Harnessing artificial intelligence to infer novel spatial biomarkers for the diagnosis of eosinophilic esophagitis

Ariel Larey[1,2†], Eliel Aknin[1,3†], Nati Daniel[1†], Garrett A. Osswald[4], Julie M. Caldwell[4], Mark Rochman[4], Tanya Wasserman[1], Margaret H. Collins[5], Nicoleta C. Arva[6], Guang-Yu Yang[7], Marc E. Rothenberg[4] and Yonatan Savir[1]*

[1]Department of Physiology, Biophysics and Systems Biology, Faculty of Medicine, Technion – Israel Institute of Technology, Haifa, Israel, [2]Faculty of Computer Science, Technion Israel Institute of Technology, Haifa, Israel, [3]Faculty of Industrial Engineering, Technion – Israel Institute of Technology, Haifa, Israel, [4]Division of Allergy and Immunology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, United States, [5]Division of Pathology, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, United States, [6]Department of Pathology and Laboratory Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States, [7]Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

Eosinophilic esophagitis (EoE) is a chronic allergic inflammatory condition of the esophagus associated with elevated esophageal eosinophils. Second only to gastroesophageal reflux disease, EoE is one of the leading causes of chronic refractory dysphagia in adults and children. EoE is a clinicopathologic disorder and the histological portion of the diagnosis requires enumerating the density of esophageal eosinophils in esophageal biopsies, and evaluating additional features such as basal zone hyperplasia is helpful. However, this task requires time-consuming, somewhat subjective manual analysis, thus reducing the ability to process the complex tissue structure and infer its relationship with the patient's clinical status. Previous artificial intelligence (AI) approaches that aimed to improve histology-based diagnosis focused on recapitulating identification and quantification of the area of maximal eosinophil density, the gold standard manual metric for determining EoE disease activity. However, this metric does not account for the distribution of eosinophils or other histological features, over the whole slide image. Here, we developed an artificial intelligence platform that infers local and spatial biomarkers based on semantic segmentation of intact eosinophils and basal zone distributions. Besides the maximal density of eosinophils [referred to as Peak Eosinophil Count (PEC)] and a maximal basal zone fraction, we identify the value of two additional metrics that reflect the distribution of eosinophils and basal zone fractions. This approach enables a decision support system that predicts EoE activity and potentially classifies the histological severity of EoE patients. We utilized a cohort that includes 1,066 biopsy slides from 400 subjects to validate the system's performance and achieved a histological severity classification accuracy of 86.70%, sensitivity of 84.50%, and specificity of 90.09%.

Our approach highlights the importance of systematically analyzing the distribution of biopsy features over the entire slide and paves the way toward a personalized decision support system that will assist not only in counting cells but can also potentially improve diagnosis and provide treatment prediction.

1. Introduction

Eosinophilic esophagitis (EoE) is a chronic immune system disease associated with esophageal tissue inflammation and injury characterized by a large number of eosinophils, which are found in the lining of the esophagus, called the esophageal mucosa (1). EoE is allergen-driven and mainly caused by a reaction to food (2). The damaged esophageal tissue leads to symptoms, such as pain and trouble swallowing (3). In particular, EoE is becoming a more common cause of dysphagia in adults and vomiting, failure to thrive, and abdominal pain in children (3). EoE can be treated by dietary restriction, proton pump inhibitor (PPI) (4) therapy or topical steroids, and in more severe conditions, an endoscopic dilation intervention, specifically stricture dilation, is used.

Currently, the diagnosis of EoE relies on performing an upper endoscopy and obtaining esophageal mucosal biopsies. The hematoxylin and eosin (H&E) stained slides (5) are examined by pathologists. The physicians typically manually examine the slide using a microscope, identify the area of the tissue with the greatest eosinophil density, and count the number of intact eosinophils in that high-power field (HPF), i.e., the peak eosinophil count (PEC). The gold standard, histologic criterion, to date, is to define patients with EoE as having active disease if their $PEC \geq 15$ (6).

Yet, the PEC score captures only the maximal eosinophil count and not other properties such as the distribution of the eosinophils within the tissue, and it does not account for other cellular features that are captured by the EoE histology scoring system (EoEHSS) (7). This method includes eight features that are relevant to EoE and accounts not only for the maximal severity of these features, but also for their distribution. This includes, for example, quantifying the percentage of HPFs within the slide that exceed the threshold of ≥ 15 eosinophils. However, estimating such a metric visually poses a significant challenge. Another example of the importance of accounting for features in addition to the maximal eosinophil count is the development of a histological severity score that was used to diagnose remission (EoEHRS) (8). In this case, both $PEC < 15/HPF$ and total grade and stage scores from all EoEHSS features ≤ 3 are required to define remission.
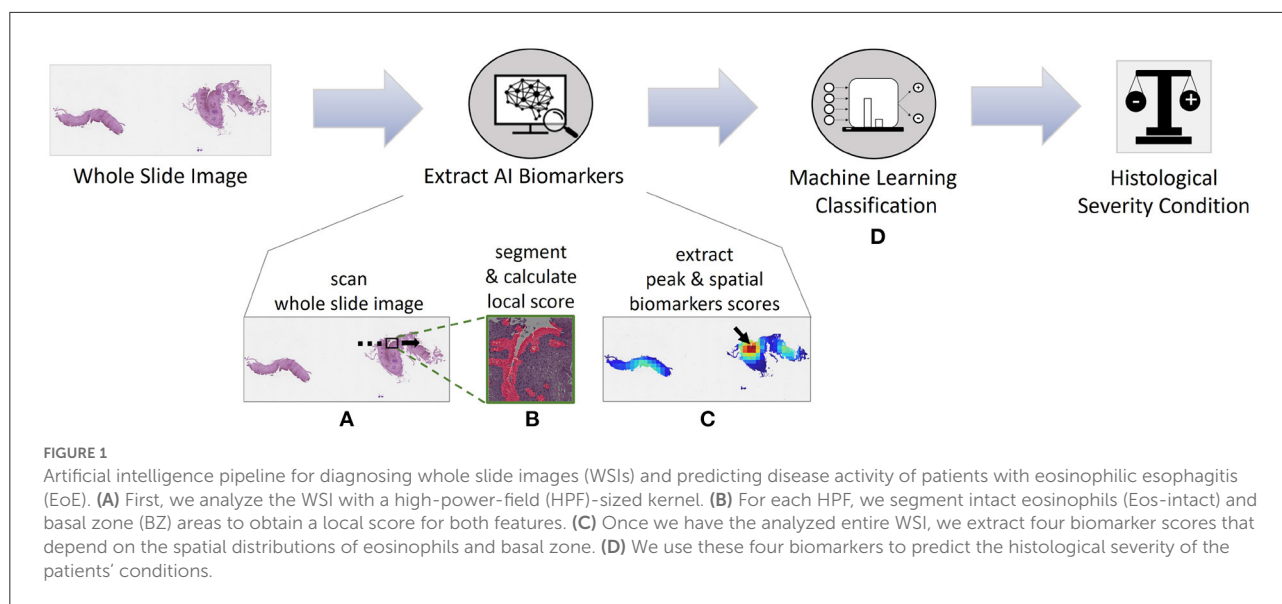
Whereas processing the features of the entire whole slide improves diagnostic metrics, current manual approaches limit it. Counting PEC and scoring EoE histology is time-consuming, requires trained personnel, and can lead to variability between pathologists upon EoE biopsy diagnosis (6, 9, 10). Hence, in recent years, considerable effort has been dedicated to build a robust and trustworthy process of inferring pathological biomarkers in health and disease. This includes harnessing machine learning in general and deep learning specifically (11–20). We have recently applied a dual approach toward diagnosing EoE: the first one is assigning a global label for the pathology images that is based on the patient condition (21). The second one is based on segmenting and counting inflammatory cells, such as Intact eosinophils and Not-Intact eosinophils for EoE biopsy diagnosis using a deep convolutional neural network (DCNN) (22).

Here, we developed an artificial intelligence (AI) approach using machine learning for extracting novel biomarkers and used it to predict the histological severity condition (Figure 1). The pipeline has a state-of-the-art segmentation performance with a mean intersection over union metric (mIoU) score of 83.85% based on basal zone (BZ) and intact eosinophils (Eos-Intact) features. We show that derived biomarkers significantly correlate with manually obtained HSS scores. Using a cohort of 1,066 biopsy slides from 400 patients, we demonstrate that AI biomarkers estimate histological severity achieving an accuracy of 86.70%, sensitivity of 84.50%, and specificity of 90.09%.

2. Materials and methods

2.1. Dataset and clinical scores

The dataset is part of the Consortium of Eosinophilic Gastrointestinal Disease Researchers (CEGIR) (23), a national collaborative network in the U.S. of 16 academic centers caring for adults and children with eosinophilic gastrointestinal disorders. The institutional review boards approved this study of the participating institutions *via* a central institutional review board at Cincinnati Children's Hospital Medical Center (CCHMC IRB protocol 2015-3613). Participants provided

FIGURE 1
Artificial intelligence pipeline for diagnosing whole slide images (WSIs) and predicting disease activity of patients with eosinophilic esophagitis (EoE). **(A)** First, we analyze the WSI with a high-power-field (HPF)-sized kernel. **(B)** For each HPF, we segment intact eosinophils (Eos-intact) and basal zone (BZ) areas to obtain a local score for both features. **(C)** Once we have the analyzed entire WSI, we extract four biomarker scores that depend on the spatial distributions of eosinophils and basal zone. **(D)** We use these four biomarkers to predict the histological severity of the patients' conditions.

written informed consent. The dataset contains subjects with a history of EoE undergoing endoscopy (EGD) for standard-of-care purposes ($n = 419$). Distal, mid, or proximal esophageal biopsies (1–3 per anatomical site) per patient were placed in 10% formalin; the tissue was then processed and embedded in paraffin. Sections (4 μm) were mounted on glass slides and subjected to hematoxylin and eosin (H&E) staining. Slides were scanned on the Aperio scanner at 400X magnification and were saved in SVS format. Each slide of esophageal tissue was analyzed by an anatomic pathologist who is a member of the CEGIR central pathology core. In addition to determining peak eosinophil count per 400X HPF (PEC), the pathologist subjected each slide to eosinophilic esophagitis histological scoring system (EoE HSS) analysis to assess the severity (grade) and extent (stage) of a set of histological abnormalities using a 4 point scale (0 normal; 3 maximum change) [7]. These features included eosinophilic inflammation (EI), basal zone hyperplasia (BZH), dilated intercellular spaces (DIS), eosinophilic abscess (EA), eosinophil surface layering (SL), surface epithelial alteration (SEA), dyskeratotic epithelial cells (DEC), and lamina propria fibrosis (LPF) [7]. The BZH grade score is determined by the amount of total epithelial thickness occupied by the basal zone, where 0 indicates that BZH is not present, 1 indicates that basal zone occupies >15% but <33% of the total epithelial thickness, 2 indicates that the basal zone occupies 33–66% of the total epithelial thickness, and 3 indicates that the basal zone occupies >66% of the total epithelial thickness. The BZH stage score indicates the amount of biopsy that showed any degree of BZH, where 0 indicates that BZH is not present, 1 indicates that <33% of the epithelium exhibits any BZH with grade >0, 2 indicates that 33–66% of the epithelium exhibits any BZH with grade >0, and 3 indicates that >66% of the epithelium exhibits any BZH with grade >0 [7].

2.2. Semantic labeling

To train and validate the models, we labeled 23 patients' whole slide images (WSIs). The dataset consists of large WSIs with median length and width of 150,000 and 56,000 pixels, respectively. We cropped each WSI into small patches with a size of 1200×1200 pixels. Patches with a small amount of tissue, less than 15% of the patch area, were filtered. A total of $n = 10,170$ patches was used for semantic labeling. Those patches were analyzed and annotated by an expert using VIA [24] and then were verified by three different experts. For each patch, the intact eosinophils' centers and the basal zone area were marked. The result was two semantic masks. In the first, the pixels in the area of a circle with a radius of 25 pixels around the intact eosinophils center were labeled as Eos-Intact [22]. In the second, pixels within the marked basal zone polygons were labeled as BZ. That is, each pixel was classified either as a BZ type, Eos-Intact type, both of them, or as none. In total, about 570 million pixels were labeled as BZ, and about 78.47 million pixels were labeled as Eos-Intact. 8.6% of the images contained BZ, where their area was, on average, 45.45% of the image size. Eos-Intact were found in 22.8% of the images, with an average area fraction of 2.35%.

2.3. Semantic segmentation

We trained two models, one using the Eos-Intact masks and one using the BZ masks. For both models, the annotated patches were divided into two groups; 80% of the data were dedicated to training the segmentation model, and the rest, 20%, for testing the model. The segmentation model was based on the UNet++ architecture [25]. It was developed in the PyTorch framework [26] and was trained on a single NVIDIA

GeForce RTX 2080 Ti GPU. During the training phase, the 1200 × 1200-pixel patches were divided into 448 × 448-pixel sub-patches with an overlap of 72 pixels between them. Different sub-patch sizes were tested, and this size was optimal in terms of precision and recall (see segmentation metrics section of the systems). In addition, multiple hyperparameters were tested. The optimal parameters were batch size of 5, "Cosine Annealing" learning rate scheduler, and a 0.5 softmax threshold. The optimization loss function contains two terms, the Dice and Binary cross-entropy (BCE), where each term is weighted. After exploring different weights, we applied the weights 1 and 0.5 to the Dice and BCE, respectively. For inference, the test image was cropped into 448 × 448-pixel sub-patches as described above. To reduce segmentation noise, contiguous regions labeled as Eos-Intact or BZ that were smaller than an area of 1800 pixels, in the case of Eos-Intact, or area of 2007 (1% out of the sub-patch size), in the case of BZ, were re-labeled as none.

2.4. Semantic metrics

To estimate the segmentation performances, we used the following metrics,

$$mIoU = \frac{1}{I \cdot C} \sum_i \sum_c \frac{TP_{i,c}}{TP_{i,c} + FP_{i,c} + FN_{i,c}} \qquad (1)$$

$$mPrecision = \frac{1}{I \cdot C} \sum_i \sum_c \frac{TP_{i,c}}{TP_{i,c} + FP_{i,c}} \qquad (2)$$

$$mRecall = \frac{1}{I \cdot C} \sum_i \sum_c \frac{TP_{i,c}}{TP_{i,c} + FN_{i,c}} \qquad (3)$$

$$mSpecification = \frac{1}{I \cdot C} \sum_i \sum_c \frac{TN_{i,c}}{TN_{i,c} + FP_{i,c}} \qquad (4)$$

where the c index iterates over the different classes in the image, and the i index iterates over the different images in the dataset. C is the total number of classes, and I is the total number of images. TP, TN, FP, and FN are classification elements that denote true positive, true negative, false positive, and false negative of the areas of each image, respectively.

2.5. Calculating WSI AI scores

To evaluate the eosinophil and basal zone distribution within each WSI, we use an iterative process to scan over the entire slide. At each step, an image the size of a HPF is processed. The area of an HPF corresponds to a size of 2144 × 2144 pixels (548 × 548 μm). The stride step between constitutive HPFs is

500 pixels. Each HPF is divided into 25 sub-patches (448 × 448 pixels—corresponding to the network input size) with an overlap of 24 pixels. Each sub-patch is segmented and the HPF segmentation mask is assembled from them. The pixels' identity in the areas overlapping between sub-patches is determined by using OR function. After segmentation, each HPF is assigned two local scores: the number of intact eosinophils (22) and the BZ area rate, which is the ratio of the number of BZ pixels in the HPF mask, to the HPF size. After scanning the entire WSI, we produce score maps for both features—an Intact-Eosinophils map and a BZ map, where every pixel in these maps represents the score of the matching HPF. Based on the score maps, we can produce four WSI scores (Figure 1C):

- Peak Eosinophil Count (PEC)—The number of eosinophils in the HPF with the densest area of eosinophils within the WSI. This score is used in the clinic to diagnose active EoE (6, 22). A patient with a PEC greater than or equal to 15 is considered to have active EoE. The EI grade score is a proxy for this measure.
- Spatial Eosinophil Count (SEC)—The ratio of the number of HPFs with an Intact-Eosinophil count that is greater than or equal to 15 to the total number of HPFs in the feature map. The EI stage score is a proxy for this measure.
- Peak Basal Zone (PBZ)—The maximum HPF BZ area rate. This score is the maximal density of basal cells per HPF in the WSI. The BZH grade score is a proxy for this measure.
- Spatial Basal Zone (SBZ)—The ratio of the number of HPFs with local BZ score that is greater than or equal to 15% to the number of tissue HPFs in the feature map. The BZH stage score is a proxy to this measure.

2.6. Classifying whole slide image

2.6.1. Features-based classification

We previously presented a pipeline for classifying WSIs using only the predicted PEC directly (22). In this paper, we leverage the spatial information, for both eosinophils and basal cells that was revealed by segmenting the entire WSI. We used this information to devise four WSI scores and to predict the histological severity condition of the patient (Figure 1D). We explored different machine learning models—support vector machine (SVM), and linear discriminant analysis (LDA). In addition, various architectures of multi-layer perceptron (MLP) were examined, particularly, all combinations of layers in the size of 10, 20, 50, 100 tiled up to four hidden layers. We used these types of classifiers because of their better capability to handle tabular data (in contrast to convolutional-neural-networks, for example, that support sequential data). The cohort contains 1,066 WSIs that were not used for the segmentation training. Classifier training was done using 80% of the data, whereas the rest were used for validation. For
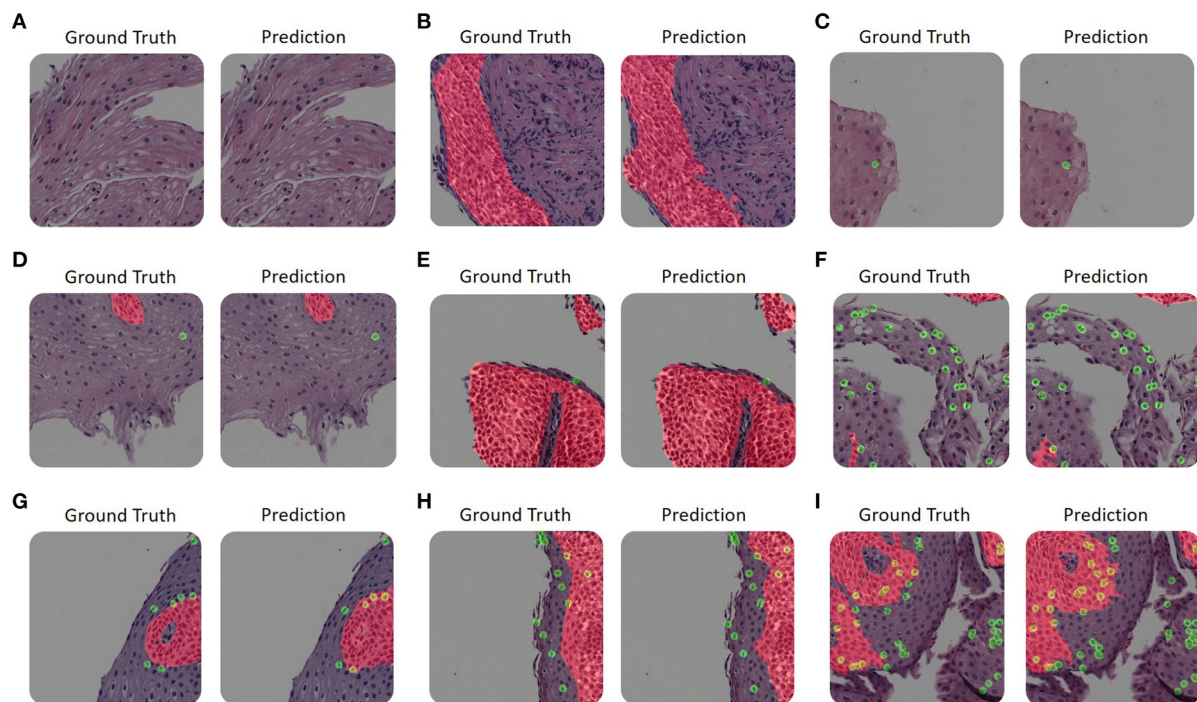
FIGURE 2
Examples of our platform semantic segmentation. **(A–I)** The size of each image is 1200 × 1200 pixels. Each panel's left-hand side is colored according to the ground truth as annotated by trained experts. The right-hand side is colored with its corresponding network prediction mask. Basal zone (BZ) pixels are colored with red, intact eosinophils (Eos-Intact) pixels are colored with green, and pixels associated with both (that is, eosinophils within a BZ area) are colored with yellow. **(A–C)** The upper row shows examples with only one label or none. **(D)** An example of an image that contains both a small number of basal cells and intact eosinophils. **(E)** An example of an image with a large basal zone and a small number of intact eosinophils. **(F)** An example that contains a small area of basal zone and a large number of intact eosinophils. **(G–I)** The bottom row displays examples with large basal zones and also a large number of intact eosinophils.

each model, we repeated the training procedure 20 times with different random seeds for splitting the data, and reported the median results.

2.6.2. Multi-classification

To improve the histological severity classification performance, different classifiers were used for regions having different eosinophil density. We define two regions of PEC scores,

$$classifier = \begin{cases} C_{in} & (PEC \geq 15 - \Delta) \text{ and } (PEC \leq 15 + \Delta) \\ C_{out} & (PEC < 15 - \Delta) \text{ or } (PEC > 15 + \Delta) \end{cases}$$
$$(5)$$

where C_{in} and C_{out} denote the classifier inside the window and outside of the window, respectively. The hyperparameter Δ defines the window size. The training procedure is as described above. To avoid bias, the contribution of each region to the 80%-20% split is proportional to the region size, ensuring that each region contributes points to the training and validation. We examined Δ values in the range of (1, 12).

3. Results

3.1. Local segmentation results

Figure 2 illustrates a few examples of our platform semantic segmentation compared with ground truth labeling by a trained researcher. Table 1 summarizes the segmentation metrics over the whole validation-set, 1, 2, 3, and 4.

3.2. WSI features scores

One of the main advantages of the described approach is that it allows scoring that is based not only on a limited number of regions probed by the pathologist but on the entire whole slide image (Figure 3). To process the entire whole slide image, we used dynamics convolution to scan the slide using windows with a HPF size with a stride of about 1/4 of the HPF size (Section 2.5). We computed the score maps for 1,066 WSIs from 400 patients that were not part of the semantic segmentation training and validation sets. The pipeline produces two feature-score maps for each WSI, one for the Eos-Intact score map and the second
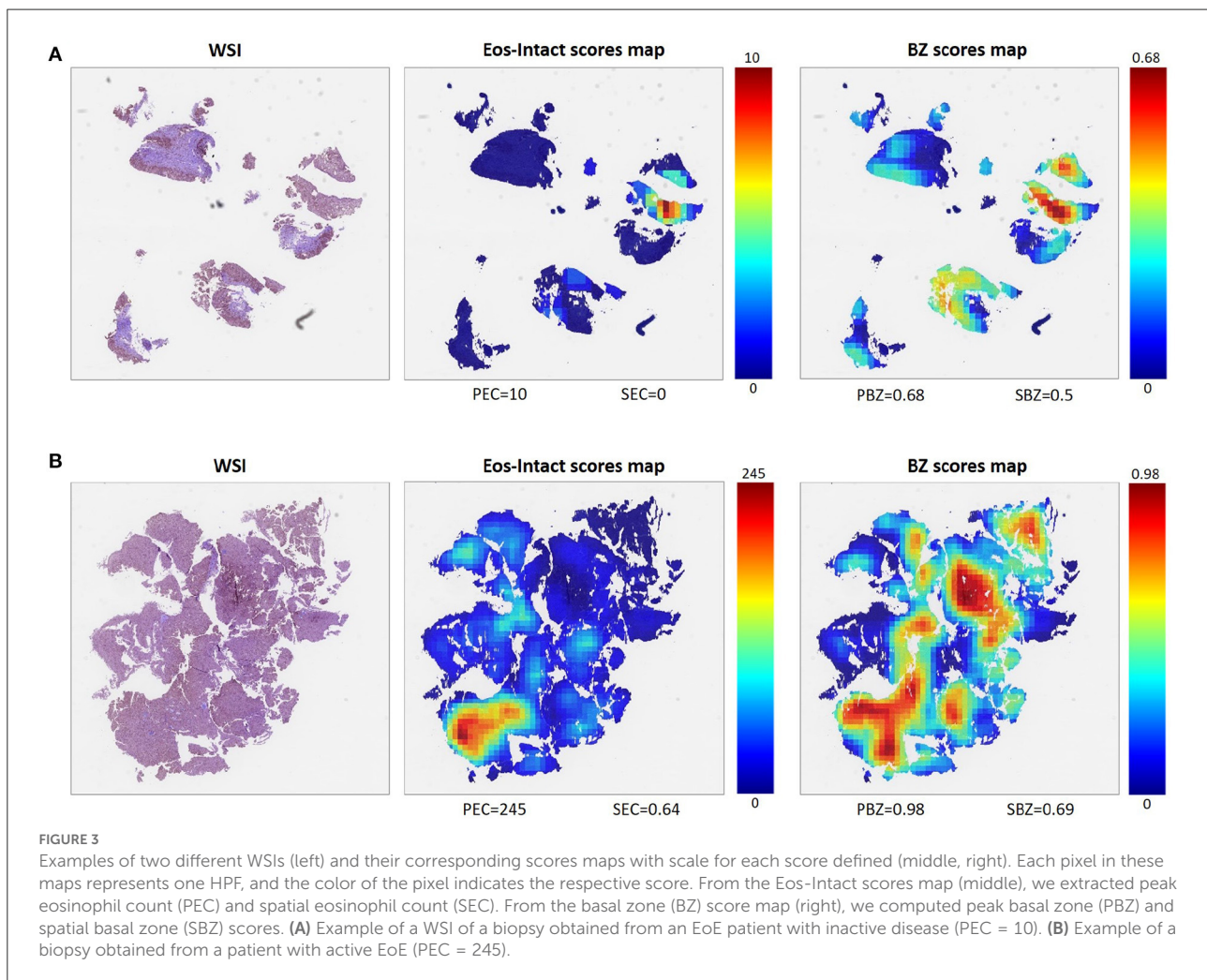
for the BZ score map. Figure 3 shows examples of two features score maps computed from two different WSIs. We computed four scores based on the semantic segmentation of the WSI;
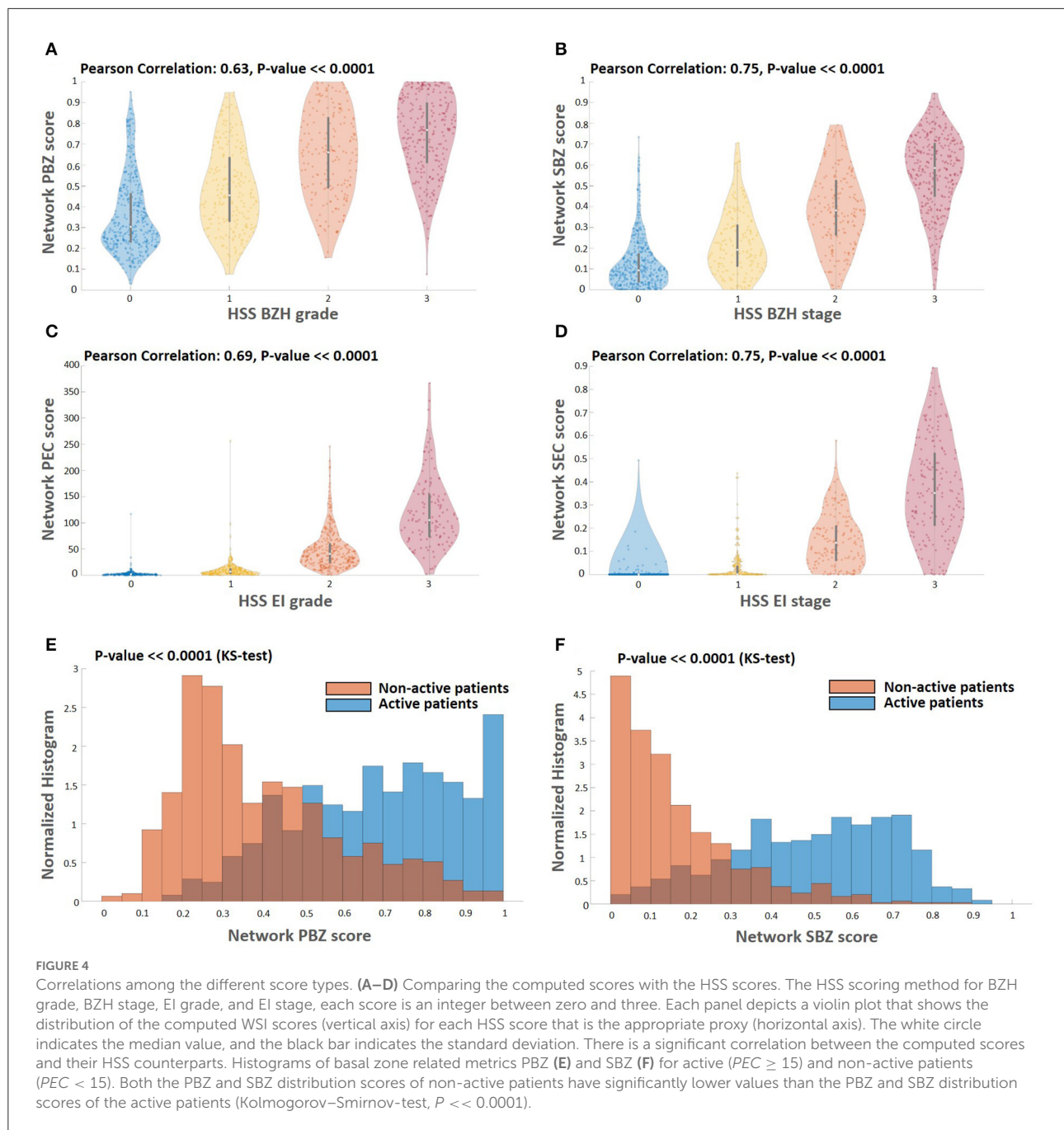
TABLE 1 Four segmentation metrics measured at the pixel level.

Metric	Eos-intact	BZ	Overall
mIoU (Equation 1)	0.93	0.75	0.84
mPrecision (Equation 2)	0.95	0.8	0.88
mRecall (Equation 3)	0.97	0.94	0.95
mSpecificity (Equation 4)	0.998	0.82	0.91

IoU denotes the Intersection Over Union between the Ground Truth and the prediction. Recall denotes the fraction of the True-Positive pixels among the total Ground Truth pixels in the image, whereas Precision denotes the fraction between the True-Positive pixels and the prediction pixels. The fraction between the True-Negative pixels and the total negative pixels in the image is coined Specificity. mIoU, mRecall, mPrecision, and mSpecificity are obtained by averaging IoU, Recall, Precision, and Specificity, respectively, over the validation set. The metrics are presented for the Eos-Intact and BZ classes separately in addition to their average per image as the overall score. The compared patches size is the network's input size—448×448 pixels.

this included two local ones (peak eosinophil counts [PEC] and peak basal zone [PBZ]), and two global ones (spatial eosinophil counts [SEC] and spatial basal zone [SBZ]) (Section 2.5). We compared the different WSI scores with the relevant HSS score estimated by the pathologists. We compared PBZ, SBZ, PEC, and SEC with HSS BZH grade, HSS BZH stage, HSS EI grade and HSS EI stage, respectively (Section 2.5). Our scores showed a significant correlation with the human estimated metrics (Figures 4A–D). We then analyzed the relationship between the two types of biomarkers: the number of eosinophils and the area of the basal zone. It was suggested that these features have some correlation between them (7). A standard condition for the classification of a patient as having active EoE is having a PEC that is greater than or equal to 15. We show that the PBZ distribution of non-active patients has significantly lower values than the PBZ score distribution of the active patients (Figure 4E). A similar trend is observed when analyzing the SBZ distribution (Figure 4F). Yet, there are still patients with high PEC scores and low PBZ / SBZ scores, and vice-versa. This

FIGURE 3
Examples of two different WSIs (left) and their corresponding scores maps with scale for each score defined (middle, right). Each pixel in these maps represents one HPF, and the color of the pixel indicates the respective score. From the Eos-Intact scores map (middle), we extracted peak eosinophil count (PEC) and spatial eosinophil count (SEC). From the basal zone (BZ) score map (right), we computed peak basal zone (PBZ) and spatial basal zone (SBZ) scores. **(A)** Example of a WSI of a biopsy obtained from an EoE patient with inactive disease (PEC = 10). **(B)** Example of a biopsy obtained from a patient with active EoE (PEC = 245).

FIGURE 4
Correlations among the different score types. **(A–D)** Comparing the computed scores with the HSS scores. The HSS scoring method for BZH grade, BZH stage, EI grade, and EI stage, each score is an integer between zero and three. Each panel depicts a violin plot that shows the distribution of the computed WSI scores (vertical axis) for each HSS score that is the appropriate proxy (horizontal axis). The white circle indicates the median value, and the black bar indicates the standard deviation. There is a significant correlation between the computed scores and their HSS counterparts. Histograms of basal zone related metrics PBZ **(E)** and SBZ **(F)** for active ($PEC \geq 15$) and non-active patients ($PEC < 15$). Both the PBZ and SBZ distribution scores of non-active patients have significantly lower values than the PBZ and SBZ distribution scores of the active patients (Kolmogorov–Smirnov-test, $P << 0.0001$).
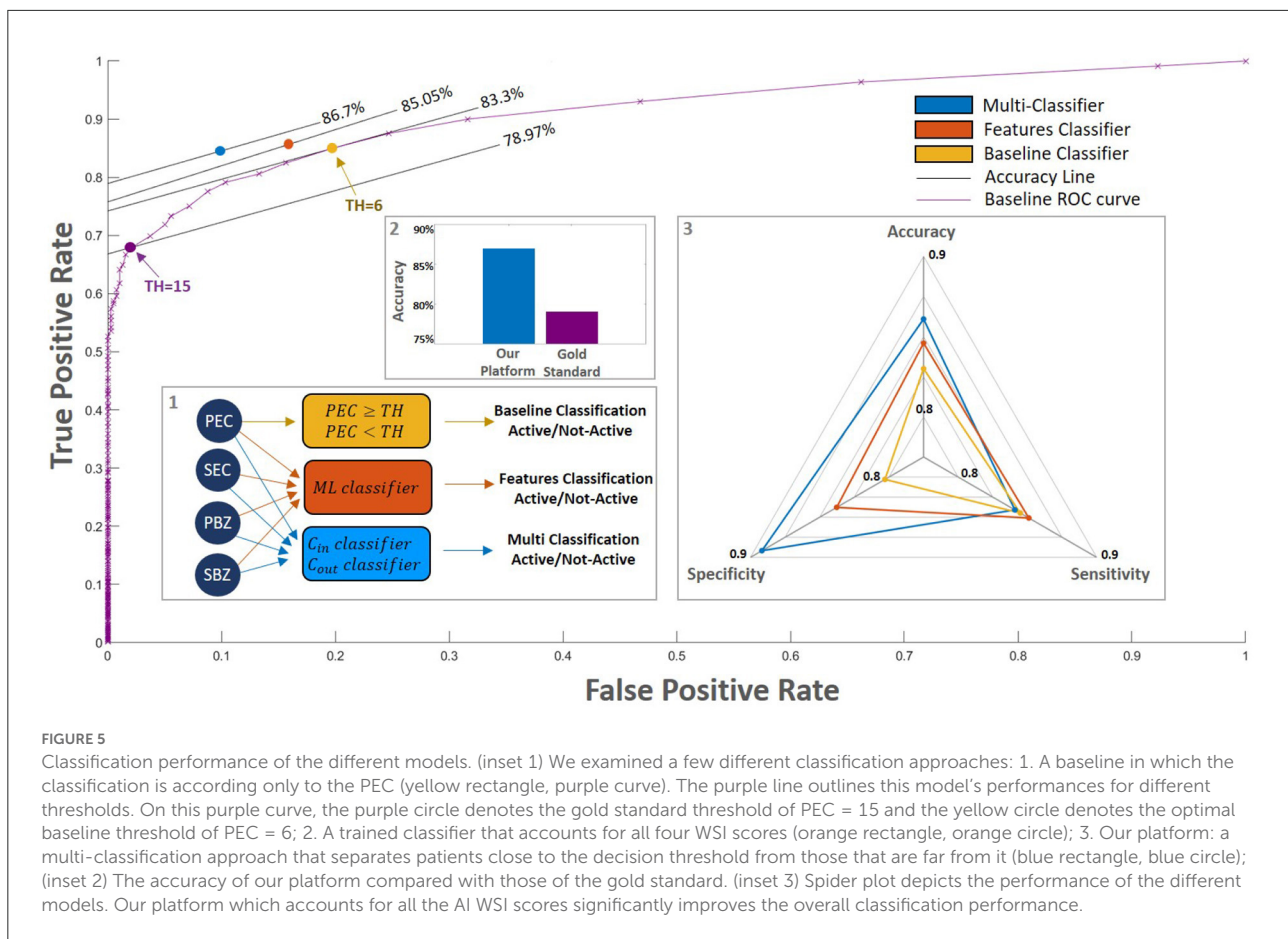
raises the question of whether a combination of basal zone-based metrics can better predict the patient clinical status and treatment outcome.

3.3. Histological severity classification

The naive approach for diagnosing patients' histological severity condition uses only PEC information. In this approach, if the patient's PEC is greater than or equal to 15, the patient is considered to have active EoE. Similar criteria are also applied to determine whether a patient who underwent treatment responded and is in remission. Recent studies suggested using basal zone histological information improves the estimation of the disease's histological severity. For example, it was suggested that patients with low PEC values, i.e., greater than 0 but less than 15, but with basal zone hyperplasia would not be considered as patients in remission (8). To test the performance of our pipeline in integrating all four WSI scores, we used as the ground truth (GT) a standard clinical histological severity metric that

FIGURE 5

Classification performance of the different models. (inset 1) We examined a few different classification approaches: 1. A baseline in which the classification is according only to the PEC (yellow rectangle, purple curve). The purple line outlines this model's performances for different thresholds. On this purple curve, the purple circle denotes the gold standard threshold of PEC = 15 and the yellow circle denotes the optimal baseline threshold of PEC = 6; 2. A trained classifier that accounts for all four WSI scores (orange rectangle, orange circle); 3. Our platform: a multi-classification approach that separates patients close to the decision threshold from those that are far from it (blue rectangle, blue circle); (inset 2) The accuracy of our platform compared with those of the gold standard. (inset 3) Spider plot depicts the performance of the different models. Our platform which accounts for all the AI WSI scores significantly improves the overall classification performance.

defines a histologically severe patient as one who is not in histologic remission, i.e., that has a PEC of greater than or equal to 15 or an HSS total score of more than 3 (8). This metric is stringent when examining whether a patient is in remission or not compared to taking into account only the PEC score.

First, as a baseline classifier, we calculated the accuracy of the histological severity classification when it was based only on the PEC score. The best accuracy (83.3%) was obtained when the threshold criteria was PEC = 6. We recently showed that when taking only PEC as a metric for classification of the patient state (i.e., active EoE vs. non-active EoE), the AI-based PEC score provides a classification accuracy of 94.75%. Moreover, the optimal PEC threshold that provided the best accuracy in that case was 15 (22), the same as the gold standard threshold (6). Thus, the current results suggest that to compensate for the cases in which low PEC are still considered histologically severe, the system converges to more tight PEC criteria for histological severity classification.

Next, we trained a classifier that takes into account all four metrics we calculated from the WSI score maps (i.e., PEC, SEC, PBZ, SBZ). We used several training approaches: support vector machine (SVM), linear discriminant analysis (LDA), and multi-layer perceptron (MLP). The best results were obtained using MLP with three hidden layers where each layer has 20, 50 and 100 neurons, respectively. Integrating all the metrics yields an improvement in accuracy to 85.05%. Moreover, the false alarm rate decreased by about 20% compared to the baseline classifier, whereas the miss rate decreased by about 5% (Figure 5).

A possible factor that may impede the prediction performances is the fact that our data contain patients with a large range of eosinophil counts. To further improve the prediction, we took a multi-classification approach where patients with a PEC level that is near the decision threshold are classified separately from patients that have a PEC level that is far from it. The best results were achieved when patients with PEC values within the range (6, 24) were analyzed separately (Section 2.6.2). This approach led to an accuracy of 86.70% and a significant reduction in the false-alarm rate to 9.91% (Figure 5). In this case, the best results were given by an MLP with three hidden layers in the size of 100, 20, and 100, respectively, for both classifiers.

To gain insight into the role of each of our four WSI scores, we explored the effect of training a classifier with a limited

TABLE 2 Classification results of multiple models (SVM, LDA, and MLP) with different combinations of input features (PEC, SEC, PBZ, and SBZ).

INPUT - WSI AI features scores				OUTPUT - classification models results		
PEC	SEC	PBZ	SBZ	SVM med/std	LDA med/std	MLP med/std
+	+			0.8364 / 0.0247	0.75 / 0.0787	0.8388 / 0.027
		+	+	0.7991 / 0.0236	0.8061 / 0.0227	0.806 / 0.0227
+	+	+	+	0.8341 / 0.0233	0.8155 / 0.0208	0.8505 / 0.0285

Each model was trained and validated 20 times with different train-validation random splits, the median (med) results are reported with the standard deviation (std).
PEC, Peak Eosinophil Count; SEC, Spatial Eosinophil Count; PBZ, Peak Basal Zone; SBZ, Spatial Basal Zone. SVM, Support Vector Machine; LDA, Linear Discriminant Analysis; MLP, Multi-Layer Perceptron.

subset of them (Table 2). In all configurations, the best accuracy was obtained by the MLP model. As expected, the highest classification score was achieved when we used all four WSI AI features scores. Yet, accounting only of Eos-intact scores (PEC and SEC) provides better accuracy than using only BZ scores (PBZ and SBZ).

4. Discussion

Biopsy-based diagnosis often requires the identification of features that are on the single-cell scale. One of the promises of digital pathology, besides automating manual tasks, is the ability to process the entire WSI and infer novel biomarkers that capture the spatial distribution of the relevant features. In the case of EoE, the diagnosis procedure involves counting eosinophils and estimating their density. As a typical whole slide image contains at least tens of high-power fields, gold standard scores usually do not account for the entire features' distribution. In the case of EoE, the gold-standard of clinical diagnosis is based on Peak Eosinophil Count (PEC). As quantifying the number of eosinophils in the slide using manual microscopy, the common practice involves locating by eye the densest high-power fields and taking the maximal number of eosinophils per field as the number that represents the sample. This is a limited biomarker since it considers peak local features (not the entire distribution of eosinophils), and it takes into account only one cellular feature. Indeed, previous histological studies (such as the EoEHSS scoring system) suggested that accounting for more cellular features (such as basal hyperplasia), and taking into account not only the maximal number of eosinophils (or other cellular features) but also accounting for the quantized fraction of high-power fields with threshold levels of eosinophils assessed manually.

In a previous study (22), we showed that our pipeline is able to recapitulate the gold-standard PEC score with state-of-the-art performance. In this work, we go beyond recapturing the current manual histological gold standard. In this study, we introduce an artificial intelligence system that infers novel local and spatial biomarkers based on semantic segmentation of intact eosinophils and basal zone. To test the platform, we utilized a cohort that includes 1,066 biopsy slides from 400 subjects. Whereas the decision of whether EoE is active or not depends on a gold standard cutoff of 15 eosinophils per high power field, the histological severity score (mainly used to estimate whether a patient was in histologic remission after a treatment) also accounts for the basal zone properties. Indeed, using only the PEC of greater than or equal to 15 as a threshold to predict histological severity yields an accuracy of only 78.97% (Figure 5). The PEC cutoff that provides the best accuracy for histological severity, which was 83.3%, is 6 eosinophils/HPF (Figure 5). This reflects the fact that adding the basal zone criteria results in a stronger criteria for the PEC.

Our platform provides a complete quantification of the eosinophils and basal cells fraction over the entire slide. We are therefore able to not only quantify the peak count and basal cell fraction (PEC and PBZ) but also the percent of high-power fields that have more than 15 eosinophils (SEC) and the percent of high-power fields that have more than 25% basal cells within them (SBZ). These metrics have a significant clinical impact – they allow us to predict the histological severity of the patients better than the gold-standard method (86.7% accuracy compared with 78.97% accuracy, Figure 5). Therefore, these new metrics are important for pathologists and gastroenterologists when accounting for the remission status of the patients.

To improve the performance, we used a few machine learning approaches that take our metrics as an input. We show that taking the eosinophil metrics alone yields an accuracy of 83.4% whereas taking the basal zone metrics alone gives an accuracy of 80.6%. Putting all the metrics together gives an accuracy of 85.05%. That is, using all the metrics together gives better performances than each of the metrics alone and also better than a naïve approach of changing the PEC cutoff. Finally, we also constructed a multi-classifier approach that is based on the fact that patients around the $PEC = 15$ cutoffs are more prone to errors. Altogether, our platform yields a classification accuracy of 86.70%, sensitivity of 84.50%, and specificity of 90.09%. Interestingly, while there is no dependence of the error rate with the number of biopsies and their spatial orientation, the disagreement between the AI and the manual decision is

higher when the total of area of the tissue in the slide is bigger. One potential cause for this disagreement could be the difficulty of manually probing a large area. Our approach highlights the importance of systematically analyzing the distribution of biopsy features over the entire slide image and putting together metrics based on them. Our platform paves the way toward a personalized decision support system that will assist in not only counting cells but also in providing treatment prediction.

Data availability statement

The data that support the conclusions of this study will be made available upon request from the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Cincinnati Children's Hospital Medical Center (CCHMC IRB protocol 2015-3613). Written informed consent to participate in this study was provided by the participants' legal guardian.

Author contributions

YS and MER conceived and designed the research. YS and AL designed the pipeline. YS, AL, EA, and ND designed and coded the platform code. AL, EA, ND, TW, and YS analyzed the data and validated it. AL, EA, ND, and YS performed all the mathematical analyses. GO and JC contributed to the pipeline clinical aspects, annotated and validated the segmentation data, and organized and analyzed the data. MC, NA, and G-YY annotated the CEGIR slides. MC supervised the data annotation. MC, MR, NA, and G-YY contributed to the pipeline clinical aspects. YS, AL, EA, ND, TW, and JC wrote the draft of the paper, which was reviewed, modified, and approved by all authors. All authors contributed to the article and approved the submitted version.

Funding

Conflict of interest

Author MER is a consultant for Pulm One, Spoon Guru, ClostraBio, Serpin Pharm, Allakos, Celldex, Nextstone One, Bristol Myers Squibb, Astra Zeneca, Ellodi Pharma, GlaxoSmith Kline, Regeneron/Sanofi, Revolo Biotherapeutics, and Guidepoint and has an equity interest in the first seven listed, and royalties from reslizumab (Teva Pharmaceuticals), PEESSv2 (Mapi Research Trust) and UpToDate. MER is an inventor of patents owned by Cincinnati Children's Hospital. Author MC: Allakos, Arena, AstraZeneca, Bristol Myers Squibb, Calypso, EsoCap, GlaxoSmithKline, Regeneron Pharmaceuticals, Inc., Sanofi, Shire—consultant; Receptos/Celgene/Bristol Myers Squibb, Regeneron Pharmaceuticals, Inc., Shire a Takeda company—research funding.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Dellon ES, Liacouras CA, Molina-Infante J, Furuta GT, Spergel JM, Zevit N, et al. Updated international consensus diagnostic criteria for eosinophilic esophagitis. *Gastroenterology*. (2018) 155:1022–33.e10. doi: 10.1053/j.gastro.2018.07.009

2. O'Shea KM, Aceves SS, Dellon ES, Gupta SK, Spergel JM, Furuta GT, et al. Pathophysiology of eosinophilic esophagitis. *Gastroenterology*. (2018) 154:333–45. doi: 10.1053/j.gastro.2017.06.065

3. Miehlke S. Clinical features of Eosinophilic esophagitis in children and adults. *Best Pract Res Clin Gastroenterol.* (2015) 29:739–48. doi: 10.1016/j.bpg.2015.09.005

4. Dhar A, Haboubi HN, Attwood SE, Auth MK, Dunn JM, Sweis R, et al. British Society of Gastroenterology (BSG) and British Society of Paediatric Gastroenterology, Hepatology and Nutrition (BSPGHAN) joint consensus guidelines on the diagnosis and management of eosinophilic oesophagitis in children and adults. *Gut.* (2022) 71:1459–87. doi: 10.1136/gutjnl-2022-327326

5. Bancroft JD, Gamble M. *Theory and Practice of Histological Techniques. 6th ed.* New York, NY: Churchill Livingstone/Elsevier (2008). p. 744.

6. Dellon ES, Fritchie KJ, Rubinas TC, Woosley JT, Shaheen NJ. Inter- and intraobserver reliability and validation of a new method for determination of eosinophil counts in patients with esophageal eosinophilia. *Digest Dis Sci.* (2010) 55:1940–9. doi: 10.1007/s10620-009-1005-z

7. Collins MH, Martin LJ, Alexander ES, Todd Boyd J, Sheridan R, He H, et al. Newly developed and validated eosinophilic esophagitis histology scoring system and evidence that it outperforms peak eosinophil count for disease diagnosis and monitoring. *Dis Esophagus.* (2017) 30:1–8. doi: 10.1111/dote.12470

8. Collins MH, Martin LJ, Wen T, Abonia JP, Putnam PE, Mukkada VA, et al. Eosinophilic esophagitis histology remission score: significant relations to measures of disease activity and symptoms. *J Pediatr Gastroenterol Nutr.* (2020) 70:598. doi: 10.1097/MPG.0000000000002637

9. Dellon ES. Eosinophilic esophagitis: diagnostic tests and criteria. *Curr Opin Gastroenterol.* (2012) 28:382. doi: 10.1097/MOG.0b013e328352b5ef

10. Stucke EM, Clarridge KE, Collins MH, Henderson CJ, Martin LJ, Rothenberg ME. Value of an additional review for eosinophil quantification in esophageal biopsies. *J Pediatr Gastroenterol Nutr.* (2015) 61:65–8. doi: 10.1097/MPG.0000000000000740

11. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. *Annu Rev Biomed Eng.* (2017) 19:221–48. doi: 10.1146/annurev-bioeng-071516-044442

12. Serag A, Ion-Margineanu A, Qureshi H, McMillan R, Saint Martin MJ, Diamond J, et al. Translational AI and deep learning in diagnostic pathology. *Front Med.* (2019) 6:185. doi: 10.3389/fmed.2019.00185

13. Chauhan NK, Singh K. A review on conventional machine learning vs deep learning. In: *2018 International Conference on Computing, Power and Communication Technologies.* GUCON (2018). p. 347–52. doi: 10.1109/GUCON.2018.8675097

14. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform.* (2016) 7:29. doi: 10.4103/2153-3539.186902

15. Hart SN, Flotte W, Norgan AP, Shah KK, Buchan ZR, Mounajjed T, et al. Classification of melanocytic lesions in selected and whole-slide images via convolutional neural networks. *J Pathol Inform.* (2019) 10:5. doi: 10.4103/jpi.jpi_32_18

16. Azevedo Tosta TA, Neves LA, do Nascimento MZ. Segmentation methods of H&E-stained histological images of lymphoma: a review. *Inform Med Unlocked.* (2017) 9:35–43. doi: 10.1016/j.imu.2017.05.009

17. Cui Y, Zhang G, Liu Z, Xiong Z, Hu J. A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images. *Med Biol Eng Comput.* (2019) 57:2027–43. doi: 10.1007/s11517-019-02008-8

18. Wang J, MacKenzie JD, Ramachandran R, Chen DZ. A deep learning approach for semanticsegmentation in histology tissue images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer (2016). p. 176–84.

19. Djuric U, Zadeh G, Aldape K, Diamandis P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *NPJ Precis Oncol.* (2017) 1:1–5. doi: 10.1038/s41698-017-0022-1

20. Adorno 3rd W, Catalano A, Ehsan L, von Eckstaedt HV, Barnes B, McGowan E, et al. Advancingeosinophilic esophagitis diagnosis and phenotype assessment with deep learning computer vision. Biomedical engineering systems and technologies, international joint conference, BIOSTEC... revised selected papers. *BIOSTEC (Conference).* (2021) 2021:44–55.

21. Czyzewski T, Daniel N, Rochman M, Caldwell JM, Osswald GA, Collins MH, et al. Machine learning approach for biopsy-based identification of eosinophilic esophagitis reveals importance of global features. *IEEE Open J Eng Med Biol.* (2021) 2:218–223. doi: 10.1109/OJEMB.2021.3089552

22. Daniel N, Larey A, Aknin E, Osswald GA, Caldwell JM, Rochman M, et al. A deep multi-label segmentation network for eosinophilic esophagitis whole slide biopsy diagnostics. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society.* EMBC (2022). p. 3211–7. doi: 10.1109/EMBC48229.2022.9871086

23. Gupta S, Falk G, Aceves S, Chehade M, Collins M, Dellon E, et al. Consortium for eosinophilic researchers (CEGIR): advancing the field of eosinophilic GI disorders through collaboration. *Gastroenterology.* (2018) 156:838–42. doi: 10.1053/j.gastro.2018.10.057

24. Dutta A, Zisserman A. The VIA annotation software for images, audio and video. In: *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia.* ACM (2019). p. 2276–9. doi: 10.1145/3343031.3350535

25. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support.* Springer (2018). p. 3–11.

26. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv [Preprint].* (2019). arXiv: 1912.07103. Available online at: https://arxiv.org/pdf/1912.01703.pdf