# Supplementary Materials

## Machine learning approach for biopsy-based identification of eosinophilic esophagitis reveals importance of global features

Tomer Czyzewski[‡], Nati Daniel[‡], Mark Rochman, Julie M. Caldwell, Garrett A. Osswald, Margaret H. Collins, Marc E. Rothenberg, and Yonatan Savir[*]

THIS Supplementary Material provides a detailed evaluation of the impact of image size and downscaling factor on performance, compares our approach to baseline classical classification methods, and explores the effect of additional aggregation methods.

### I. Validation and Training Set Content

The number of samples obtained and used in the training and validation sets for the four different approaches (Table S1).

### II. The Effect of Image Size and Down-Scale Factor

Our dataset is composed out of three types of image sizes with the same resolution: 1024X1360, 1548x2070, and 3096X4140. Table S2 shows the breakdown of the metrics for the different methods. As the downscaling factor (the ratio between the initial area and the final area) is larger, the accuracy is lower. For downscaling factors larger than 10, the accuracy is lower than 80% (Fig. S1A). Moreover, as the image is larger, and thus contains more patches, the accuracy is larger (Fig. S1B).

### III. Aggregation Methods

Besides the majority vote aggregation patches shown in the manuscript (Table I, Figure 3), we have implemented more approaches. First, we examined the effect of different thresholds, that is, the label assigned to the image was positive if the number of patches that were predicted to be positive was larger than a certain threshold. Figure S2A shows the resulting ROC curve for the two cropping approaches. Taking a threshold that is not 0.5 (the majority vote case) does not improve the results. We also used an aggregation based on the mean. The probability for an image to be positive was the mean of the patches probabilities, and the label was positive if the averaged probability was above half. Finally, we applied a hierarchical clustering approach. We built an agglomerative hierarchical cluster tree using MATLAB 'linkage', divided the patches into two clusters accordingly, and assigned the image label as positive if the averaged probability over the biggest cluster was more than half. All approaches yielded the same TNR and similar TPR (Fig. S2B).

### IV. Comparison to Baseline Methods Using Well Known Global Features

To benchmark our results, we implemented classification based on linear discriminant analysis, logistic regression, and linear SVM, based on textural properties of the image. We used a set of 20 well-known textural features [1], [2]:

1. Autocorrelation $=\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} i\cdot j\cdot P(i,j)$.
   Autocorrelation is a measure of the coarseness of an image and evaluates the linear spatial relationships between texture primitives.
2. Cluster Prominence $=\sum_{i=1}^{Ng}\sum_{j=1}^{Ng}(i+j-\mu_x(i)-\mu_y(j))^4\cdot P(i,j)$.
   Cluster Prominence is a measure of the skewness and asymmetry of the GLCM. A higher values implies more asymmetry about the mean while a lower value indicates a peak near the mean value and less variation about the mean.
3. Cluster Shade $=\sum_{i=1}^{Ng}\sum_{j=1}^{Ng}(i+j-\mu_x(i)-\mu_y(j))^3\cdot P(i,j)$.
   Cluster Shade is a measure of the skewness and uniformity of the GLCM. A higher cluster shade implies greater asymmetry about the mean.
4. Contrast $=\sum_{i=1}^{Ng}\sum_{j=1}^{Ng}|i-j|^2\cdot P(i,j)$.
   Contrast is a measure of the local intensity variation, favoring values away from the diagonal. A larger value correlates with a greater disparity in intensity values among neighboring voxels.
5. Correlation $=\frac{\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} i\cdot j\cdot P(i,j)-\mu_x(i)-\mu_y(j)}{\sigma_x(i)\cdot\sigma_y(j)}$
   Correlation is a value between 0 (uncorrelated) and 1 (perfectly correlated) showing the linear dependency of gray level values to their respective voxels in the GLCM.
6. Difference entropy $=\sum_{i=0}^{N_g-1} p_{x-y}(i)\cdot log_2(p_{x-y}(i))$.
   Difference Entropy is a measure of the randomness/variability in neighborhood intensity value differences.
7. Difference variance $=\sum_{i=0}^{N_{g-1}} i^2\cdot p_{x-y}(i)$.
   Difference Variance is a measure of heterogeneity that places higher weights on differing intensity level pairs that deviate more from the mean.

8. Dissimilarity $=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}|\mathrm{i}-\mathrm{j}|\cdot P(i,j)$.
   Dissimilarity measures the relationship between occurrences of pairs with similar intensity values and occurrences of pairs with differing intensity values.
9. Energy $=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}P(i,j)^2$.
   Energy is a measure of homogeneous patterns in the image (Energy is equal to 1 for a constant image). A greater Energy implies that there are more instances of intensity value pairs in the image that neighbor each other at higher frequencies. The property Energy is also known as textural uniformity, uniformity of energy, and angular second moment (ASM).
10. Entropy $=-\sum_{i=0}^{N_g-1}P(i,j)\cdot log_2(P(i,j))$.
    Entropy is a measure of the randomness/variability in neighborhood intensity values.
11. Homogeneity $=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{P(i,j)}{1+(i-j)^2}$.
    Homogeneity is a measure of the local homogeneity of an image. IDM weights are the inverse of the Contrast weights (decreasing exponentially from the diagonal in the GLCM).
12. Informational Measure of Correlation (IMC) $=\frac{-\sum_{i=0}^{Ng-1}P(i,j)\cdot log_2(P(i,j))-HXY}{\max\{HX,HY\}}$.
    Informational measure of Correlation is the correlation between the probability distributions of $i$ and $j$ (quantifying the complexity of the texture), using mutual information
13. Inverse Difference Moment Normalized (IDMN) $=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{P(i,j)}{1+(\frac{|i-j|^2}{N_g^2})}$.
    IDMN (inverse difference moment normalized) is a measure of the local homogeneity of an image. IDMN weights are the inverse of the Contrast weights (decreasing exponentially from the diagonal in the GLCM). Unlike Homogeneity, IDMN normalizes the square of the difference between neighboring intensity values by dividing over the square of the total number of discrete intensity values.
14. Inverse Difference Normalized (IDN) $=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{P(i,j)}{1+(\frac{|i-j|}{N_g})}$.
    IDN (inverse difference normalized) is another measure of the local homogeneity of an image. Unlike Homogeneity, IDN normalizes the difference between the neighboring intensity values by dividing over the total number of discrete intensity values.
15. Maximum Probability $=\max\{P(i,j)\}$.
    Maximum Probability is occurrences of the most predominant pair of neighboring intensity values.
16. Sum average $=\sum_{i=2}^{2N_g}i\cdot p_{x+y}(i)$.
    Sum Average measures the relationship between occurrences of pairs with lower intensity values and occurrences of pairs with higher intensity values.
17. Sum entropy $=-\sum_{i=2}^{2N_g}p_{x+y}(i)\cdot log_2(p_{x+y}(i))$.
    Sum Entropy is a sum of neighborhood intensity value differences.
18. Sum variance $=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+\sum_{i=2}^{2N_g}p_{x+y}(i)\cdot log_2(p_{x+y}(i)))^2\cdot p_{x+y}(i)$.
    Sum Variance is a measure of groupings of voxels with similar gray-level values.
19. Smoothness $=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{P(i,j)}{(1+\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu_x(i))^2\cdot P(i,j))}$.
    Smoothness is a property measured by the number of derivatives it has that are continuous.
20. Variance $=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu_x(i))^2\cdot P(i,j)$.
    Variance is a measure, which puts relatively high weights on the elements that differ from the average value of $P(i,j)$.

Where,

$P(i,j)$ - $(i,j)$th entry in at the co-occurrence matrix.
$p_x(i)$ - $i$the entry in the marginal-probability matrix obtained by summing the rows of $P(i,j), =\sum_{j=1}^{N_g}P(i,j)$.
$p_y(j)$ - $j$the entry in the marginal-probability matrix obtained by summing the columns of $P(i,j), =\sum_{i=1}^{N_g}P(i,j)$.
$p_{x+y}(k)=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}P(i,j),\quad i+j=k,\quad k=2,3,\dots,2N_g.$
$p_{x-y}(k)=\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}P(i,j),\quad |i-j|=k,\quad k=0,1,\dots,N_g-1.$
$N_g$ – Number of distinct gray levels in the quantized image.
$\mu$ – be the mean of $P(i,j)$.
$\mu_x(i)=\sum_i\sum_j i\cdot P(i,j)$ - be the mean of row $i$.
$\mu_y(j)=\sum_i\sum_j j\cdot P(i,j)$ - be the mean of column $j$.
$\sigma_x(i)=\sum_i\sum_j(i-\mu_x)^2\cdot P(i,j)$ - be the standard deviation of row $i$.
$\sigma_y(j)=\sum_i\sum_j(j-\mu_y)^2\cdot P(i,j)$ - be the standard deviation of column $j$.
$HXY=-\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}P(i,j)\cdot\log(p_x(i)\cdot p_y(j))$.
$HX=-\sum_{i=1}^{N_g}p_x(i)\cdot log_2(p_x(i))$.
$HY=-\sum_{j=1}^{N_g}p_y(j)\cdot log_2(p_y(j))$.

Each image is represented as a vector in this texture vector space. The training was done using MATLAB Classification Learner using LDA, LR, and linear SVM classification models. The training and validation sets were identical to the ones used to train and validate the DCNN described in the manuscript. The training was done using 5-fold cross-validation. Estimating the accuracy distribution was done by bootstrapping each model against the validation set 10000 times. The p-value is the probability of getting an accuracy that is higher than the corresponding DCNN accuracy. The results are summarized in Table S3.

## V. Additional Performance Measures

Additional performance measures to the ones shown in table I in the manuscript (Table S4).

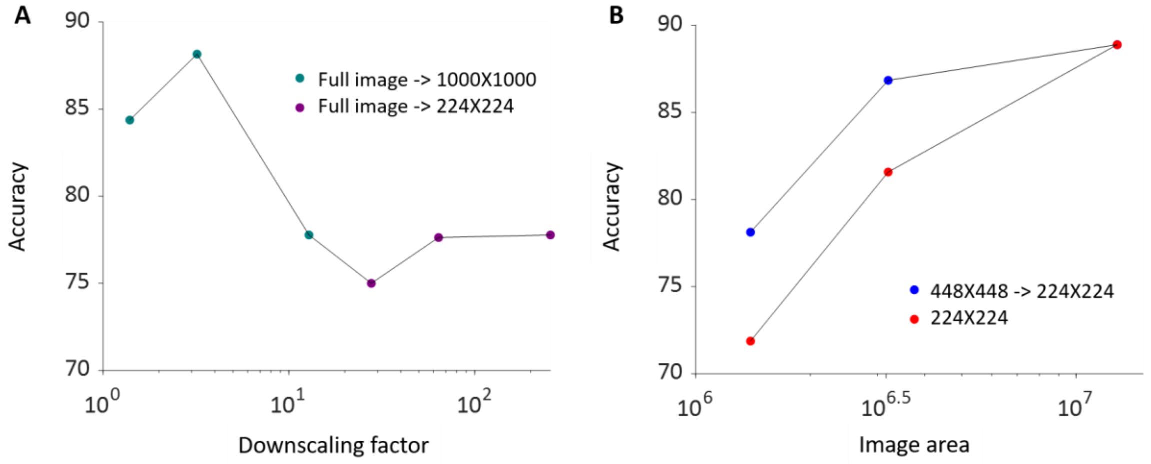

Fig. S1. The effect of downscaling factor (A), and image size (B) on the accuracy for the various methods.

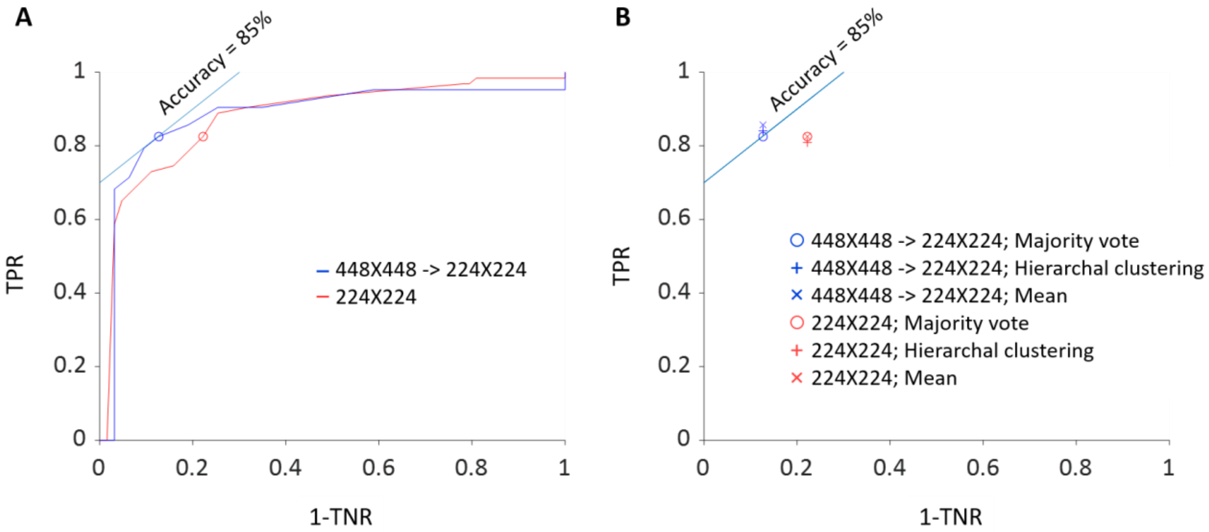

Fig. S2. The effect of different aggregation methods. (A) The results of assigning the image label as positive if the number of positive patches is larger than a threshold fraction. The circles are the majority vote case - the threshold fraction is half. (B) Overlay of the results for a majority vote, hierarchal clustering, and mean approach.

Number Of Training And Validation Generated Image Patches

| Original Image | Final DCNN input image size | Training Set | | Validation Set | |
|---|---|---|---|---|---|
| | | Active EoE | Non-EoE | Active EoE | Non-EoE |
| Full Image | 1000x1000 (Downscale) | 147 | 147 | 63 | 63 |
| Full Image | 224x224 (Downscale) | 147 | 147 | 63 | 63 |
| Patch = 448x448 | 224x224 (Downscale) | 4130 | 4159 | 478 | 497 |
| Patch = 224x224 | 224x224 | 14109 | 14365 | 1626 | 1692 |

TABLE S1. The number of samples obtained and used in the training and validation sets for the four different approaches.

Whole image prediction for the different image sizes

| Image Size | Original Image | Final DCNN input image size | Active EoE (TPR) | Non-EoE (TNR) | ACC | Predicted Prevalence (PP) |
|---|---|---|---|---|---|---|
| 1024X1360 | Full Image | 1000x1000 (Downscale) | 75% | 93.75% | 84.3% | 0.40 |
| 1024X1360 | Full Image | 224x224 (Downscale) | 62.5% | 87.5% | 75.0% | 0.37 |
| 1024X1360 | Patch = 448x448 | 224x224 (Downscale) | 75.0% | 81.2% | 78.1% | 0.46 |
| 1024X1360 | Patch = 224x224 | 224x224 | 87.5% | 56.2% | 71.8% | 0.65 |
| 1548X2070 | Full Image | 1000x1000 (Downscale) | 78.9% | 97.3% | 88.1% | 0.40 |
| 1548X2070 | Full Image | 224x224 (Downscale) | 68.4% | 86.8% | 77.6% | 0.40 |
| 1548X2070 | Patch = 448x448 | 224x224 (Downscale) | 86.8% | 86.8% | 86.8% | 0.5 |
| 1548X2070 | Patch = 224x224 | 224x224 | 81.5% | 81.5% | 81.5% | 0.4 |
| 3096X4140 | Full Image | 1000x1000 (Downscale) | 55.5% | 100% | 77.7% | 0.27 |
| 3096X4140 | Full Image | 224x224 (Downscale) | 55.5% | 100% | 77.7% | 0.27 |
| 3096X4140 | Patch = 448x448 | 224x224 (Downscale) | 77.7% | 100% | 88.8% | 0.38 |
| 3096X4140 | Patch = 224x224 | 224x224 | 77.7% | 100% | 88.8% | 0.38 |

TABLE S2. Whole image classification results for four downscale and/or crop approaches. The validation cohort of images (n = 63 active EoE; n = 63 non-EoE) was the same for each of the classifiers. True positive rate (TPR; number of images classified as active EoE / number of active EoE images x 100), true negative rate (TNR; number of images classified as non-EoE / number of non-EoE images x 100), accuracy (number of images accurately classified as either active EoE or non-EoE / total number of images x 100), and predicted prevalence (total number of images classified as active [i.e., true positive + false positive number of images] / total number of images) for each method are shown. DCNN, deep convolutional neural network. ACC, accuracy.

Whole image Training And Prediction
of Three Different Baseline Classification Learners
(Linear Discriminant Analysis/ Logistic Regression/ Linear SVM)

| Original Image | Image size | Active EoE (TPR) | Non-EoE (TNR) | ACC | Predicted Prevalence (PP) |
|---|---|---|---|---|---|
| Full Image | 1000x1000 (Downscale) | 65.0% / 63.3% / 66.6% | 77.7% / 79.3% / 73.0% | 71.4% (P = 0.0001) /<br>71.4% (P = 0.0001) /<br>69.8% (P < 0.0001) | 0.43 / 0.42 / 0.46 |
| Full Image | 224x224 (Downscale) | 63.4% / 63.4% / 52.3% | 66.6% / 68.2% / 66.6% | 65.0% (P = 0.0009) /<br>65.8% (P = 0.0001) /<br>59.5% (P < 0.0001) | 0.48 / 0.47 / 0.42 |
| Patch = 448x448 | 224x224 (Downscale) | 57.7% / 57.5% / 53.3% | 72.8% / 73.8% / 72.0% | 65.4% (P = 0.0002) /<br>65.8% (P = 0.0004) /<br>62.8% (P < 0.0001) | 0.42 / 0.41 / 0.40 |
| Patch = 224x224 | 224x224 | 52.5% / 51.9% / 48.5% | 66.9% / 67.9% / 73.4% | 59.9% (P < 0.0001) /<br>60.0% (P < 0.0001) /<br>61.2% (P = 0.0002) | 0.42 / 0.41 / 0.37 |

TABLE S3. Whole image classification results for four downscale and/or crop approaches. The training cohort of images training (n = 147 active EoE; n = 147 non-EoE), and the validation cohort of images (n = 63 active EoE; n = 63 non-EoE) was the same for each of the classifiers.
Training is performed on 3 different classification baseline learners with 5-fold cross-validation, which protects against overfitting by portioning the training data set in 5-folds and estimating the model accuracy (number of images accurately classified as either active EoE or non-EoE / total number of images x 100)) on each fold, true positive rate (TPR; number of images classified as active EoE / number of active EoE images x 100), true negative rate (TNR; number of images classified as non-EoE / number of non-EoE images x 100), test accuracy (number of images accurately classified as either active EoE or non-EoE / total number of images x 100), and predicted prevalence (total number of images classified as active [i.e., true positive + false positive number of images] / total number of images) for each method are shown. DCNN, deep convolutional neural network. ACC, accuracy.

Whole image prediction

| Original Image | Final DCNN input image size | Recall | Precision | F1-Score |
|---|---|---|---|---|
| Full Image | 1000x1000 (Downscale) | 74.6% | 95.9% | 0.83 |
| Full Image | 224x224 (Downscale) | 65.1% | 85.4% | 0.73 |
| Patch = 448x448 | 224x224 (Downscale) | 82.5% | 86.6% | 0.84 |
| Patch = 224x224 | 224x224 | 82.5% | 78.7% | 0.80 |

TABLE S4. Whole image classification results for four downscale and/or crop approaches. The validation cohort of images (n = 63 active EoE; n = 63 non-EoE) was the same for each of the classifiers. Recall (number of images classified as active EoE / number of active EoE images x 100), Precision (number of images actually are active EoE / number of images classified as active EoE x 100), F1-score (a weighted average of the precision and recall, formulated as = 2 * Precision * Recall / (Precision + Recall)). DCNN, deep convolutional neural network.

## REFERENCES

[1] R. M. Haralick and L. G. Shapiro, “Computer and Robot Vision: Vol. 1,” in *Addison-Wesley*, Vol. 1., Addison-Wesley, 1992, p. 459.

[2] R. M. Haralick, I. Dinstein, and K. Shanmugam, “Textural Features for Image Classification,” *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, 1973, doi: 10.1109/TSMC.1973.4309314.