HHS Public Access

Author manuscript

Information (Basel). Author manuscript; available in PMC 2021 August 06.

Published in final edited form as:

Information (Basel). 2020 June ; 11(6): . doi:10.3390/info11060318.

HMIC: Hierarchical Medical Image Classification, A Deep Learning Approach

Kamran Kowsari[1,2,3,*], **Rasoul Sali**[1], **Lubaina Ehsan**[4], **William Adorno**[1], **Asad Ali**[5], **Sean Moore**[4], **Beatrice Amadi**[6], **Paul Kelly**[6,7], **Sana Syed**[4,5,8,*], **Donald Brown**[1,8,*]

[1]Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904, USA;

[2]Office of Health Informatics and Analytics, University of California, Los Angeles (UCLA), CA 90095, USA

[3]Sensing Systems for Health Lab, University of Virginia, Charlottesville, VA 22911, USA

[4]Department of Pediatrics, School of Medicine, University of Virginia, Charlottesville, VA 22903, USA;

[5]Department of Paediatrics and Child Health, The Aga Khan University, Karachi 74800, Pakistan;

[6]Tropical Gastroenterology and Nutrition Group, University of Zambia School of Medicine, 32379 Lusak, Zambia;

[7]Blizard Institute, Barts and The London School of Medicine, Queen Mary University of London, London E1 4NS, UK

[8]School of Data Science, University of Virginia, Charlottesville, VA 22904, USA

Abstract

Image classification is central to the big data revolution in medicine. Improved information processing methods for diagnosis and classification of digital medical images have shown to be successful via deep learning approaches. As this field is explored, there are limitations to the performance of traditional supervised classifiers. This paper outlines an approach that is different from the current medical image classification tasks that view the issue as multi-class classification. We performed a hierarchical classification using our Hierarchical Medical Image classification (HMIC) approach. HMIC uses stacks of deep learning models to give particular comprehension at each level of the clinical picture hierarchy. For testing our performance, we use biopsy of the small bowel images that contain three categories in the parent level (Celiac Disease, Environmental

*Correspondence: kk7nc@virginia.edu (K.K.); sana.syed@virginia.edu (S.S.); deb@virginia.edu (D.B.); Tel.: +1-202-812-3013 (K.K.).

Author Contributions: K.K., S.S., and D.B. worked on the Concept and design of the platform. K.K. worked on the implementation of these models. K.K., S.S., and L.E. worked on the analysis and interpretation of data. K.K. worked on the drafting of the manuscript. K.K., R.S., and W.A. worked on the critical revision of the manuscript for important intellectual content. D.B., S.S., B.A., S.M. and A.A. obtained funding. This work was under the supervision of S.S., P.K., A.A., S.M., and D.B. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Enteropathy, and histologically normal controls). For the child level, Celiac Disease Severity is classified into 4 classes (I, IIIa, IIIb, and IIIC).

Keywords

deep Learning; hierarchical classification; hierarchical medical image classification; medical imaging

1. Introduction and Related Works

Automatic diagnosis of diseases based on medical image categorization has become increasingly challenging over the last several years [1–3]. Areas of research involving deep learning architectures for image analysis have grown in the past few years with an increasing interest in their exploration and understanding of the domain application [3–7]. Deep learning models achieved state-of-the-art results in a wide variety of fundamental tasks such as image classification in the medical domain [8,9]. This growth has raised questions regarding classification of sub-types of disease across a range of disciplines including Cancer (e.g., stage of cancer), Celiac Disease (e.g., Marsh Score Severity Class), and Chronic Kidney Disease (e.g., Stage 1–5) among others [10]. Therefore, it is important to not just label medical images-based specialized areas, but to also organize them within an overall field (i.e., name of disease) with the accompanying sub-field (i.e., sub-type of disease) which we have done in this paper via Hierarchical Medical Image Classification (HMIC). Hierarchical models also combat the problem of unbalanced medical image datasets for training the model and have been successful for other domains [11,12].

In the literature, few efforts have been made to leverage the hierarchical structure of categories. Nevertheless, hierarchical models have shown better performance compared to flat models in image classification across multiple domains [13–15]. These models exploit the hierarchical structure of object categories to decompose the classification tasks into multiple steps. Yan et al. proposed HD-CNN by embedding deep CNNs into a category hierarchy [13]. This model separates easy classes using a coarse category classifier while distinguishing difficult classes using fine category classifiers. In a CNN, shallow layers capture low-level features while deeper layers capture high level ones. Zhu and Bain proposed Branch Convolutional Neural Network (B-CNN) [16] based on this characteristic of CNNs. This model instead of employing different classifiers for different levels of class hierarchy, exploits the hierarchical structure of layers in a CNN and embeds different levels of class hierarchy on a single CNN. B-CNN outputs multiple predictions ordered from coarse to fine along concatenated convolutional layers corresponding to hierarchical structure of the target classes. Sali et al. employed B-CNN model for the classification of gastrointestinal disorders on histopathological images [17].

Our paper uses the HMIC approach for assessment of small bowel enteropathies; Environmental Enteropathy (EE) versus Celiac Disease (CD) versus histologically normal controls. EE is a common cause of stunting in Low-to-Middle Income Countries (LMICs), for which there is no universally accepted, clear diagnostic algorithms or non-invasive biomarkers for accurate diagnosis [18], making this a critical priority [19]. Linear growth

failure (or stunting) is associated with irreversible physical and cognitive deficits, with profound developmental implications [18]. Interestingly, CD, a common cause of stunting in the United States, with an estimated 1% prevalence, is an autoimmune disorder caused by a gluten sensitivity [20] and has many shared histological features with EE (such as increased inflammatory cells and villous blunting) [18]. This resemblance has led to the major challenge of differentiating clinical biopsy images for these similar but distinct diseases. CD severity is further assessed via Modified Marsh Score Classification. It takes into account the architecture of the duodenum as having finger-like projections (called “villi”) which are lined by cells called epithelial cells. Between the villi are crevices called crypts that contain regenerating epithelial cells. Normal villus to crypt ratio is between 3:1 and 5:1 and a healthy duodenum (first part of the small intestine) has no more than 30 lymphocytes interspersed per 100 epithelial cells within the villus surface layer (epithelium). Marsh I comprises of normal villus architecture with an increase in the number of intraepithelial lymphocytes. Marsh II has increased intraepithelial lymphocytes along with crypt hypertrophy (crypts appear enlarged). This is usually rare since patients typically rapidly progress from Marsh I to IIIa. Marsh III is sub-divided into IIIa (partial villus atrophy), Marsh IIIb (subtotal villus atrophy) and Marsh IIIc (total villus atrophy) along with crypt hypertrophy and increased intra-epithelial lymphocytes. Finally, in Marsh IV, villi are completely atrophied [21].

The HMIC approach is shown in Figure 1. The parent level is a model trained based on the parent level of data; EE, CD or Normal. The child level model is trained for sub-classes of CD based on Modified Marsh Score based on severity; I, IIIa, IIIb, and IIIc).

The rest of this paper is organized as follows: In Section 2, the different data sets used in this work, as well as, the required pre-processing steps are described. The architecture of the model is explained in Section 5. Empirical results are elaborated in Section 6. Finally, Section 7 concludes the paper along with outlining future directions.

2. Data Source

As shown in Table 1, the biopsies were already obtained from 150 children in this study with a median (interquartile range) age of 37.5 (19.0 to 121.5) months and a roughly equal sex distribution; 77 males (51.3%), and LAZ/ HAZ (Length/ Height-for-Age Z score) of the EE participants were −2.8 (inter-quartile range (*IQR*) : −3.6 to −2.3) and −3.1 (IQR: −4.1 to −2.2). LAZ/ HAZ of the Celiac participants were −0.3 (IQR: −0.8 to 0.7). and LAZ/ HAZ for Normal were −0.2 (IQR: −1.3 to 0.5). Duodenal biopsy samples were developed into 461 whole-slide biopsy images and labeled as either Normal, EE, or CD. The biopsy slides for EE patients were collected from the Aga Khan University Hospital (AKUH) in Karachi, Pakistan ($n = 29$ slides from 10 patients), and the University of Zambia Medical Center in Lusaka, Zambia ($n = 16$). The slides for Normal patients ($n = 63$) and CD ($n = 34$) were collected from The University of Virginia (UVa). Normal and CD slides were transformed into a whole-slide at 40× amplification using the Leica SCN 400 slide scanner (Meyer Instruments, Houston, TX, USA) at UVa, and the digitized EE slides of 20× and shared by means of the Environmental Enteric Dysfunction Biopsy Investigators (EEDBI) Consortium shared WUPAX server. The patient populace is as per the following:

The median age of ($Q1$, $Q3$) of our whole investigation populace was 37.5 (19.0, 121.5) months, and we had a generally equivalent dispersion of females (48%, $n = 49$) and males (52%, $n = 53$). Most of our examination populace were histologically Normal controls (37.7%), followed by CD patients (51.8%), and EE patients (10.05%).

239 Hematoxylin and eosin (H&E) stained duodenal biopsy samples were collected from the archived biopsies of 63 CD patients from the University of Virginia (UVa) in Charlottesville, VA, USA. The sample were converted into whole-slide images at 40× magnification using the Leica SCN 400 slide scanner (Meyer Instruments, Houston, TX, USA) at the Biorepository and Tissue Research Facility at UVa. The median age of the UVa patient populace is 130 months with interquartile ranges of 85.0 and 176.0 months for $Q1$ and $Q3$, respectively. UVa images had a generally equivalent circulation of females (54%, $n = 54$) and male (46%, $n = 29$). The biopsy labels for this research were determined by two clinical experts and approved by a pathologist with considerable authority in gastroenterology. Our dataset is ranged from Marsh I to IIIc with no biopsy declared as Marsh II.

Based on Table 2, the biopsy images are patched in to 91,899 total images which contain 32,393 normal patches, 29,308 EE patches, and 30,198 CD patches. In the child level of the medical biopsy patches, CD contains 4 severities of disease (Type I, IIIa, IIIb, and IIIc) which has 7125 Type I patches, 6842 Type IIIa patches, 8120 Type IIIb patches, and 8111 Type IIIb patches. The training set for normal and EE contains 22,676 and 20,516 patches, respectively, and for testing 9717 and 8792 patches, respectively. For CD, we have two sets of training and testing where one belongs to the parent model and the other belongs to child level. The parent set contains 21,140 patches for training and 9058 image patches for testing with the common label of CD for all. In the CD child dataset, we have four severity types of this disease (I, IIIa, IIIb, and IIIc). Type I of CD contains 4988 patches in the training set and 2137 patches in the test set. Type IIIa of CD contains 4790 patches in the training set and 2052 patches in the test set. Type IIIb of CD contains 5684 patches in the training set and 2436 patches in the test set. Finally, IIIc of CD contains 5678 patches in the training set and 2137 patches in the test set.

3. Pre-Processing

In this section, we explain the entirety of the pre-processing steps which includes medical image patching, image clustering to remove useless information, and color balancing to solve the staining problem. The biopsy images are unstructured, can vary in size, and are often very high resolution to even consider processing with deep neural systems. Therefore, it becomes necessary to tile the whole-slide images into smaller image subsets called patches. Many of the patches created after tiling the whole-slide image will not contain useful biopsy tissue data. For example, some patches only contain the white or light-gray background area. In the image clustering section, the process to select useful images is described. Lastly, color balancing is used to address staining problems which is a typical issue in histological image preparation.

3.1. Image Patching

Although the effectiveness of CNNs in image classification has been shown in various studies in different domains, training on high-resolution Whole Slide Tissue Images (WSI) is not commonly preferred due to a high computational cost. Applying CNNs on WSI can also lead to losing a large amount of discriminative data because of severe down-sampling [22]. Due to cellular level contrasts between Celiac Disease, Environmental Enteropathy, and Normal cases, an image classification model performed on patches can perform at least similarly to a WSI-level classifier [22]. For this study, patches are labeled with the same class as the associated WSI. The CNN models are trained to predict the presence of disease or disease severity at the patch-level.

3.2. Clustering

As shown in Figure 2, after each biopsy the whole image is divided into patches; many of these patches are not useful input for a deep image classification model. These patches tend to contain only connective tissue, are located on the border region of the tissue, or consist entirely of image background [2]. A two-stage clustering process was applied to recognize the immaterial patches. For the initial step, a convolutional autoencoder was used to learn a vectorized representation of features of each patch and in the second step, we used k-means clustering to assign patches into two groups: helpful and not useful patches. In Figure 3, the pipeline of our clustering strategy is depicted which contains both the autoencoder and *k*-means clustering.

3.2.1. Autoencoder—An autoencoder is a form of a neural network that is intended to output a reconstruction of the model's input [23]. The autoencoder has achieved incredible success as a dimensionality reduction technique [24]. The primary version of the autoencoder was presented by DE. Rumelhart et al. [25] in 1985. The fundamental concept is that one hidden layer acts as a bottle-neck and has far fewer nodes than other layers in the model [26]. This condensed hidden layer can be used to represent the important features of the image with a smaller amount of data. With image inputs, autoencoders can convert the unstructured data into feature vectors that can be processed through other machine learning methods such the k-means clustering algorithm.

Encode: A CNN-based autoencoder can be isolated into two principle steps [27]: encoding and interpreting. This condition is:

$$O_m(i,j) = a\left(\sum_{d=1}^{D}\sum_{u=-2k-1}^{2k+1}\sum_{v=-2k-1}^{2k+1} F_{m_d}^{(1)}(u,v) I_d(i-u, j-v)\right) \quad m = 1, \cdots, n \tag{1}$$

where $F \in \left\{F_1^{(1)}, F_2^{(1)}, \ldots, F_n^{(1)},\right\}$ is a convolutional filter, with convolution among an input volume defined by $I = \{I_1, \ldots, I_D\}$ which it learns to represent the input by combining non-linear functions:

$$z_m = O_m = a\left(I * F_m^{(1)} + b_m^{(1)}\right) \quad m = 1, \ldots, m \tag{2}$$

where $b_m^{(1)}$ is the bias, and the number of zeros we want to pad the input with is such that: dim(I) = dim(decode(encode(I))). Finally, the encoding convolution is equal to:

$$\begin{aligned} O_w = O_h &= (I_w + 2(2k+1) - 2) - (2k+1) + 1 \\ &= I_w + (2k+1) - 1 \end{aligned} \tag{3}$$

Decode: The decoding convolution step produces n feature maps $z_{m=1,\ldots,n}$. The reconstructed results $\hat{I}$ is the result of the convolution between the volume of feature maps $Z = \{z_{i=1}\}^n$ and this convolutional filters volume $F^{(2)}$ [28,29].

$$\tilde{I} = a\left(Z * F_m^{(2)} + b^{(2)}\right) \tag{4}$$

$$O_w = O_h = (I_w + (2k+1) - 1) - (2k+1) + 1 = I_w = I_h \tag{5}$$

where Equation (5) shows the decoding convolution with I dimensions. The input's dimensions are equal to the output's dimensions.

3.2.2. K-Means—

K-means clustering is one of the most popular clustering algorithms [30–34] for data in the form $D \in \{x_1, x_2, \ldots, x_n\}$ in d dimensional vectors for $x \in f^d$. K-means had been applied to perform image and data clustering for information retrieval [30,35,36]. The aim is to identify groups of similar data points and assign each point to one of the groups. There are many other clustering algorithms, but the k-means approach works well for this problem, because there are only two clusters and it is computationally inexpensive compared to other methods.

As an unsupervised approach, one measure of effective clustering is to sum the distances of each data point from the centroids of the assigned clusters. The goal of K-means is to minimize ξ, the sum of these distances, by determining optimal centroid locations and cluster assignments. This algorithm can be difficult to optimize due to the volatility of cluster assignments as the centroid locations change. Therefore, the K-means algorithm is a greedy-like approach that iteratively adjusts these locations to solve the minimization.

Minimize ξ with respect to A and μ by:

$$\xi = \sum_{j=1}^{k} \sum_{x_i} \|x_i - \mu_j\|^2 = \sum_{j=1}^{k} \sum_{i=1}^{n} A_{ij} \|x_i - \mu_j\| \tag{6}$$

where x_i are values from the autoencoder feature representation, μ_j is the centroid of each cluster, and A_{ij} is the cluster assignment of each data point i with cluster j. A_{ij} can only take on binary values and each data point can only be assigned to a single cluster.

The centroid μ of each cluster is calculated as follows:

$$\mu(w) = \frac{1}{|w|} \sum_{\bar{x} \in w} \bar{x} \tag{7}$$

Finally, as shown in Figure 4, all patches are assigned into two clusters which one of them contains useful information and the other one is empty or does not have medical information. The Algorithm 1 indicates kmeans algorithm for two clusters medical images.

Algorithm 1 K-means algorithm for 2 clusters medical images

Input: $D = \{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_n}\}$
Output: $\mu = \{\vec{\mu_1}, \vec{\mu_2}\}$
$S = \{\vec{s_1}, \vec{s_2}\}$ set random seeds
$(\{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_n}\}, K)$
for $i \leftarrow 1$ *to* K **do**
  $\vec{\mu_i} \leftarrow \vec{s_i}$
endfor
while *Criterion has not been met* **do**
  for $i \leftarrow 1$ *to* K=2 **do**
    $w_k \leftarrow \{\}$
  endfor
  for $n \leftarrow 1$ *to* N **do**
    $j \leftarrow arg\,\min_{j'} |\vec{\mu_{j'}} - \vec{x_n}|$
    $w_j \leftarrow w_j \bigcup \{\vec{x_n}\}$
  endfor
  for $i \leftarrow 1$ *to* K=2 **do**
    $\mu_i \leftarrow \frac{1}{|w_i|} \sum_{\vec{x} \in w_i} \vec{x}$
  endfor
endwhile

3.3. Medical Image Staining

Hematoxylin and eosin (H&E) stains have been used for at least a century and are still essential for recognizing various tissue types and the morphologic changes that form the basis of contemporary CD, EE, and cancer diagnosis [37]. H&E is used routinely in histopathology laboratories as it provides the pathologist/researcher a very detailed view of the tissue [38]. Color variation has been a very important problem in histopathology based on light microscopy. A range of factors makes this problem even more complex such as the use of different scanners, variable chemical coloring/reactivity from different manufacturers/batches of stains, coloring being dependent on staining procedure (timing, concentrations, etc.), and light transmission being a function of section thickness [39]. Different H&E staining appearances within machine learning inputs can cause the model to focus only on the broad color variations during training. For example, if images with a certain label all have a unique stain color appearance, because they all originated from the same location, the machine learning model will likely leverage the stain appearance to classify the images rather than the important medical cellular features.

3.3.1. Color Balancing—The idea of color balancing for this study is to convert images in to a similar color space to represent variations in H&E staining. The images can be represented with the illuminant spectral power distribution as shown by $I(\lambda)$, the surface spectral reflectance $S(\lambda)$, and the $C(\lambda)$ is sensor spectral sensitivities [40,41]. Using these notations [41], the sensor reactions at the pixel with coordinates of (x, y) which can be presented as:

$$p(x, y) = \int_w I(x, y, \lambda) S(x, y, \lambda) C(\lambda) d\lambda \tag{8}$$

where w is the wavelength range of the visible light spectrum, p and $C(\lambda)$ are three-component vectors.

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix}_{out} = \left(\alpha \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \times \begin{bmatrix} r_i & 0 & 0 \\ 0 & g_i & 0 \\ 0 & 0 & b_i \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}_{in} \right)^{\gamma} \quad (9)$$

where RGB_{in} stand for the raw images from medical images, and the diagonal matrix diag(r_i, g_i, b_i) is the channel-independent gain compensation of the illuminant [41]. In addition, RGB_{out} is output results that be send to input feature space of CNN models. γ is the gamma correction defined for the RGB color space and RGB_{out} are the output RGB values. In the following, a more compact version of Equation (9) is used:

$$RGB_{out} = (\alpha A I_w \,.\, RGB_{in})^{\gamma} \quad (10)$$

where a stand for exposure compensation gain, and the diagonal matrix for the illuminant compensation shows by I_w and the color matrix transformation is shown by matrix A which is a diagonal matrix for the illuminant compensation and the color matrix transformation [41].

Figure 5 indicates the output results of three classes (CD, EE, and Normal) for color balancing (CB) with various color balancing percentage in range between 0.01 and 50.

3.3.2. Stain Normalization—Histological images can have significant variations in stain appearance that will cause biases during model training [1]. The variations occur due to many factors such as contrasts in crude materials and assembling procedures of stain vendors, staining conventions of labs, and color reactions to digital scanners [1,42]. To solve this problem, the stains of all images are normalized to a single stain appearance. Different staining normalization approaches have been proposed in research projects. In this paper, we used the methodology proposed by Vahadane et al. [42] for the CD severity child-level since all images are collected from one center. This methodology is designed to preserve the structure of cellular features of images after stain normalization and accomplishes stain separation with non-negative matrix factorization. Figure 6 shows an example outputs before and after applying this method on biopsy patches.

4. Baseline

4.1. Deep Convolutional Neural Networks

A Convolutional Neural Network (CNN) performs hierarchical medical image classification for each individual image. The original version of the CNN was built for image processing with an architecture similar to the visual cortex. In this basic CNN baseline for image processing, an image tensor is convolved with a set of $d \times d$ kernels size. These convolution layers are called feature maps and these provide multiple filters which could be stacked on the input. We used a flat CNN (non-hierarchical CNN) as one of our baselines.

4.2. Deep Neural Networks

A Deep Neural Network (DNN) or multilayer perceptron is designed to be trained by multiple layers of connections. Each individual hidden layer can receive connection from the previous hidden layers' nodes and only can provide connections to the next layer. The input is a connection of flattened feature space (RGB). The output layer is number of classes for multi-class classification (six nodes). Our baseline implementation of DNN (multilayer perceptron) is a discriminative trained model that uses a standard back-propagation algorithm with sigmoid (Equation (12)) and Rectified Linear Units (ReLU) [43] (Equation (13)) activation functions. The output layer for classification task uses the *Softmax* function due to having multi-class output as shown in Equation (14).

5. Method

In this section, we explain our concept of Deep Convolutional Neural Networks (CNN) containing the convolutional layers, activation functions, pooling-layers, and finally, the optimizer. Then, we describe our Deep Convolutional Neural Networks architecture to diagnose Celiac disease and environmental enteropathy. As shown in Figure 7, the input layer consists of image patches with size of (1000 × 1000 pixels) and it follows the connection to the convolutional layer (*Conv* 1). Conv 1 connects to the its following pooling layer (*MaxPooling*). The pooling layer is connected to second convolutional layer *Conv* 2. The last convolutional layer (*Conv* 3) has been flattened and connected to a fully connected multi-layer perceptron. The final layer includes three nodes where each individual node represents one class.

5.1. Convolutional Neural Networks

5.1.1. Convolutional Layer—Convolutional Neural Networks are deep learning models that can be used for the hierarchical classification tasks, especially, image classification [44]. Initially, CNNs were designed for image and computer vision with a similar design as the visual cortex. CNNs have been used successfully for clinical image classification. In CNNs, an image tensor is convolved with set of $d \times d$ kernels. These convolutions ("Feature Maps") can be stacked to represent many different features detected by the filters in that layer. The feature dimensions of output and input networks can be different [45]. The procedure for processing a solitary output of a matrix is characterized as follows:

$$A_j = f\left(\sum_{i=1}^{N} I_i * K_{i,j} + B_j\right) \tag{11}$$

Each individual matrix I_i is convolved with its corresponding kernel matrix $K_{i,j}$, and bias of B_j. Finally, a activation function (non-linear activation function is explained in Section 5.1.3) is applied to each individual element [45].

The biases and weights are adjusted to constitute competent feature detection filters after the back-propagation step during CNN training. The feature map filters are applied across all three channels [46].

5.1.2. Pooling Layer—To diminish the computational multifaceted nature, CNNs use pooling layers which decrease the size of the output layer from its input with one layer then onto the next in the networks. Distinctive pooling procedures are used to decrease output while safeguarding significant features [47]. The most widely recognized pooling technique is a max-pooling technique where the largest activation is chosen in the pooling window.

5.1.3. Neuron Activation—The CNN is implemented as a discriminative method that uses a back-propagation algorithm derived from sigmoid (Equation (12)), or (Rectified Linear Units (ReLU) [43] (Equation (13)) activation functions. The final layer contains one node with sigmoid activation function for binary classification multiple nodes for each class and a *Softmax* activation function for multi-class problems (as demonstrated in Equation (14)).

$$f(x) = \frac{1}{1 + e^{-x}} \in (0, 1) \tag{12}$$

$$f(x) = \max(0, x) \tag{13}$$

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \forall j \in \left\{1, \ldots, K\right\} \tag{14}$$

5.1.4. Optimizer—For our CNN architecture, we use the *Adam* optimizer [48]. This is a stochastic gradient descent that uses the norm of the initial two moments of gradient (v and m, appeared in Equations (15)–(18)). It can deal with non-stationarity of the target in a similar fashion to RMSProp, while defeating the sparse gradient problem constraint of RMSProp [48].

$$\theta \leftarrow \theta - \frac{\alpha}{\sqrt{\hat{v}} + \epsilon}\hat{m} \tag{15}$$

$$g_{i,t} = \nabla_\theta J(\theta_i, x_i, y_i) \tag{16}$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_{i,t} \tag{17}$$

$$m_t = \beta_2 v_{t-1} + (1 - \beta_2) g_{i,t}^2 \tag{18}$$

where m_t is the first moment and v_t indicates second moment that both are estimated. $\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$ and $\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$.

5.1.5. Network Architecture—As demonstrated in Figure 7, our implementation contains three convolutional layers with each followed by a pooling layer (Max-Pooling). This method with three channel input image patches with size a of (1000 × 1000 pixels). The first convolutional layer has 32 filters with kernel size of (3, 3). Then, a pooling layer is connected with size of (5, 5) to reduce feature maps from (1000 × 1000) to (200 × 200). The next convolutional layer includes 32 filters with (3, 3) kernel. Then, a $2D$ MaxPooling layer is connected to scales down the feature space from (200 × 200) to (40 × 40). The final convolutional layers contain 64 filters that kernel size is (3, 3). This convolutional layer is connected to a $2D$ MaxPooling to scale down by (8 × 8). The feature map is flattened, and a fully connected layers is connected to our CNN with 128 nodes. The output layer has 3 nodes that represent our parent classes: (Environmental Enteropathy, Celiac Disease, and Normal). The child level of this model as shown on the bottom of Figure 7, is similar to parent level with significant difference which is that the output layer has 4 nodes that represent our child classes: (I, IIIa, IIIb, and IIIc).

The Adam (See Section 5.1.4) optimizer is used with a learning rate of 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The loss function is sparse categorical crossentropy [49]. Also, for all layers, we use a Rectified linear unit (ReLU) as the activation function except for the output layer which used a *Softmax* (See Section 5.1.3). In this technique, we use dropout in each individual layer to address over-fitting problem [50]

5.2. Whole Slide Classification

The objective of this study was to group WSIs dependent on the diagnosis of CD and EE, and CD severity on child-level by means of the adjusted Marsh score. The model was used by training it on the patch-level and is extended to WSI. To accomplish this objective, a heuristic strategy was created which aggregated crop classifications and translated them to whole-slide inferences. Each WSI in the test set was at firstly patched, those patches which did not contain any useful information were filtered out, and then stain methods were performed on the patches (color balancing applied on parent level and stain normalization applied for CD severity). After these pre-processing steps, our prepared model was applied with the objective of image classification. We meant the likelihood dissemination over potential marks, given the patches images x and training set D by $p(y|x, D)$. Finally, this classification produces a vector of length C, where C is the number of classes. In our documentation, the likelihood is contingent on the test patch x, just as, the training set D. The trained model predicts a vector of probabilities (three for parent-level and four for child-level) that represents the likelihood an image belongs in each class. Given a probabilistic result, the patch j in slide i is assigned to the most likely class label $\hat{y}_{ij}$ as shown in Equation (19).

$$\hat{y}_{ij} = \underset{c \in \{1, 2, 3, \ldots, C\}}{\arg\max} p\left(y_{ij} = c \mid x_{ij}, D\right) \tag{19}$$

where $\hat{y}$ stands for maximum a posteriori (MAP). The summation over these vectors (output vector of all patches for a single WSI) and normalizing the resultant vector made a vector that had parts demonstrating the likelihood of a vector with three elements (CD, EE, and N) seriousness for the related WSI. Equation (20), shows how the class of WSI was anticipated.

$$\hat{y}_i = \underset{c \in \{1,2,3,\ldots,C\}}{\arg\max} \sum_{j=1}^{N_i} p\left(y_{ij} = c \mid x_{ij}, D\right) \tag{20}$$

where the number of patches in slide i is shown by N_i.

5.3. Hierarchical Medical Image Classification

The main contribution of this paper is a hierarchical medical image classification of biopsies. A common multi-class algorithm is functional and efficient for a limited number of categories. However, performance drops when we have an unequal number of data-points in our classes. In our deep learning models with various levels, this issue has been solved by creating a hierarchical structure that makes deep learning approaches for their levels of the clinical hierarchy (e.g., see Figure 7).

6. Results

In this section, we have two main results: empirical results and visualizations for patches. The empirical results are mostly used for comparing our accuracy with our baseline.

6.1. Evaluation Setup

In the computer science community, shareable and commensurate performance measures to assess an algorithm are desirable. However, in real projects, such measures may only exist for a few methods. The extensive problem when assessing the medical image categorization model is the absence of standard data collection agreement. Even if a commonplace method existed, simply choosing disparate training and test sets can introduce divergencies in model achievement [51]. Performance measures widely evaluate specific aspects of image classification. In this section, we explain different performance measures and metrics that are used in this research paper. These metrics have been calculated from a “confusion matrix” that comprises false negatives (FN) true negatives (TN), true positives (TP), and false positives (FP) [52]. The importance of these four measures may shift depending on the application. The fraction of all correctly predicted over all number of test set samples is the overall accuracy (Equation (21)). The fraction of correctly predicted over all positives is called precision, i.e., positive predictive value (Equation (22)).

$$accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{21}$$

$$Precision = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L} TP_l + FP_l} \tag{22}$$

$$Recall = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L} TP_l + FN_l} \tag{23}$$

$$F1Score = \frac{\sum_{l=1}^{L} 2TP_l}{\sum_{l=1}^{L} 2TP_l + FP_l + FN_l} \quad (24)$$

6.2. Experimental Setup

The following results were obtained using a combination of central processing units (CPUs) and graphical processing units (GPUs). The processing was done on a *Core i7* – 9700*F* with 8 cores and 128*GB* memory, and the GPU cards were two *Nvidia GeForce RTX* 2080 *Ti*. We implemented our approaches in Python using the Compute Unified Device Architecture (CUDA), which is a parallel computing platform and Application Programming Interface (API) model created by *Nvidia*. We also used Keras and TensorFlow libraries for creating the neural networks [49,53].

6.3. Empirical Results

In this sub-section, as we discussed in Section 6.1, we report precision, recall, and F1-score.

Table 3 shows the results of the parent level model trained for classifying between Normal, Environmental Enteropathy (EE) and Celiac Disease (CD). The precision of normal patches is 89.97 ± 0.5973 and recall is 89.35 ± 0.6133. The F1-score of normal is 89.66 ± 0.6054. For EE, precision is 94.02 ± 0.4955, recall is 97.30 ± 0.3385, F1-score is 95.63 ± 0.4270. The CD evaluation measure for the parent level is as follows: precision is equal to 91.12 ± 0.3208, recall is equal to 88.71 ± 0.3569, and F1-score is equal to 89.90 ± 1.2778.

Table 4 shows the comparison of our techniques with three different baselines. The baseline results from Convolutional Neural Network (CNN), Deep Neural Network (Multilayer perceptron), and Deep Convolutional Neural Network (DCNN) are using in this results section. Much research has been done in this domain such as ResNet, but these novel techniques can only handle small images such as 250 × 250. In this dataset, we create 1000 patches, so we could not compare our work with ResNet, AlexNet, etc. Regarding precision, the highest is HMIC whole-slide with a mean of 88.01 percent and a confidence interval of 0.3841 followed by HMIC none whole-slide 84.13 percent and confidence interval of 0.3751. The precision of CNN is 76.76 ± 0.4985, multilayer perceptron is 76.19 ± 0.5030, and DCNN is 82.95 ± 0.4439. Regarding recall, the highest is HMIC whole-slide with a mean of 93.98 percent and a confidence interval of 0.2811 followed by HMIC non whole-slide at 93.56 percent and confidence interval of 0.291. The recall of CNN is 80.18 ± 0.4706, multilayer perceptron is 79.4 ± 0.471, and DCNN is 87.28 ± 0.3933. The highest F1-score is HMIC whole-slide with a mean of 90.89 percent and a confidence interval of 0.3804 followed by HMIC non whole-slide with 88.61 percent and confidence interval of 0.3751. The recall of CNN is 78.43 ± 0.4855, multilayer perceptron is 77.76 ± 0.4911, and DCNN is 85.06 ± 0.4207.

Table 5 shows the results by each class. For Normal images, the best classifier is DCNN with 95.14 ± 0.42 recall of 94.91 ± 0.43 F1-score of 95.14 ± 0.42. For EE, HMIC is the best classifier. The whole-slide images classifier for parent level is more robust in comparison

with non -whole slide with precision of 94.08 ± 0.49 Recall of 97.33 ± 0.42 F1-score of 98.68 ± 0.42. Although the results of Normal and EE Images are very similar to flat models such as DCNN, but the results of sub-class of CD contains 4 different stages and the margin is very high. The best flat model (non-hierarchical) is DCNN with mean of F1-score of 73.99 for I, 71.63 for IIIa, 77.74 for IIIb, and 75.71 IIIc.

The Table 5 indicates the margin for child level is very high even for the non whole-slide level of this dataset. The best results belong to the whole-slide classifier for parent level with precision with 88.73 ± 1.34 for I, 81.19 ± 1.65 for IIIa, 90.51 ± 1.24 for IIIb, 89.26 ± 1.31 for IIIc. The whole-slide classifier for parent level with recall with 85.07 ± 1.51 for I, 81.19 ± 1.65 for IIIa, 90.48 ± 1.27 for IIIb, 90.18 ± 1.26 for IIIc. The results of whole-slide classifier for parent level for recall is 85.07 ± 1.51 for I, 83.72 ± 0.78 for IIIa, 90.48 ± 0.61 for IIIb, 90.18 ± 1.26 for IIIc. Finally, The F1-score for whole-slide classifier for parent level is equal to 86.86 ± 1.43 for I, 82.44 ± 1.51 for IIIa, 90.49 ± 1.16 for IIIb, 89.72 ± 1.28.

6.4. Visualization

Grad-CAMs were generated for 41 patches (18 EE, 14 Celiac Disease, and 9 histologically normal duodenal controls) which mainly focused on distinct, yet medically relevant cellular features outlined below. Although, most heatmaps focused on medically relevant features, there were some patches that focused on too many features (n = 8) or focused on connective tissue debris (n = 10) that we were unable to categorize.

As shown in Figure 8, three categories are describe as follows:

- EE: surface epithelium with IELs and goblet cells was highlighted. Within the lamina propria, the heatmaps also focused on mononuclear cells.
- CD: heatmaps highlighted the edge of crypt cross sections, surface epithelium with IELs and goblet cells, and areas with mononuclear cells within the lamina propria.
- Histologically Normal: surface epithelium with epithelial cells containing abundant cytoplasm was highlighted.

7. Conclusions

Medical image classification is a significant problem to address, given the growing number of medical instruments to collect digital images. When medical images are organized hierarchically, multi-class approaches are difficult to apply using traditional supervised learning methods. This paper introduces a novel approach to hierarchical medical image classification, HMIC, that could use multiple deep convolutional neural networks approaches to produce hierarchical classifications, and in our experimental results, we use two level of CNNs hierarchy. Testing on a medical image data set shows that this technique produced robust results at the higher and lower level, and the accuracy is consistently higher than those obtainable by conventional approaches using CNN, Multi-layer perceptron, and DCNN. These results show that hierarchical deep learning method could provide improvements for classification and that they provide flexibility to classify these data within

a hierarchy. Hence, they provide extensions over current and traditional methods that only consider the multi-class problem.

This modeling approach can be extended in a couple of ways. Additional training and testing with other hierarchically structured clinical data will help to identify other architectures that work better for these problems. Also, deeper levels of hierarchy is another possible extension of this approach. For instance, if the stage of the disease is treated as ordered then the hierarchy continues down multiple levels. Scoring here could be performed on small sets using human judges.

Funding:

This research was supported by University of Virginia, Engineering in Medicine SEED Grant (*SS* & *DEB*), the University of Virginia Translational Health Research Institute of Virginia (*THRIV*) Mentored Career Development Award (*SS*), and the Bill and Melinda Gates Foundation (AA, OPP1138727; SRM, OPP1144149; PK, OPP1066118). Research reported in this publication was supported by [National Institute of Diabetes and Digestive and Kidney Diseases] of the National Institutes of Health under award number K23 DK117061-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Sali R; Ehsan L; Kowsari K; Khan M; Moskaluk CA; Syed S; Brown DE CeliacNet: Celiac Disease Severity Diagnosis on Duodenal Histopathological Images Using Deep Residual Networks. arXiv 2019, arXiv:1910.03084.
2. Kowsari K; Sali R; Khan MN; Adorno W; Ali SA; Moore SR; Amadi BC; Kelly P; Syed S; Brown DE Diagnosis of celiac disease and environmental enteropathy on biopsy images using color balancing on convolutional neural networks. Im Proceedings of the Future Technologies Conference; Springer: Cham, Switzerland, 2019; pp. 750–765.
3. Kowsari K Diagnosis and Analysis of Celiac Disease and Environmental Enteropathy on Biopsy Images using Deep Learning Approaches. Ph.D. Thesis, University of California, Los Angeles, CA, USA, 2020; doi:10.18130/v3-837s-3a79.
4. Kowsari K; Jafari Meimandi K; Heidarysafa M; Mendu S; Barnes L; Brown D Text Classification Algorithms: A Survey. Information 2019, 10, 150. doi:10.3390/info10040150.
5. Litjens G; Kooi T; Bejnordi BE; Setio AAA; Ciompi F; Ghafoorian M; Van Der Laak JA; Van Ginneken B; Sánchez CI A survey on deep learning in medical image analysis. Med. Image Anal 2017, 42, 60–88. [PubMed: 28778026]
6. Nobles AL; Glenn JJ; Kowsari K; Teachman BA; Barnes LE Identification of imminent suicide risk among young adults using text messages. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; ACM: New York, NY, USA, 2018; p. 413.
7. Zhai S; Cheng Y; Zhang ZM; Lu W Doubly convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 1082–1090.
8. Hegde RB; Prasad K; Hebbar H; Singh BMK Comparison of traditional image processing and deep learning approaches for classification of white blood cells in peripheral blood smear images. Biocybern. Biomed. Eng 2019, 39, 382–392.
9. Zhang J; Kowsari K; Harrison JH; Lobo JM; Barnes LE Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record. IEEE Access 2018, 6, 65333–65346.
10. Pavik I; Jaeger P; Ebner L; Wagner CA; Petzold K; Spichtig D; Poster D; Wüthrich RP; Russmann S; Serra AL Secreted Klotho and FGF23 in chronic kidney disease Stage 1 to 5: A sequence suggested from a cross-sectional study. Nephrol. Dial. Transplant 2013, 28, 352–359. [PubMed: 23129826]

11. Kowsari K; Brown DE; Heidarysafa M; Meimandi KJ; Gerber MS; Barnes LE Hdltex: Hierarchical deep learning for text classification. In Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18–21 December 2017; pp. 364–371.
12. Dumais S; Chen H Hierarchical classification of web content. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 24–28 July 2000; pp. 256–263.
13. Yan Z; Piramuthu R; Jagadeesh V; Di W; Decoste D Hierarchical Deep Convolutional Neural Network for Image Classification. U.S. Patent 10,387,773, 20 8 2019.
14. Seo Y; Shin KS Hierarchical convolutional neural networks for fashion image classification. Expert Syst. Appl 2019, 116, 328–339.
15. Ranjan N; Machingal PV; Jammalmadka SSD; Thenaknidiyoor V; Dileep A Hierarchical Approach for Breast cancer Histopathology Images Classification. 2018. Available online: https://openreview.net/forum?id=rJlGvTojG (accessed on 10 January 2019).
16. Zhu X; Bain M B-CNN: Branch convolutional neural network for hierarchical classification. arXiv 2017, arXiv:1709.09890.
17. Sali R; Adewole S; Ehsan L; Denson LA; Kelly P; Amadi BC; Holtz L; Ali SA; Moore SR; Syed S; et al. Hierarchical Deep Convolutional Neural Networks for Multi-category Diagnosis of Gastrointestinal Disorders on Histopathological Images. arXiv 2020, arXiv:2005.03868.
18. Syed S; Ali A; Duggan C Environmental enteric dysfunction in children: A review. J. Pediatr. Gastroenterol. Nutr 2016, 63, 6. [PubMed: 26974416]
19. Naylor C; Lu M; Haque R; Mondal D; Buonomo E; Nayak U; Mychaleckyj JC; Kirkpatrick B; Colgate R; Carmolli M; et al. Environmental enteropathy, oral vaccine failure and growth faltering in infants in Bangladesh. EBioMedicine 2015, 2, 1759–1766. [PubMed: 26870801]
20. Husby S; Koletzko S; Korponay-Szabó IR; Mearin ML; Phillips A; Shamir R; Troncone R; Giersiepen K; Branski D; Catassi C; et al. European Society for Pediatric Gastroenterology, Hepatology, and Nutrition guidelines for the diagnosis of coeliac disease. J. Pediatr. Gastroenterol. Nutr 2012, 54, 136–160. [PubMed: 22197856]
21. Fasano A; Catassi C Current approaches to diagnosis and treatment of celiac disease: An evolving spectrum. Gastroenterology 2001, 120, 636–651. [PubMed: 11179241]
22. Hou L; Samaras D; Kurc TM; Gao Y; Davis JE; Saltz JH Patch-based convolutional neural network for whole slide tissue image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2424–2433.
23. Goodfellow I; Bengio Y; Courville A; Bengio Y Deep Learning; MIT Press: Cambridge, MA, USA, 2016; Volume 1.
24. Wang W; Huang Y; Wang Y; Wang L Generalized autoencoder: A neural network framework for dimensionality reduction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 490–497.
25. Rumelhart DE; Hinton GE; Williams RJ Learning Internal Representations by Error Propagation; Technical Report; California Univ San Diego La Jolla Inst for Cognitive Science: La Jolla, CA, USA, 1985.
26. Liang H; Sun X; Sun Y; Gao Y Text feature extraction based on deep learning: A review. EURASIP J. Wirel. Commun. Netw 2017, 2017, 211. [PubMed: 29263717]
27. Masci J; Meier U; Cireşan D; Schmidhuber J Stacked convolutional auto-encoders for hierarchical feature extraction. In International Conference on Artificial Neural Networks; Springer: Berlin/Heidelberg, Germany, 2011; pp. 52–59.
28. Chen K; Seuret M; Liwicki M; Hennebert J; Ingold R Page segmentation of historical document images with convolutional autoencoders. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR); IEEE: Washington, DC, USA, 2015; pp. 1011–1015.
29. Geng J; Fan J; Wang H; Ma X; Li B; Chen F High-resolution SAR image classification via deep convolutional autoencoders. IEEE Geosci. Remote. Sens. Lett 2015, 12, 2351–2355.
30. Jain AK Data clustering: 50 years beyond K-means. Pattern Recognit. Lett 2010, 31, 651–666.

31. Gao Q; Xu HX; Han HG; Guo M Soft-sensor Method for Surface Water Qualities Based on Fuzzy Neural Network. In Proceedings of the 2019 Chinese Control Conference (CCC), Guangzhou, China, 27–30 July 2019; pp. 6877–6881.
32. Kowsari K; Yammahi M; Bari N; Vichr R; Alsaby F; Berkovich SY Construction of fuzzyfind dictionary using golay coding transformation for searching applications. arXiv 2015, arXiv:1503.06483.
33. Kowsari K; Alassaf MH Weighted unsupervised learning for 3d object detection. arXiv 2016, arXiv:1602.05920.
34. Alassaf MH; Kowsari K; Hahn JK Automatic, real time, unsupervised spatio-temporal 3d object detection using rgb-d cameras. In Proceedings of the 2015 19th International Conference on Information Visualisation, Barcelona, Spain, 22–24 July 2015; pp. 444–449.
35. Manning CD; Raghavan P; Schutze H Introduction to Information Retrieval; Cambridge University Press: Cambridge, UK, 2008; Volume 20, pp. 405–416.
36. Mahajan M; Nimbhorkar P; Varadarajan K The Planar k-Means Problem is NP-Hard. In WALCOM: Algorithms and Computation; Das S, Uehara R, Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 274–285.
37. Fischer AH; Jacobson KA; Rose J; Zeller R Hematoxylin and eosin staining of tissue and cell sections. Cold Spring Harb. Protoc 2008, 2008, pdb–prot4986.
38. Anderson J An introduction to Routine and special staining. Retrieved 8 2011, 18, 2014.
39. Khan AM; Rajpoot N; Treanor D; Magee D A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans. Biomed. Eng 2014, 61, 1729–1738. [PubMed: 24845283]
40. Bianco S; Cusano C; Napoletano P; Schettini R Improving CNN-Based Texture Classification by Color Balancing. J. Imaging 2017, 3, 33.
41. Bianco S; Schettini R Error-tolerant color rendering for digital cameras. J. Math. Imaging Vis 2014, 50, 235–245.
42. Vahadane A; Peng T; Sethi A; Albarqouni S; Wang L; Baust M; Steiger K; Schlitter AM; Esposito I; Navab N Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans. Med Imaging 2016, 35, 1962–1971. [PubMed: 27164577]
43. Nair V; Hinton GE Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
44. Kowsari K; Heidarysafa M; Brown DE; Meimandi KJ; Barnes LE Rmdl: Random multimodel deep learning for classification. In Proceedings of the 2nd International Conference on Information System and Data Mining, Lakeland, FL, USA, 9–11 April 2018; pp. 19–28.
45. Li Q; Cai W; Wang X; Zhou Y; Feng DD; Chen M Medical image classification with convolutional neural network. In Proceedings of the 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), Singapore, 10–12 December 2014; pp. 844–848.
46. Heidarysafa M; Kowsari K; Brown DE; Jafari Meimandi K; Barnes LE An Improvement of Data Classification Using Random Multimodel Deep Learning (RMDL). arXiv 2018, arXiv:1808.08121, doi:10.18178/ijmlc.2018.8.4.703.
47. Scherer D; Müller A; Behnke S Evaluation of pooling operations in convolutional architectures for object recognition. In Proceedings of the Artificial Neural Networks–ICANN 2010, Thessaloniki, Greece, 15–18 September 2010; pp. 92–101.
48. Kingma D; Ba J Adam: A method for stochastic optimization. arXiv 2014, arXiv:1412.6980.
49. Chollet F Keras: Deep Learning Library for Theano and Tensorflow. 2015. Available online: https://keras.io/ (accessed on 19 August 2019).
50. Srivastava N; Hinton G; Krizhevsky A; Sutskever I; Salakhutdinov R Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res 2014, 15, 1929–1958.
51. Yang Y An evaluation of statistical approaches to text categorization. Inf. Retr 1999, 1, 69–90.
52. Lever J; Krzywinski M; Altman N Points of significance: Classification evaluation. Nat. Methods 2016, 13, 603–604.

53. Abadi M; Agarwal A; Barham P; Brevdo E; Chen Z; Citro C; Corrado GS; Davis A; Dean J; Devin M; et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv 2016, arXiv:1603.04467.

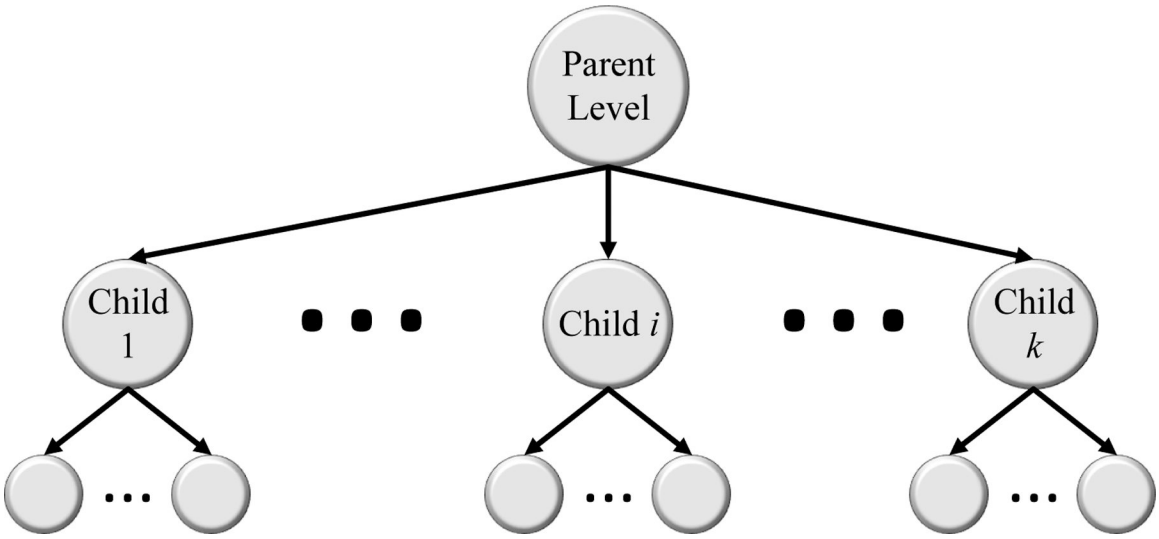

Figure 1.
HMIC: Hierarchical Medical Image Classification.

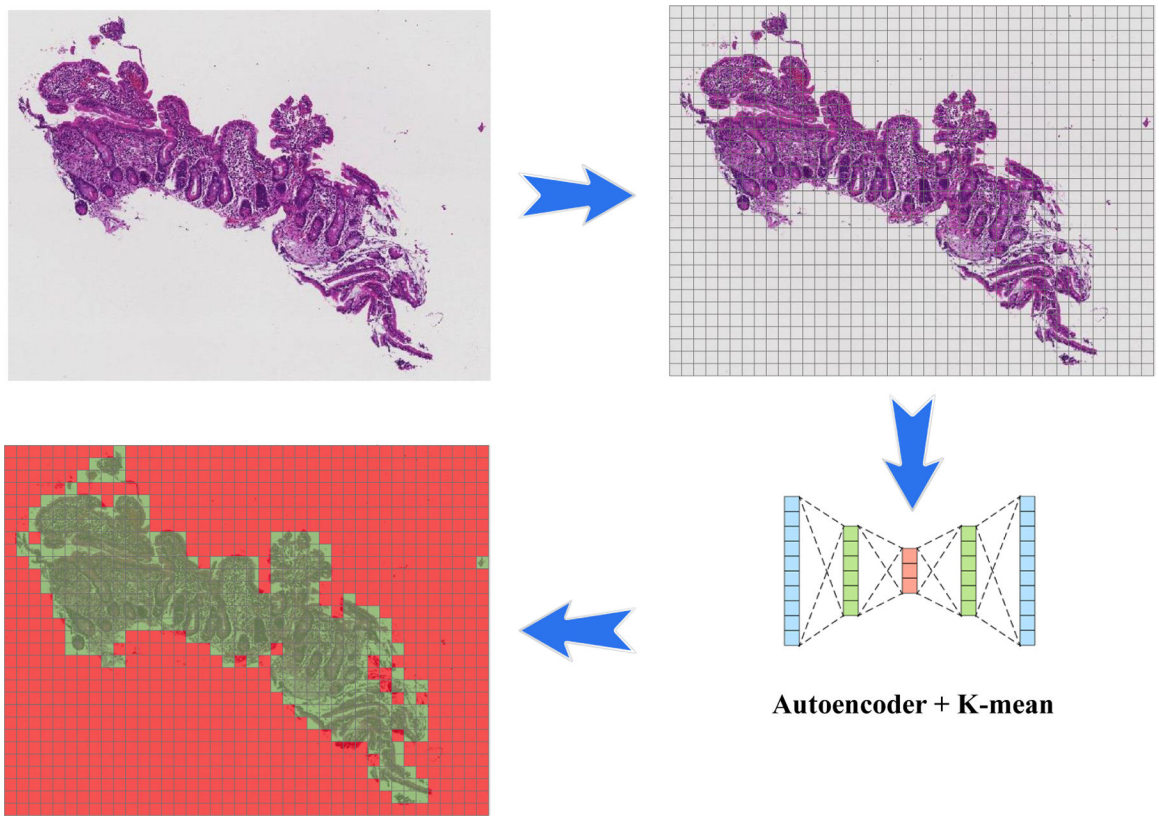

Figure 2.

Pipeline of patching and applying an autoencoder to find useful patches for the training model. The biopsy images are very large, so we need to divide into smaller patches to be used in the machine learning model. As you can see in the image, many of these patches are empty. After using an autoencoder, we can apply a clustering algorithm to discard useless patches (green patches contain useful information, while red patches do not).

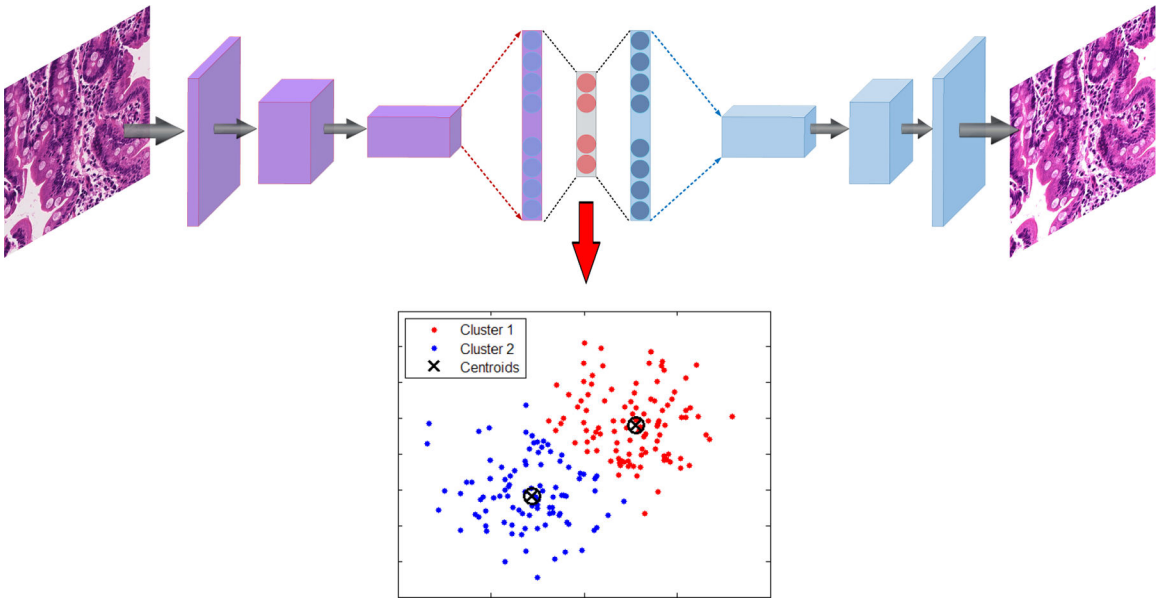

Figure 3.
Example autoencoder architecture with K-means applied on the bottle-neck layer feature vector to cluster useful and not useful patches.

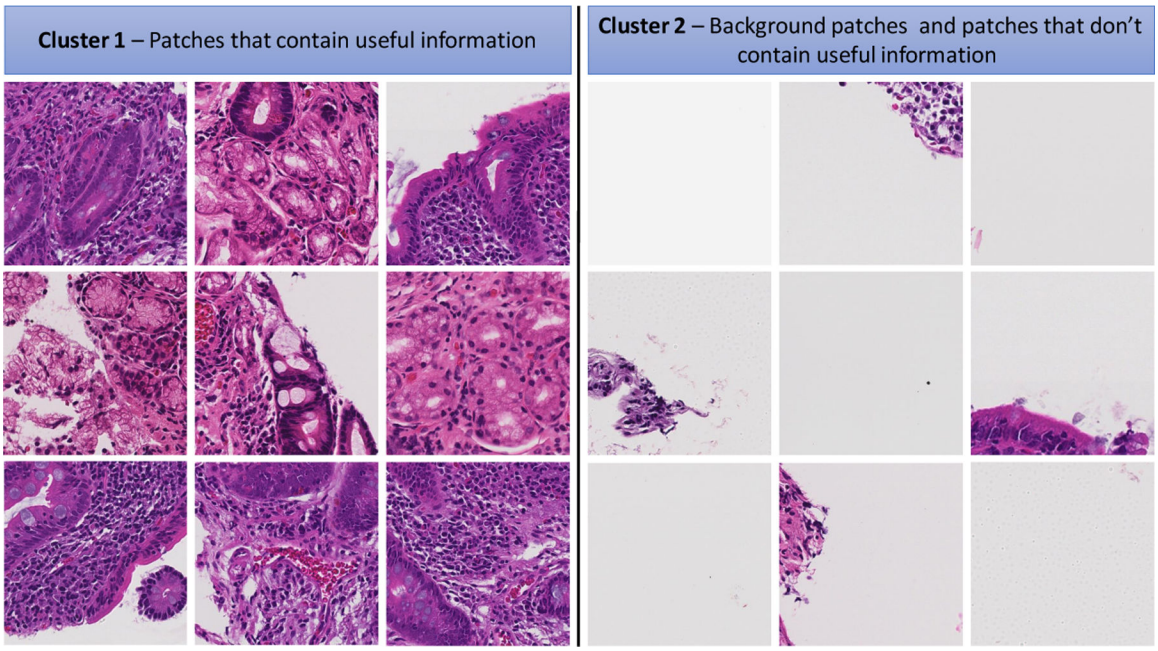

Figure 4.

Some samples of clustering results—cluster 1 includes patches with useful information and cluster 2 includes patches without useful information (mostly created from background parts of WSIs).

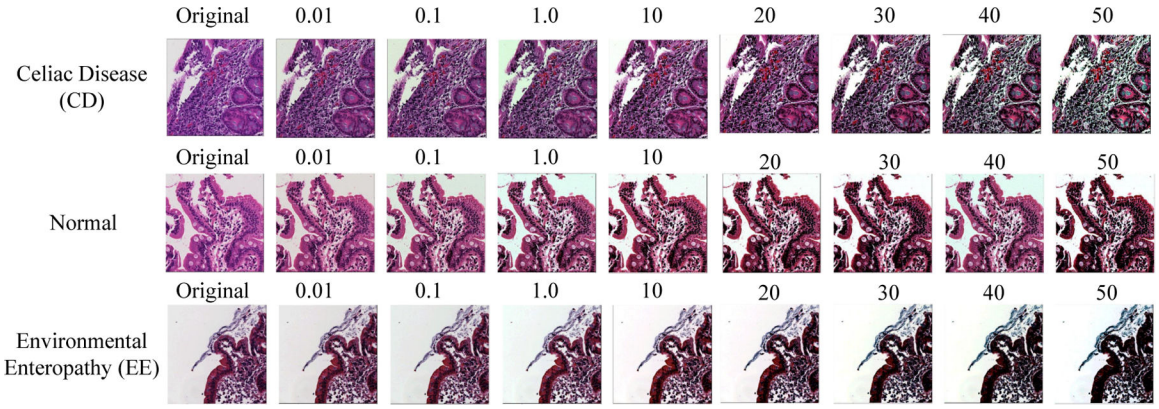

Figure 5.
Color Balancing samples for the three classes.

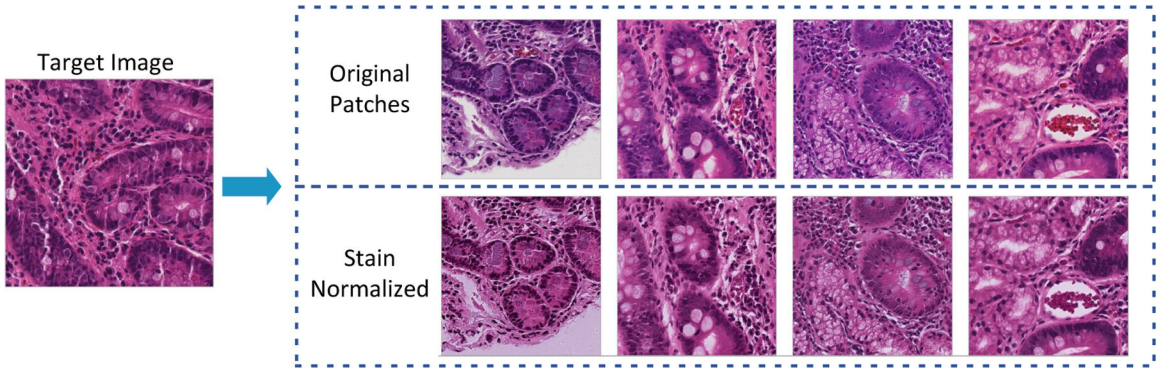

Figure 6.

Stain normalization results when using the method proposed by Vahadane et al. [42]. Images in the first row represent the source images. The source images are normalized images to the stain appearance of the target image in second row [1].

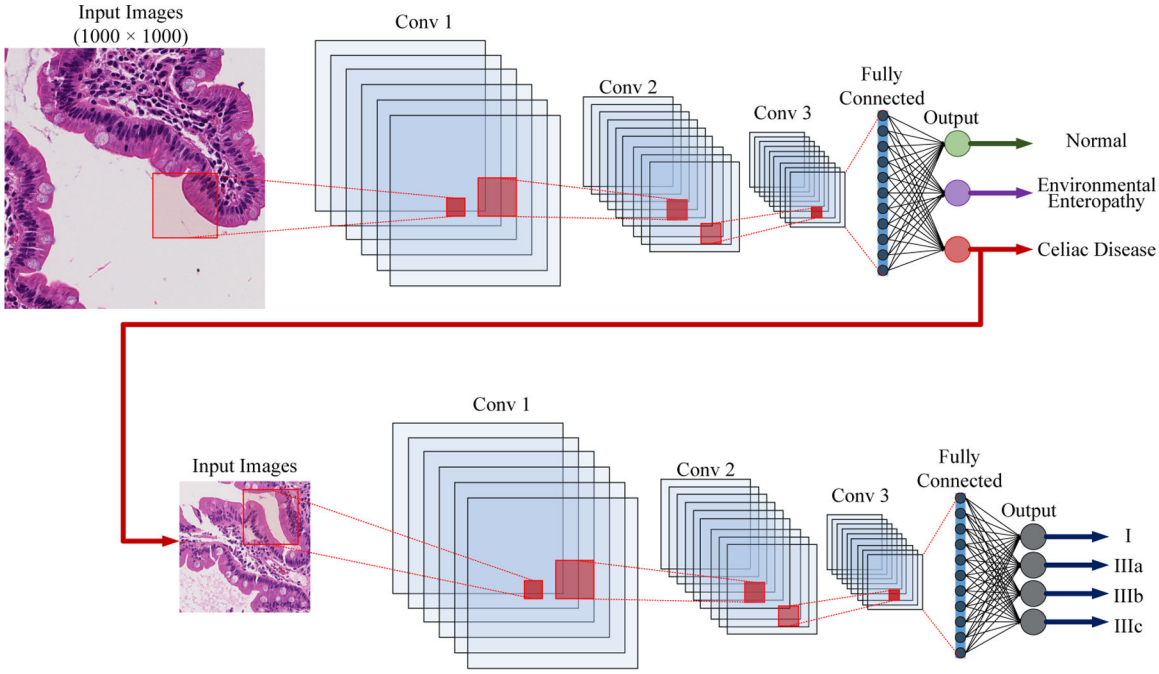

Figure 7.
Structure of Convolutional Neural Net using multiple 2D feature detectors and 2D max-pooling.

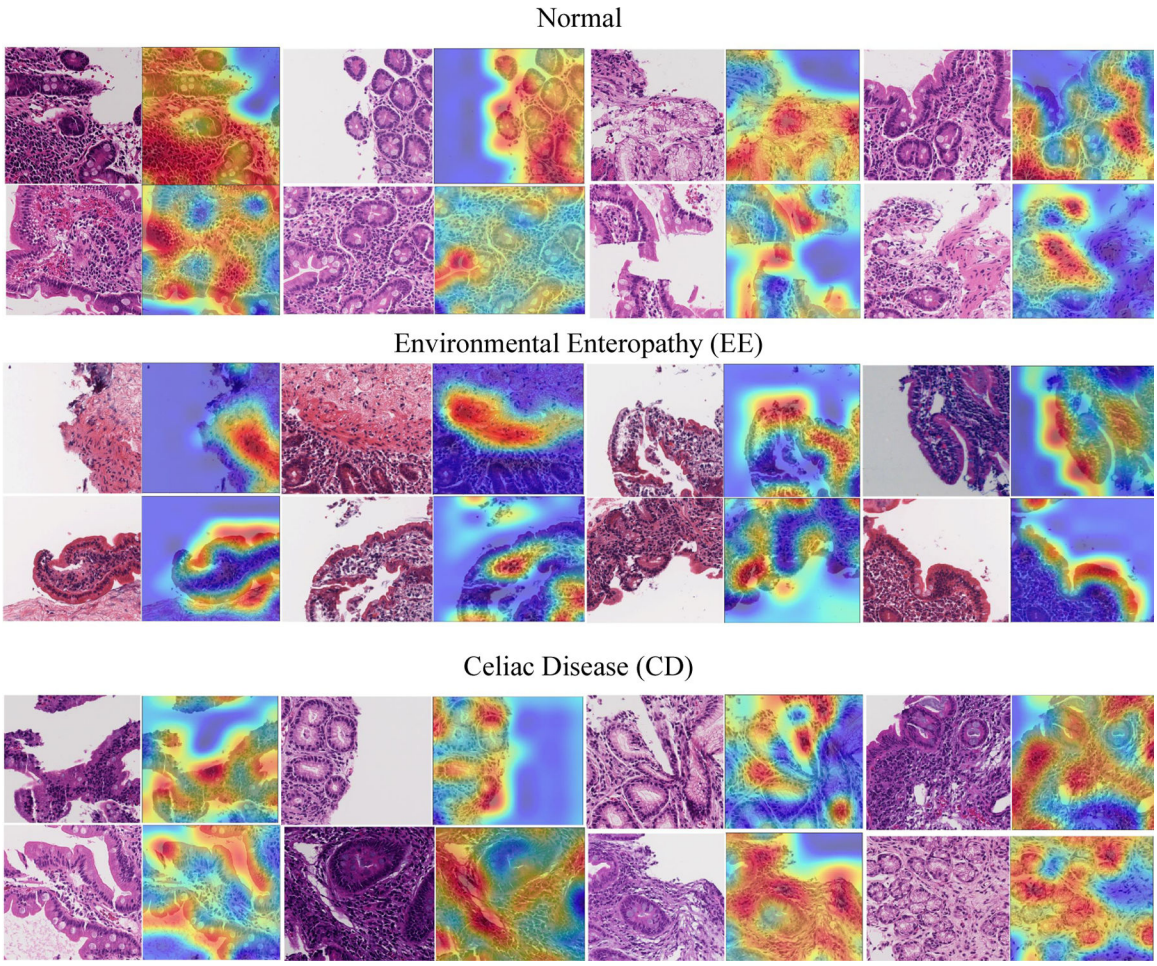

Figure 8.
Grad-CAM results for showing feature importance.

Table 1.

Population results of biopsies dataset.

	Total Population	**Pakistan**	**Zambia**	**US**	
Data	150	EE (n = 10)	EE (n = 16)	Celiac (n = 63)	Normal (n = 61)
Biopsy Images	461	29	19	239	174
Age, median (IQR), months	37.5 (19.0 to 121.5)	22.2 (20.8 to 23.4)	16.5 (9.5 to 21.0)	130.0 (85.0 to 176.0)	25.0 (16.5 to 41.0)
Gender, n (%)	M = 77 (%51.3) F = 73 (%48.7)	M = 5 (%50) F = 5 (%50)	M = 10 (%62.5) F = 6 (%37.5)	M = 29 (%46) F = 34 (%54)	M = 33 (%54) F = 28 (%46)
LAZ/ HAZ, median (IQR)	−0.6 (−1.9 to 0.4)	−2.8 (−3.6 to −2.3)	−3.1 (−4.1 to −2.2)	−0.3 (−0.8 to 0.7)	−0.2 (−1.3 to 0.5)

Table 2.

Dataset used for Hierarchical Medical Image Classification (HMIC).

Data		Train		Test		Total	
Normal		22,676		9717		32,393	
Environmental Enteropathy		20,516		8792		29,308	
		Parent	Child	Parent	Child	Parent	Child
Celiac Disease	I	21,140	4988	9058	2137	30,198	7125
	IIIa		4790		2052		6842
	IIIb		5684		2436		8120
	IIIc		5678		2433		8111

Table 3.

Result of parent level classifications for normal, environmental enteropathy, and Celiac disease.

	Precision	Recall	F1-Score
Normal	89.97 ± 0.59	89.35 ± 0.61	89.66 ± 0.60
Environmental Enteropathy	94.02 ± 0.49	97.30 ± 0.33	95.63 ± 0.42
Celiac Disease	91.12 ± 0.32	88.71 ± 0.35	89.90 ± 1.27

Table 4.

Results of HMIC with comparison with our baseline.

	Model	Precision	Recall	F1-Score
Baseline	CNN	76.76 ± 0.49	80.18 ± 0.47	78.43 ± 0.48
	Multilayer perceptron	76.19 ± 0.50	79.40 ± 0.47	77.76 ± 0.49
	Deep CNN	82.95 ± 0.44	87.28 ± 0.39	85.06 ± 0.42
HMIC	Non Whole slide	84.13 ± 0.37	93.56 ± 0.29	88.61 ± 0.37
	Whole slide	**88.01 ± 0.38**	**93.98 ± 0.28**	**90.89 ± 0.38**

Table 5.

Results per-classed of HMIC with comparison with our baseline.

Model				Precision	Recall	F1-Score
Baseline	CNN	Normal		87.83 ± 0.57	90.77 ± 0.65	89.28 ± 0.61
		Environmental Enteropathy		90.93 ± 0.61	82.48 ± 0.79	86.50 ± 0.71
		Celiac Disease	I	68.37 ± 1.98	68.62 ± 1.96	68.50 ± 1.96
			IIIa	56.26 ± 1.01	56.26 ± 2.21	59.29 ± 1.95
			IIIb	65.28 ± 0.97	98.28 ± 2.01	66.64 ± 1.87
			IIIc	62.66 ± 1.99	66.83 ± 1.99	64.68 ± 2.02
	Multilayer perceptron	Normal		87.97 ± 0.76	81.87 ± 0.76	84.81 ± 0.71
		Environmental Enteropathy		87.25 ± 0.69	90.18 ± 0.62	88.69 ± 0.66
		Celiac Disease	I	57.92 ± 2.07	60.74 ± 2.07	59.30 ± 2.09
			IIIa	62.58 ± 2.09	62.18 ± 2.09	60.89 ± 2.11
			IIIb	65.00 ± 1.89	66.09 ± 1.87	65.56 ± 1.88
			IIIc	67.97 ± 1.85	74.85 ± 1.72	71.24 ± 1.78
	DCNN	Normal		95.14 ± 0.42	94.91 ± 0.43	95.14 ± 0.42
		Environmental Enteropathy		92.22 ± 0.55	90.62 ± 0.60	91.52 ± 0.58
		Celiac Disease	I	75.41 ± 1.82	72.63 ± 1.89	73.99 ± 1.85
			IIIa	70.81 ± 1.92	72.47 ± 1.93	71.63 ± 1.79
			IIIb	81.08 ± 0.81	74.67 ± 1.84	77.74 ± 1.65
			IIIc	75.07 ± 1.83	76.37 ± 1.81	75.71 ± 1.81
HMIC	Non Whole Slide	Normal		89.97 ± 0.59	89.35 ± 0.61	89.66 ± 0.61
		Environmental Enteropathy		94.02 ± 0.49	97.30 ± 0.33	95.63 ± 0.33
		Celiac Disease	I	83.25 ± 1.58	80.91 ± 1.66	82.06 ± 1.62
			IIIa	80.34 ± 1.62	80.46 ± 1.71	80.40 ± 1.57
			IIIb	85.35 ± 1.49	81.77 ± 1.67	83.52 ± 1.47
			IIIc	85.54 ± 1.49	82.71 ± 1.60	84.10 ± 1.55
	Whole Slide	Normal		90.64 ± 0.57	90.06 ± 0.57	90.35 ± 0.58
		Environmental Enteropathy		94.08 ± 0.49	97.33 ± 0.42	98.68 ± 0.42
		Celiac Disease	I	88.73 ± 1.34	85.07 ± 1.51	86.86 ± 1.43
			IIIa	81.19 ± 1.65	81.19 ± 1.65	82.44 ± 1.51
			IIIb	90.51 ± 1.24	90.48 ± 1.27	90.49 ± 1.16
			IIIc	89.26 ± 1.31	90.18 ± 1.26	89.72 ± 1.28