G. Bhagat, Variability in small bowel histopathology reporting between different pathology practice settings: impact on the diagnosis of coeliac disease, J. Clin. Pathol. 65 (2012) 242–247.

[61] M.I.P. Sánchez, E. Smecuol, H. Vázquez, R. Mazure, E. Maurin˜o, J.C. Bai, Very high rate of misdiagnosis of Celiac Disease in clinical practice, Acta Gastroenterol. Latinoam. 39 (2009) 250–253.

[62] R. Shidrawi, R. Przemioslo, D. Davies, M. Tighe, P. Ciclitira, Pitfalls in diagnosing coeliac disease, J. Clin. Pathol. 47 (1994) 693–694.

[63] M. Wiersema, P. Vilmann, M. Giovannini, K. Chang, L. Wiersema, Endosonography-guided fine-needle aspiration biopsy: diagnostic accuracy and complication assessment, Gastroenterology 112 (1997) 1087–1095.

[64] O. Faust, Y. Hagiwara, T.J. Hong, O.S. Lih, U.R. Acharya, Deep learning for healthcare applications based on physiological signals: a review, Comput. Methods Progr. Biomed. 161 (2018) 1–13.