Inter-Class Analysis of Frequency-Band Similarity in Gastrointestinal Endoscopic Image Datasets

Zahraa Ch. Oleiwi[1,*], Zena H. Khalil[1], Salwa Shakir Baawi[1], Tara Sabah Mahdi[2]

[1]*Department of Computer Information System, College of computer science and Information Technology, University of Al-Qadisiyah, Diwanya, Iraq*

[2]*College of Medicine, University of Al-Qadisiyah, Diwanya, Iraq.*

[*]*Corresponding Author:* zahraa.chaffat@qu.edu.iq

Received 11 Nov. 2025, Accepted 1 Dec. 2025, published 30 Dec. 2025.
DOI: 10.52113/2/12.02.2025/166-187

Abstract: Gastrointestinal (GI) endoscopic examinations can detect various GI issues early. Challenges like high intra-class variability, moderate differences between classes, and biased data complicate automated classification. The study analyzes frequency-dependent image features in classification results, focusing on similarities within and between classes to better understand the dataset and identify the most difficult classes to classify. It uses a similar approach for similarity analysis with Discrete Wavelet Transform (DWT), breaking images into low-frequency (LL) and high-frequency (HH) sub-bands based on frequency ranges. Structural Similarity Index Measure (SSIM) and Mean Squared Error (MSE) inter- and intra-class similarities. Functional validation involved a classification test using the Random Forest (RF) model. Experiments on multiple GI endoscopic datasets illustrate that LL sub-bands, capturing coarse structural features, provide higher discriminative power and improve classification accuracy, while HH sub-bands, preserving fine textures, are less effective due to higher inter-class similarity. Analysis of similarity measures highlights classes with high intra-class variability, particularly minority classes, as the most challenging for classification. The frequency-aware similarity approach enhances interpretability, reveals dataset-specific issues, and automates the evaluation of gastrointestinal images.

***Keywords*:** Gastrointestinal Endoscopic Image, Intra-Class Similarities, Discrete Wavelet Transform.

1. Introduction

Similarity measure compares the distance (by a selected norm) between the data points in order to estimate the level of similarity between two images. Similarity is the extent of resemblance of two images. A powerful similarity measure is greatly reliant on the choice of the distance or comparison function. In medical imaging particularly endoscopy, analysis of image similarity is essential in order to improve diagnostic and classification of medical imaging [1]. Understanding the image resemblance or variance of patterns among classes of diseases could provide information on the fairness of image qualities and inherent complexity of the classification task. Consequently, scholars have paid more and more attention to the development of similarity measures determining

the visual attributes that have the most significant effect on automated decisions [2].

The research problem is clarified in persistent scanty research has been done on the inter-class similarity of large-scale gastrointestinal (GI) endoscopic image data. Automatic classification is a problem due to the similar visual features of disease groups with slight variation. Conventional measures of similarity are based on structural or perceptual consistency. But These measures tend to overlook frequency based variations which are important in differentiating visually similar classes.

The contributions of this proposed study will solve these issues by proposing a wavelet-based frequency decomposition of GI images into low-frequency (LL) and high-frequency (HH) sub-bands. The inter-class similarity is quantified in both LL and HH band to determine where similarities exist, in general content or structural intricacies, and how well the classification is complex by identifying which frequency bands have the strongest discriminative features. This framework that utilizes frequency-sensitive information classifies classes with inherently greater intra-class variability or subtle inter-class differences, costs less to compute and has less data dimensionality, and improves the performance, efficiency, and interpretability of classifiers in automated GI image classification.

1. Related Works

Wang et al. [3]. created SSIM in 2004, and the approach symbolizes a drastic departure of traditional error-based approaches to perceptually pertinent, structure-based evaluation. The SSIM is used to measure the brightness, contrast and structural coherence of the images and there is a great relationship with human perception. In spite of its massive success in determining image quality, SSIM is heavily space-based and does not adequately represent finer frequency-based variations that prove vital in medical-imaging, wherein high-frequency features often hold vital clinical details.

The FSIM that was presented by Zhang et al. (2011) [4] combined low-level data, such as phase congruence and the magnitude of gradient, to enhance the discriminative analysis. FSIM demonstrated strong results on most benchmark data. But in the case of SSIM, FSIM works mainly in the spatial field. Other approaches are edge detection within similarity assessment, such as the Feature-Based Structural Measure (FSM) by Shnain, Hussain and Lu (2017) [2], which enhances facial picture recognition. Although FSM is effective, it also focuses on conspicuous edges and spatial characteristics without being able to analyze frequency bands individually, and therefore its direct use to inter-class classification in medical imaging, where high-frequency variations are

often critical. Ineffectively resolves frequencies, which may ignore small but significant differences in high frequency textures or complicated structures.

These findings are supported by recent assessments of medical image analysis. Liu et al. (2021) [5] identified the following gaps to deep learning-based medical segmentation, such as the lack of frequency-domain information utilization, inter-class similarity, and intra-class variability. By illustrating that a wavelet-based feature can increase the accuracy of diagnosis of pneumoconiosis images based on their frequency-domain analysis [6], Wang (2022) identified the importance of frequency-domain analysis. Although the authors of the article focused their research on selecting the methodologies that suit certain datasets, Sai Kiran and Areeckal (2025) [7] showed that wavelet-based texture analysis enhances classification in osteoporotic X-ray images.Taken together, these studies indicate a specific trend: feature-based, as well as spatial-domain similarity measures are suitable in perceptual assessment. They may not be capable of the fine, frequency-specific variations required to conduct an accurate medical classification. Following this revelation, the present paper proposes a frequency-band-based similarity analysis through a wavelet decomposition.

2. Proposed Methodology Framework

In this section, the proposed framework that is expected to evaluate the inter-class similarity and complexity of gastrointestinal endoscopic images is described. The main objective is to find out which of the frequency bands (low or high) offers the most distinguishing characteristics to use in classification tasks. The method combines frequency decomposition using wavelets, statistical similarity measurement to select the best frequency bands to classify.

3.1 Dataset Description

Figure (1) displays a sample analysis of a contemporary gastrointestinal endoscopic dataset of multiple disease types performed as a suggested analysis. There are numerous photos of different clinical situations in each of the classes. The available set of gastrointestinal (GI) endoscopic imaging has been used in this work that combines a wide range of diseased and normal conditions around the digestive tract. It is a set of 8,000 high-resolution endoscopic images of the upper and lower gastrointestinal tract including the esophagus, stomach and colon. The assortment includes 27 distinct categories each with a specific gastrointestinal disease or condition. These classes comprise a broad range of anatomical points, clinical, and clinical anomalies,

Fig. (1): Dataset distribution

normal mucosal appearances, and instances of polyp excision, hence provide a comprehensive presentation of the real endoscopic variations [8].

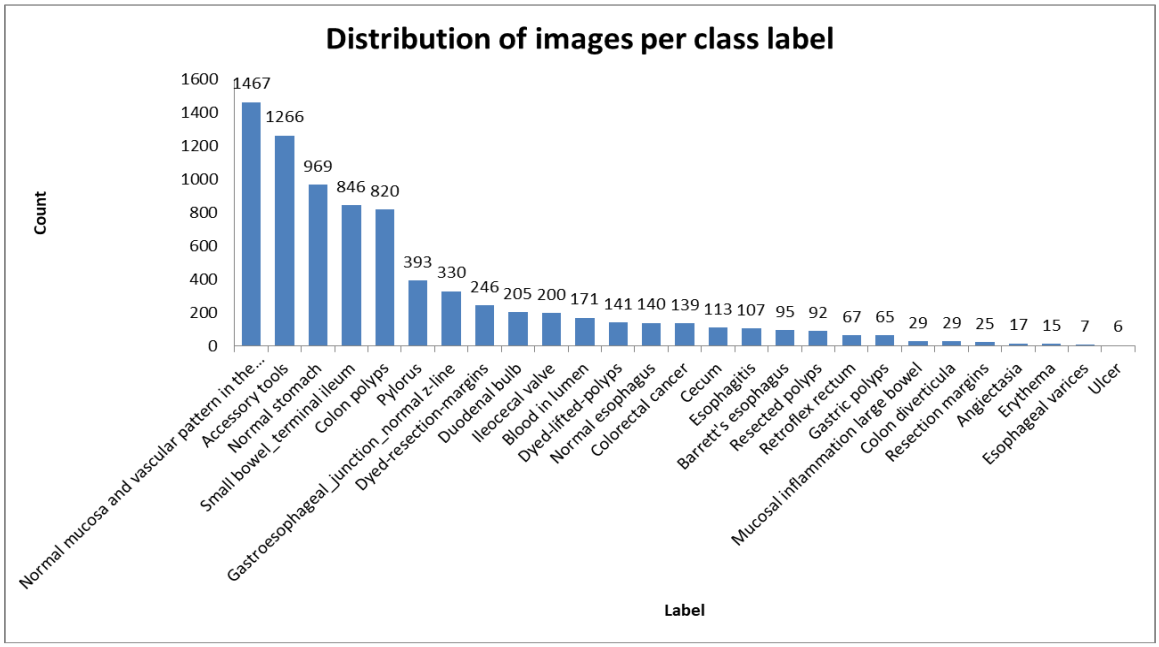

Fig. (2): Dataset distribution over 27 classes.

The images were captured through various endoscopic systems in varying lighting and viewing conditions. So, variability in imaging is usually common in clinical practice. Such variability will ensure that the dataset will include both diseased properties and natural intra-class variation and inter-class similarity within real-world medical images.

The collection contain an equal number of diseased cases and normal cases. This distribution is a good standard of classification algorithms, similarity measures, and frequency-based analytical processes. The diverse visual characteristics of the classes make it particularly suitable when analyzing image similarity, discriminability, and classification difficulty and these are the main goals of the given research [9].

Wavelet Decomposition

Wavelet transform is a method of multi-resolution analysis which represents an image at the same time in both spatial and frequency domain. This representation allowing a detailed consideration of the textures, edges and changes in illumination. Unlike in Fourier Transform, which can only provide global frequencies of an image, the Discrete Wavelet Transform (DWT) divides an image into hierarchical sub-bands and

retains both spatial localization and frequency content[10].

The separable low-pass and high-pass filters used on rows and columns of an image in using the two-dimensional discrete wavelet transform (DWT2) are sequential. The process produces four groups of coefficients at each decomposition stage, one low-pass group, known as the Approximation Coefficients (A or LL) and three high-pass groups, known as the Detail Coefficients (LH, HL, HH), which are the vertical, horizontal and diagonal orientations, respectively. The LL band (Low-Low) maintains the coarse structure of an image and the overall illumination whereas the HH band (High-High) captures fine details such as edges, fine textures, and local contrast variations. Vertical and horizontal details information is included in the LH and HL subbands. The down-sampling procedure of the DWT ensures that after each filtering the number of coefficients is halved in each dimension. This reduces redundancy of data and enhances a hierarchical arrangement where the lower levels contain generalized attributes whereas the high levels preserve specifics [11].

This decomposition can be iterated across numerous layers (L levels) to achieve progressively abstract representations of the visual structure. All images in this study were broken down using the DWT to produce multi-resolution representations. We examined two main subbands:-

-LL (Low–Low): encompasses the low-frequency elements, representing the entire configuration and illumination.

-HH (High–High): encompasses high-frequency components associated with edges, textures, and intricate features.

This decomposition provides insights into frequency-level changes in structural similarity and visual complexity by independently analyzing similarity patterns between the detailed (HH) and coarse (LL) information domains.

3.2 Similarity Measurement

The similarity of the image pairings among different classes was measured on the basis of two complementary and known values: Mean Squared Error (MSE), which is the measure of pixel-level distance, and Structural Similarity Index (SSIM), which is the measure of perceptual distance [2].

The Mean Squared Error for two image sub-band arrays X and Y of size N is defined as [12]: -

$$\mathrm{MSE}(X,Y)=\frac{1}{N}\sum_{i=1}^{N}(X_i - Y^2)^2 \qquad (1)$$

The MSE offers a straightforward and comprehensible metric for the average difference between matching pixel values; lower

MSE values signify greater similarity across images. To find the MSE on the wavelet sub-bands, each sub-band is first min-max scaled to a uniform range (e.g., [0,1]). This ensures that MSE values can be compared across different pictures and sub-band types. The normalization was executed via Eq. (2) [13].

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

The Structural Similarity Index (SSIM) is defined in Eq. (3) as : -

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)} \quad (3)$$

The constants C_1 and C_2 are defined as follows:

$$C_1 = (K_1 L)^2, C_2 = (K_2 L)^2$$

K_1 and K_2 are significant constants, while L equals 255, the maximum pixel value [14].

SSIM can be calculated on a local level with the help of a Gaussian-weighted sliding window, which is averaged across the whole image to obtain a single global similarity index. Similarly, the normalized map of coefficients of the wavelet sub-bands (LL and HH) is processed using a windowed operation, and data-range is 1.0 to produce similar and comparable values of SSIM across various frequency bands.

3.3 Proposed Similarity and Complexity Assessment Technique

To increase the capacity to discriminate and inter-class separability between gastrointestinal tract images, a composite similarity-complexity analysis methodology was created that improved the capacity to discriminate and interclass differentiation. Besides applying the wavelet-based subband decomposition and pair-wiser similarity analysis method. A methodological workflow diagram was demonstrated the broad sequence of suggested methods as in Figure (2).

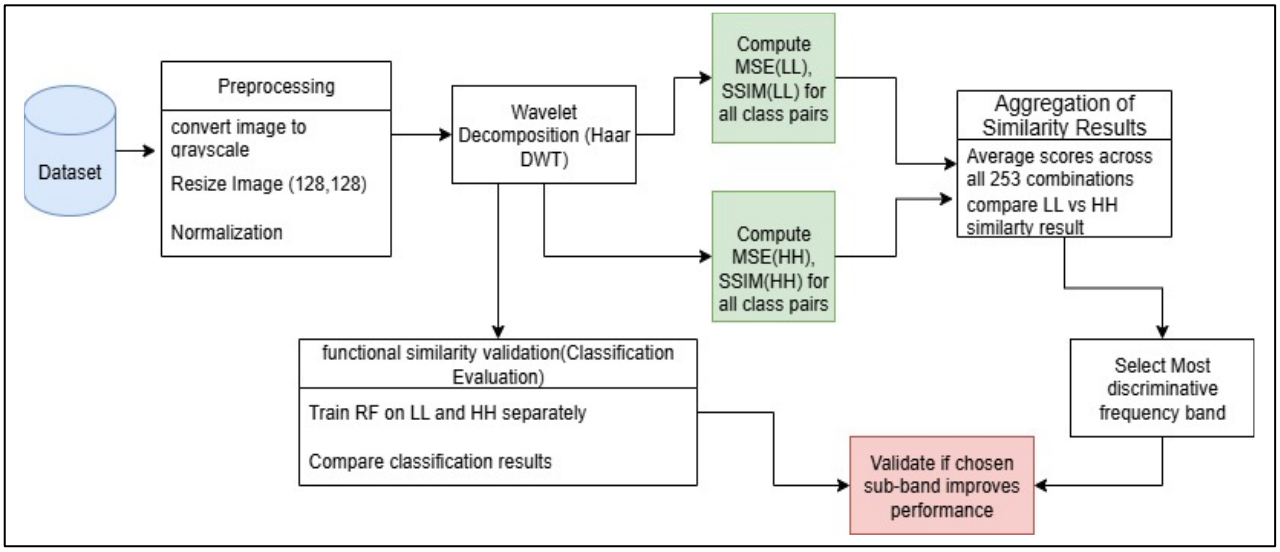

Fig. (3): Proposed Frequency-Based Similarity and Classification Framework Workflow

Step1. Image preprocessing: During this step, a conversion into grayscale followed by resizing all the images to 128X128 pixels was done so that the spatial resolution of each image in the collection remained similar. This is done by a normalizing process so that the samples all have the same dimensionality, making sub-band coefficients comparisons easier. Also, the scaling maintains valuable structural data that is essential in disease classification and reducing computational expenses.

Step2. Pairwise Similarity Computation: Similarity score was calculated in pair-wise variables according to the Discrete Wavelet Transform (DWT) of each image pair of different classes, according to the low-frequency (LL) and the high-frequency (HH) sub-bands. Each of the preprocessed images was decomposed using Haar mother wavelets, i.e. Discrete Wavelet Transforms (DWTs). The Haar wavelet was chosen because it is computationally simple, orthogonal and highly separates low frequency structural information and high frequency detail. Haar gives a localized and sharp basis, so it is especially useful on medical images like gastrointestinal endoscopy, where sudden changes in intensity are significant anatomical boundaries. Also, the use of rectangular step-shaped basis functions in Haar enables it to be easily decomposed at very low cost, a feature necessary when computing large volumes of images, using extensive comparisons of pairwise similarities. A non-redundant representation has the advantage of leaving in LL sub-bands only those global structural components necessary to evaluate inter-class separability, and in HH sub-bands, only fine detail that can be used to measure intra-class variability. Since the main task of the study was to determine the strength of

discriminatory stimuli over frequency bands and the extent of inter/intra-class similarity, Haar was the best fit with the aim of carrying out this exploratory study as it provided an ideal ratio of interpretability, efficiency, and robustness.

Two related similarity indices were applied, Mean Squared Error (MSE), used to measure pixel-wise differences between sub-bands and Structural Similarity Index (SSIM) used to measure perceptual similarity using brightness, contrast and structure. Each of the sub-bands was then normalised using the min-max scaling to the range of [0,1], thus ensuring that the bands and classes are comparable. Solved the zero-variance case by replacing a zero matrix with a zero array to prevent similarity calculation instability.

Step3. Aggregation of Similarity Results: All image permutations in the LL band and HH band were determined and the SSIM and MSE calculated separately per class pair. Aggregated scores are statistically the average similarity of perception and structural difference in classes. The data were summarized and presented in a way that indicates how inter-class similarity is different in low and high frequency domains.

Step4. Frequency Band Selection: Comparative SSIM values between the HH and LL components were used to determine the appropriate sub-band in which the image should be classified as demonstrated by the similarity analysis in frequency bands. A less similar demonstration of the LL band than the HH band indicates that more features of low frequencies (global forms and smooth patches) are differentiated between classes; the LL sub-band is therefore selected as the input to the classifier. On the other hand, in the case where the HH band shows reduced similarity, high-frequency textures and edges have stronger discriminative signals, which causes the choice of HH sub-band. In similar case when the similarity within a band is approximately moderate (e.g. close to 0.5) it often indicates confusing structural overlap between classes and makes discrimination difficult.

In particular, in the case when a substantial resemblance is observed in the LL band, the process of classification becomes more difficult as a strong resemblance in low-frequency structures would signal similar global morphology among the classes leading to reduced separability. On the other hand, high similarity in the HH band is not important, since this is commonly an indication of superficial textures and not fundamental similarities in structure.

Frequency bands with lower similarity are usually preferred, as this would capture more unique and discriminative information, and, thus, better class separation and better classification.

3.4 Evaluation Strategy

To verify the proposed similarity-based complexity analysis, the results were compared with traditional similarity methods (SSIM-only and MSE-only). The assessment was developed to: -

- Determine which bands (low or high frequency) exhibit the greatest discriminative potential.
- Examine how classification difficulty and inter-class similarity relate to one another.

The evaluation and comparison were subsequently structured into two complementary scenarios: -

- Intra-class similarity assessment: The initial scenario assesses the proposed technique by quantifying intra-class similarity, namely the degree of similarity across pictures within the same class, to examine internal class consistency and variability.

This elucidates the internal composition of each category and its impact on categorization complexity.

- Classification performance-based functional similarity validation: The third scenario involves carrying out a controlled classification using multiple classifiers on different representations of data.

The raw images are first classified and then the individual classification of the LL and HH sub-bands are carried out. Results obtained in the form of accuracies and patterns of confusion are analyzed to determine the most common misclassifications and their association with complexity results according to similarity. Specific focus was given to the imbalance in the classes and limited access to the data in certain GastroVision categories. As a result, the frequency bands (LL and HH) were evaluated separately to determine whether one of the bands gives more categorization results. The obtained results were then contrasted with the suggested band-selection method to determine whether the frequency band determined as more discriminative through the similarity analysis truly produced better classification. It is a multi-stage test, a functional contrast, which explicitly connects similarity-based complex analysis with quantitative categorization outcomes, which justifies not only the interpretability but also the practical importance of the suggested method.

3. Results and Discussion

This section breaks down and examines the experimental data derived from the suggested similarity-based complexity evaluation and its validation via classification performance. The analysis intends to: -

- Measure the variation of similarity between frequency sub-bands (LL and HH).
- Analyze the correlation between these similarity patterns and class separability as well as the challenges of classification.

4.1 Intra-class and Inter-class Similarity Patterns

In this sub-section, this method given the average and distributional similarity measures (MSE and SSIM) of the LL and HH sub-bands across all 253 combinations of class pairs which are formed out of the 23 used classes. Though the initial dataset has 27 classes, four of them were not subjected to analysis because of their inadequate sample sizes. Such classes were under-represented which would have resulted in unreliable calculation of frequency-based similarity measures and would have produced unreliable or biased statistics. Thus, it analyzed the rest 23 classes, which give sufficient representation in calculating the similarity of LL and HH among all the 253 combinations of classes. The results showed the differences in internal homogeneity (intra-class) and the mutual separability (inter-class) among classes, and the frequency bands that provide more effective discriminative indicators. The average values of SSIM within the LL sub-band of each class pair combination are presented in Figure (3). The reduction in the values of SSIM means that there is less structural similarity among classes, which points to an increased discriminative capacity in the low-frequency domain.

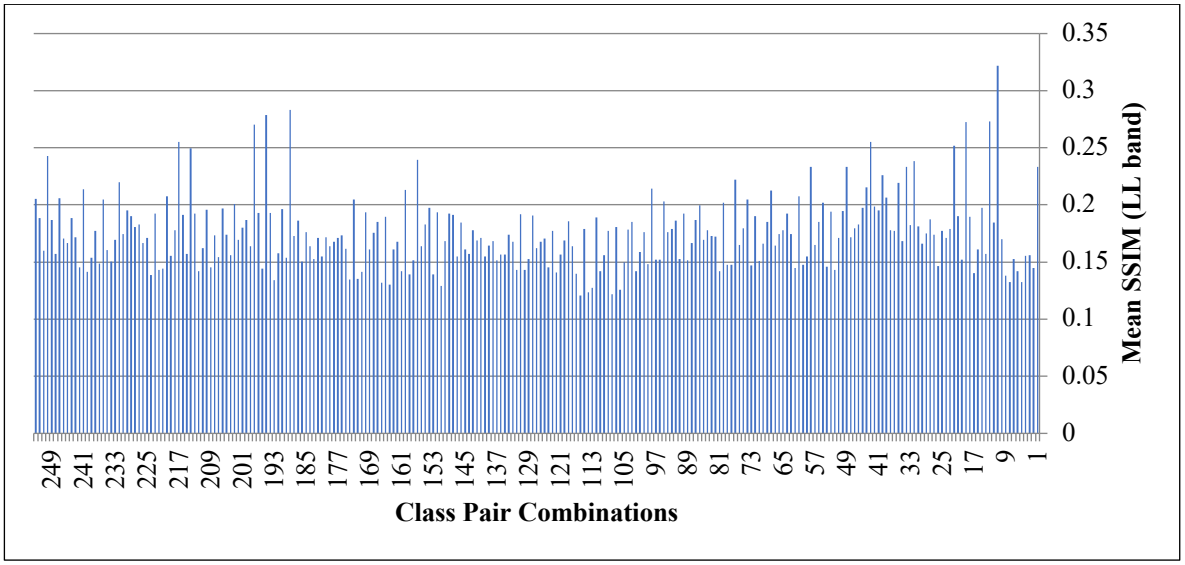

Fig. (4): Mean Structural Similarity (SSIM) over the LL sub-band for all class pair combinations

Figure (4) depicts the average SSIM values calculated for the HH sub-band across all 253 class pair combinations. The results elucidate the high-frequency similarity behavior, emphasizing discrepancies in textural and edge-based discriminative information across various gastrointestinal states.

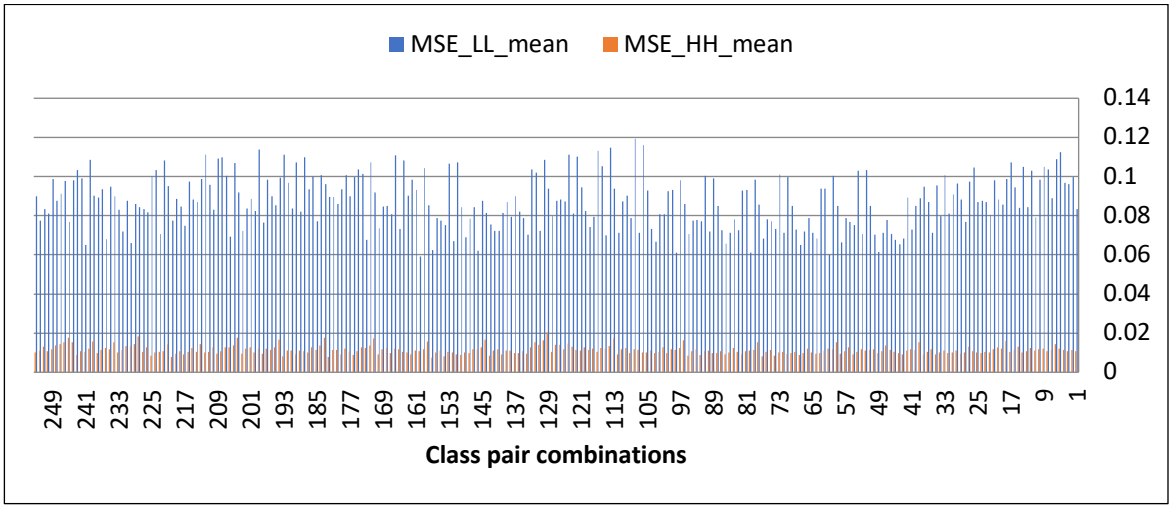

Fig. (5): Mean Structural Similarity (SSIM) over the HH sub-band for all class pair combinations

Figure (5) summarizes the average inter-class Mean Squared Error (MSE) values for the LL (low-frequency) and HH (high-frequency) sub-bands over all 253 class pair combinations. Reduced MSE values signify more pixel-level similarity between class pairings. The HH sub-band typically has elevated MSE values, indicating heightened sensitivity to subtle textural differences, while the LL band encompasses more extensive structural similarities.

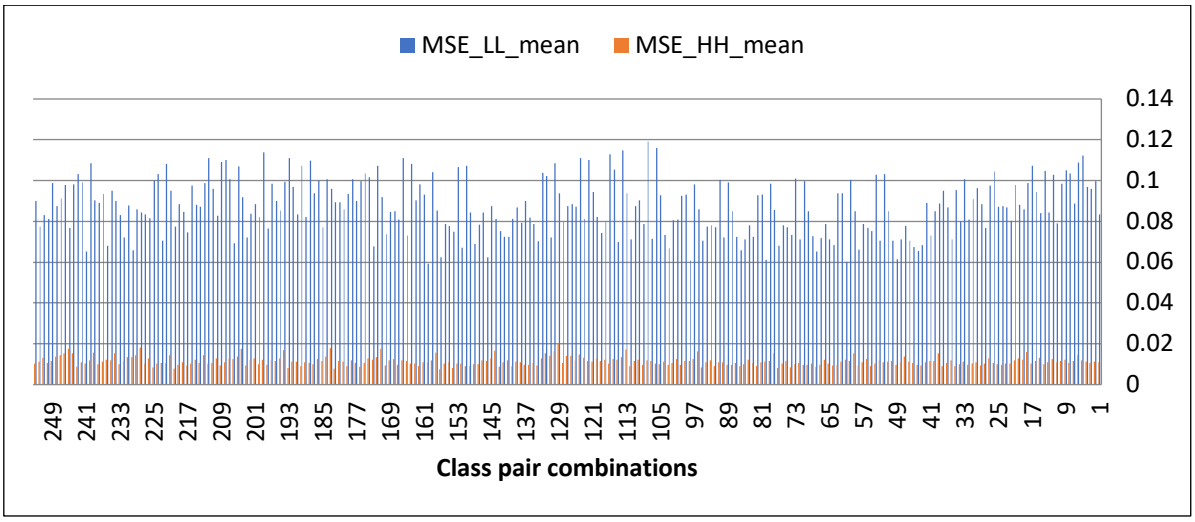

Fig. 6: Comparison of mean inter-class MSE values between LL and HH sub-bands.

The results suggest that similarity values of the low-frequency (LL) bands are worse than that of the high-frequency (HH) bands, whereas the values of the Mean Squared Error (MSE) are very high in the LL bands.

This can be interpreted to mean that HH bands are more similar across classes, and the LL bands have greater inter-class diversity.

The similarity is also present, which suggests classification problems of some pairs of classes, but, overall, the similarity values are lower in the LL bands, which makes them more useful to classification, as they have more efficient discriminative representation to all combinations of classes.

Table (1) demonstrates the statistical qualities of the processed similarity measures (SSIM and MSE) of all 253 combinations of classes-pair involving the two sub-bands (LL and HH).

Table 1: Illustrates the statistical attributes of the calculated similarity measures (SSIM and MSE) across all 253 class-pair combinations for both LL and HH sub-bands

Statistic	Average SSIM (LL Sub-band)	Average SSIM (HH Sub-band)	Average MSE (LL Sub-band)	Average MSE (HH Sub-band)
Minimum value	0.1209	0.2310	0.0591	0.0076
Maximum value	0.3216	0.4736	0.1193	0.0203
Mean value	0.1757	0.3082	0.0872	0.0113

They indicate that the values of the SSIM are always lower in the LL bands (mean= 0.1757) than in the HH bands (mean = 0.3082), and that there are few similarities and high discriminative capacity within the low-frequency domain. In the meantime, the MSE values are greater in the LL bands (mean = 0.0872), which suggests that more classes of variance would be found in the LL subbands. On the other hand, the HH bands are more similar (SSIM) and less reconstruction error (MSE) which is an indication that high-frequency bands retain more shared structures and textures across classes. Such findings confirm the earlier finding that LL sub-bands are more effective in classification since they exhibit less inter-class similarities and more variabilities hence more category separability.

The intra-class similarity statistics of the entire 23 categories in the sample are shown in Table 2. The results show that SSIM values are higher in the HH sub-bands than in the LL sub-bands in most of the intra-class couples indicating that high-frequency components maintain more similarity among the structures in the same class. Conversely, the MSE values are high in the LL sub-bands, which implies that the low-frequency components have a higher degree of variability and less similarity.

Table 2: Similarity metrics (SSIM and MSE) for intra-class couples within LL and HH sub-bands

Pair	SSIM_LL	SSIM_HH	MSE_LL	MSE_ HH
Barretts esophagus vs Barretts esophagus	0.3630	0.3998	0.0848	0.0117
Dyed-lifted-polyps vs Dyed-lifted-polyps	0.2549	0.4008	0.0656	0.0079
Ileocecal valve vs Ileocecal valve	0.2143	0.2885	0.0745	0.0098
Normal mucosa and vascular pattern in the large bowel vs Normal mucosa and vascular pattern in the large bowel	0.2667	0.3325	0.0545	0.0095
Cecum vs Cecum	0.2946	0.3186	0.0424	0.0095
Accessory tools vs Accessory tools	0.2193	0.3162	0.0716	0.0123
Duodenal bulb vs Duodenal bulb	0.2964	0.3584	0.0775	0.0173
Small bowel_terminal ileum vs Small bowel_terminal ileum	0.1922	0.2909	0.0820	0.0146
Resected polyps vs Resected polyps	0.2270	0.3246	0.0688	0.0092
Gastric polyps vs Gastric polyps	0.1831	0.3076	0.0875	0.0173
Esophagitis vs Esophagitis	0.3084	0.4720	0.0652	0.0113
Retroflex rectum vs Retroflex rectum	0.2374	0.3794	0.0749	0.0132
Normal esophagus vs Normal esophagus	0.4112	0.5511	0.0646	0.0057
Normal stomach vs Normal stomach	0.1980	0.3488	0.0926	0.0116
Dyed-resection-margins vs Dyed-resection-margins	0.2720	0.3662	0.0539	0.0084
Mucosal inflammation large bowel vs Mucosal inflammation large bowel	0.2083	0.2978	0.0774	0.0135
Resection margins vs Resection margins	0.2714	0.4103	0.0662	0.0068
Blood in lumen vs Blood in lumen	0.1764	0.2835	0.0905	0.0086
Colorectal cancer vs Colorectal cancer	0.2178	0.2912	0.0643	0.0105
Gastroesophageal_junction_normal z-line vs Gastroesophageal_junction_normal z-line	0.3113	0.4723	0.0713	0.0127
Pylorus vs Pylorus	0.3177	0.3767	0.0552	0.0165
Colon polyps vs Colon polyps	0.2539	0.3428	0.0577	0.0111
Colon diverticula vs Colon diverticula	0.2092	0.3422	0.0616	0.0101

Table (3) illustrates that the intra-class similarity measures reveal SSIM values are predominantly elevated in the HH subband relative to the LL subband, whereas MSE values are more pronounced in the LL subband. This indicates a higher resemblance in the high-frequency components and more discernible distinctions in the low-frequency components, offering insights into intra-class variability, which may also aid in further categorization with other classes.

Table 3: Comparison of Similarity Metrics (SSIM and MSE) for LL and HH Sub-bands within Intra-class Pairs

Statistic	Average SSIM (LL Sub-band)	Average SSIM (HH Sub-band)	Average MSE (LL Sub-band)	Average MSE (HH Sub-band)
Minimum value	0.1764	0.2835	0.0424	0.0057

Maximum value	0.4112	0.5511	0.0926	0.0173
Mean value	0.2567	0.3597	0.0698	0.0113

Intra-class analysis of the LL sub band indicates that the value of SSIM is comparatively low; nevertheless, the value of MSE is considerably high. That is, even components of the same category existing at low frequency have significant differences; this is an indication of differences in general structure, texture, or light. Such variation denotes intra-class variance, which is a crucial variable to reach the classification models or determine the homogeneity of the classes.

In order to examine the reliability of the proposed method of similarity assessment, we point out that the values of intra-class similarity (within the same class) are usually larger than inter-class similarity. This is expected, because all of the photos are of the same class, as it is in keeping with the properties of medical images in such gastrointestinal dataset, where photos of similar anatomical or pathological classes share inherent structural and textural features.

The intra-class heterogeneity is effectively captured by the proposed similarity evaluation method as can be observed by the results in Tables 2 and 3. This difference in the values of SSIM and MSE within a single group can be explained by the fact that there is a biological and anatomic diversity in gastrointestinal medical pictures. This variable is an indicator of the accuracy and reliability of the proposed similarity measurement method, how it is sensitive to subtle differences within the same class, thus making the classification models better at analyzing gastrointestinal data. The current observation aligns with the recent study conducted by Cambay et al. (2024), that highlights the fact that intra-class variability is an issue of significant issues in medical picture classification, particularly in the gastrointestinal domain. The authors in the research attribute this variation to other factors such as difference in lighting, imaging angles, and heterogeneity of the tissues and this necessitates consideration of this variability in the designing of the classification algorithms [15].

4.2 Classification-based Functional Evaluation

The findings of RF of images of LL and HH are provided in this section. Performance measures (accuracy, recall, precision, and F1-score) are contrasted with the similarity based complexity predictions to establish whether classes with higher similarity exhibit lower classification accuracy. The effect of class imbalance and limited sample diversity in GastroVision dataset is also discussed to put the classification performance into context.

The RF model was used in every classification experiment in a controlled setup to ascertain consistency and interpretability of frequency-band testing. In particular, the trees size (n estimators) was to be 10, which is lightweight but expressive enough to detect simple decision patterns without creating too many computational costs. Such a small sample size of trees is consistent with the exploratory nature of this analysis where the most important goal is to compare the discriminative roles of the LL and HH sub-bands as opposed to optimizing the level of model performance. Moreover, a fixed random seed (42) allows getting a complete reproducibility of all classification outcomes. This controlled environment enables the differences observed in the performance of the various frequency sub-bands to be associated with the properties of the data instead of differences in the complexity of the model.

In order to obtain a similar and representative assessment, an 80/20 train and test split was applied to the data. The stratification sampling strategy was used in order to maintain the original proportions of classes in both subsets, which is specifically necessary due to the high imbalance of the dataset. A fixed random seed (42) has been used to split to ensure the reproducibility of the experiments is complete. The stratification was used to maintain the original distribution of classes and to obtain a fair assessment because the imbalance of classes in the data is enormous.

Cross-validation was not used due to the main aim of the study to not receive the most optimized or generalized classification results but to have the decomposition of frequency-bands (LL vs. HH) on the separability of classes and to examine how patterns of similarity connect to the results of classification. The fixed train-test split enabled us to have a controlled and constant evaluation factor that would ensure that the performance differences realized between LL and HH sub-bands could be directly attributed to the intrinsic discriminatory nature of the bands and not due to resampling manipulation.

This sub-section will provide the performance of classification attained with the Random Forest classifier against the two sub-bands, namely, the LL and HH. These findings allow a practical comparison of the functionality of low- and high-frequency representations, showing how each frequency band has contributed to the class separability and overall classification effectiveness.

The randomization of the RF model using the LL (Low-Low) subband summarizes the classification performance of the RF model as shown in table 4. According to the macro and weighted F1-scores, the total accuracy was 43%.

Table 4: Presents the classification report for the LL sub-band, encompassing precision, recall, F1-score, and total accuracy across all classes

Class	Precision	Recall	F1-score	Support
Barretts esophagus	0.06	0.05	0.06	19
Dyed-lifted-polyps	0.00	0.00	0.00	28
Ileocecal valve	0.21	0.10	0.14	40
Normal mucosa and vascular pattern in the large bowel	0.45	0.72	0.55	293
Cecum	0.00	0.00	0.00	23
Accessory tools	0.58	0.60	0.59	253
Duodenal bulb	0.23	0.27	0.25	41
Small bowel_terminal ileum	0.43	0.56	0.49	169
Resected polyps	0.00	0.00	0.00	19
Gastric polyps	0.14	0.08	0.10	13
Esophagitis	0.07	0.05	0.06	21
Retroflex rectum	0.50	0.08	0.13	13
Normal esophagus	0.56	0.32	0.41	28
Normal stomach	0.52	0.59	0.55	194
Dyed-resection-margins	0.21	0.12	0.15	49
Mucosal inflammation large bowel	0.00	0.00	0.00	6
Resection margins	0.00	0.00	0.00	5
Blood in lumen	0.25	0.03	0.05	34
Colorectal cancer	0.00	0.00	0.00	28
Gastroesophageal_junction_normal z-line	0.36	0.33	0.35	66
Pylorus	0.30	0.20	0.24	79
Colon polyps	0.32	0.23	0.27	164
Colon diverticula	0.00	0.00	0.00	6
Accuracy			**0.43**	**1591**
Macro avg	**0.23**	**0.19**	**0.19**	**1591**
Weighted avg	**0.39**	**0.43**	**0.39**	**1591**

Higher recall values were noted for broad structural classifications such as Normal mucosa and Normal stomach, however texture-rich or pathologically analogous categories like Barrett's esophagus and Colorectal cancer had diminished scores. The results demonstrate that the LL sub-band proficiently catches coarse structural patterns and lighting signals, although it fails to include the finer textural characteristics necessary for distinguishing visually comparable disease situations.

The classification results of the Random Forest (RF) model for the HH (High–High) sub-band are presented in Table 5. The model attained an overall accuracy of 32%, with macro and weighted F1-scores of 0.12 and 0.29, respectively.

Table 5: displays the relevant data for the HH sub-band under identical testing conditions

Class Name	Precision	Recall	F1-score	Support
Barrett's esophagus	0.00	0.00	0.00	19
Dyed-lifted-polyps	0.00	0.00	0.00	28
Ileocecal valve	0.00	0.00	0.00	40
Normal mucosa and vascular pattern in the large bowel	0.37	0.53	0.44	293

Cecum	0.00	0.00	0.00	23
Accessory tools	0.41	0.43	0.42	253
Duodenal bulb	0.00	0.00	0.00	41
Small bowel_terminal ileum	0.27	0.31	0.29	169
Resected polyps	0.00	0.00	0.00	18
Gastric polyps	0.00	0.00	0.00	13
Esophagitis	0.08	0.05	0.06	22
Retroflex rectum	0.00	0.00	0.00	13
Normal esophagus	0.25	0.29	0.27	28
Normal stomach	0.39	0.49	0.44	194
Dyed-resection-margins	0.14	0.04	0.06	49
Mucosal inflammation large bowel	0.00	0.00	0.00	6
Resection margins	0.00	0.00	0.00	5
Blood in lumen	0.00	0.00	0.00	34
Colorectal cancer	0.20	0.04	0.06	28
Gastroesophageal junction_normal z-line	0.23	0.21	0.22	66
Pylorus	0.25	0.27	0.26	79
Colon polyps	0.20	0.30	0.24	164
Colon diverticula	0.00	0.00	0.00	6
Accuracy			**0.32**	**1591**
Macro avg	**0.12**	**0.13**	**0.12**	**1591**
Weighted avg	**0.27**	**0.32**	**0.29**	**1591**

While certain structural classifications, including normal mucosa and vascular patterns in the large bowel, normal stomach, and accessory tools, had relatively superior memory, the majority of fine-textured or low-contrast categories produced poor or negligible recollection.

This result indicates that the HH sub-band preserves high-frequency texture details but is more susceptible to noise and fluctuations in illumination, resulting in unstable class separability and less overall discriminative power.

The classification results in Tables (4) and (5) demonstrate that the LL sub-band regularly surpasses the HH sub-band, especially for the predominant classes of Normal mucosa and vascular pattern in the large bowel, Normal stomach, and accessory tools. This supports the use of LL characteristics as the main input for classification and is consistent with the suggested similarity-based analysis, which found decreased inter-class similarity in the LL band for these classes, as seen in Table (6). Consequently, the three classes were able to perform better in terms of classification accuracy, partially due to the larger sample size. As it can be seen in Figure 1, the dataset is highly skewed, which is usually a challenge to classifiers. However, even though one can observe a large intra-class variability in these classes (Table (2)), the large size of the sample relieved the imbalance effect, resulting in the model achieving a relatively stable higher performance.

Table 6: . Inter-class similarity metrics (SSIM and MSE) for the three principal categories: Normal mucosa and vascular pattern in the large bowel, Normal stomach, and Accessory tools compared to all other classes

Reference Class	Compared Class	SSIM_LL	SSIM_HH	MSE_LL	MSE_ HH
Normal mucosa and vascular pattern in the large bowel	Barretts esophagus	0.1555	0.2915	0.0969	0.0113
Normal mucosa and vascular pattern in the large bowel	Pylorus	0.1773	0.2920	0.0871	0.0101
Normal mucosa and vascular pattern in the large bowel	Dyed-resection-margins	0.1972	0.2879	0.0675	0.0106
Normal mucosa and vascular pattern in the large bowel	Resected polyps	0.1926	0.2776	0.0683	0.0095
Normal mucosa and vascular pattern in the large bowel	Colorectal cancer	0.1727	0.2662	0.0725	0.0105
Normal mucosa and vascular pattern in the large bowel	Colon polyps	0.1779	0.2870	0.0782	0.0123
Normal mucosa and vascular pattern in the large bowel	Colon diverticula	0.1692	0.2847	0.0712	0.0100
Normal mucosa and vascular pattern in the large bowel	Dyed-lifted-polyps	0.1995	0.2874	0.0657	0.0092
Normal mucosa and vascular pattern in the large bowel	Ileocecal valve	0.1870	0.2522	0.0726	0.0107
Normal mucosa and vascular pattern in the large bowel	Blood in lumen	0.1667	0.2865	0.0848	0.0097
Normal mucosa and vascular pattern in the large bowel	Normal esophagus	0.1514	0.3245	0.0992	0.0098
Normal mucosa and vascular pattern in the large bowel	Retroflex rectum	0.1927	0.3138	0.0720	0.0110
Normal mucosa and vascular pattern in the large bowel	Gastroesophageal_junction_normal z-line	0.1524	0.3060	0.1003	0.0109
Normal mucosa and vascular pattern in the large bowel	Accessory tools	0.1860	0.2748	0.0772	0.0089
Normal mucosa and vascular pattern in the large bowel	Small bowel terminal ileum	0.1790	0.2453	0.0779	0.0118
Normal mucosa and vascular pattern in the large bowel	Mucosal inflammation large bowel	0.1764	0.2700	0.0774	0.0109
Normal mucosa and vascular pattern in the large bowel	Resection margins	0.2033	0.3120	0.0706	0.0086
Normal mucosa and vascular pattern in the large bowel	Normal stomach	0.1520	0.2700	0.0859	0.0162
Normal mucosa and vascular pattern in the large bowel	Esophagitis	0.1523	0.3212	0.0982	0.0124
Normal mucosa and vascular pattern in the large bowel	Cecum	0.2143	0.2677	0.0610	0.0115
Normal mucosa and vascular pattern in the large bowel	Gastric polyps	0.1481	0.2780	0.0932	0.0116
Normal mucosa and vascular pattern in the large bowel	Duodenal bulb	0.1763	0.2514	0.0926	0.0097

It was found that the use of LL features enhances discriminative power and provides a more reliable input space on which further classification should be performed, and accuracy and F1-scores are improved by approximately 10 percent, in comparison to the HH subband.

Random Forest tests showed poor classification of several classifications (Colon diverticula, Colorectal cancer, Mucosal inflammation of the large bowel, Resected polyps, Cecum, Dyed-lifted polyps, and Resection margins) (Table (4)). The results from Table (2) and Figure (1)

confirm that two compounding challenges define these categories:

- Elevated intra-class variability, indicated by increased SSIM and MSE values in the LL and HH sub-bands, demonstrates considerable variance across images within the same class.
- Insufficient sample size, specifically as they constitute minor classes within the dataset, which intensifies the effects of imbalance on model training and diminishes the classifier's capacity to acquire differentiating characteristics.

All these factors explain the poor performance of these small classes in classification even though feature extraction based on similarity is more effective in other larger classes. This brings out the importance of addressing class imbalance and within-class variability in medical imaging databases and particularly in gastrointestinal endoscopic pictures.

The intra-class variation of the Colon polyps as in Table 2 is clear in this case, although it is also a major class not affected by the sample imbalance that affected the smaller classes. The similarity to other classes is relatively low, which means that inter-class differences are maintained sufficiently. However, there was a low classification accuracy because there was a great disparity within the class alone.

On the other extreme, the six minor categories, such as, Colon diverticula, Colorectal cancer, Mucosal inflammation of the large intestine, Resected polyps, Cecum, and Dyed-lifted polyps, achieved a classification accuracy of zero, primarily because they showed a lot of intra-class variability and had a small sample size. Although this variability difficulty is linked to the colon polyps, the larger sample size facilitated the model to overcome this limitation to some extent resulting in high performance relative to the smaller classes.

4. Conclusion and Future work

The paper presents an elaborate discussion on gastrointestinal (GI) endoscopic image classification with a particular focus on the relationship between frequency-dependent characteristics, intra-class variation, inter-class similarity, and sample size. It can be seen that low-frequency structural cues provide more discriminative power by comparing the proposed frequency-band-based framework that uses Discrete Wavelet Transform (DWT) to differentiate between LL (low-frequency) and HH (high-frequency) sub-bands. At the same time, high-frequency details, still containing textures and edges, have the effect of raising the similarity between classes and are not as effective at distinguishing between classes.Its most significant conclusion is that intra-class heterogeneity can have a significant effect on

classification difficulty. Although there is physical separation between classes (low inter-class SSIM), the internal diversity, especially of minority classes (colon diverticula, colorectal cancer, and dyed-lifted polyps), has a 0% classification accuracy. Conversely, most of the classes, including colon polyps, had a good f1-score (approximately 27%), regardless of the heterogeneity and it is therefore clear that sample size reduces the negative impact of heterogeneity partially. The study also demonstrates that frequency-sensitive selection of features enhances interpretability and efficacy of the classifier. The focus on LL sub-bands is associated with a better classification performance, which suggests that automated gastrointestinal image analysis should be focused on global structured information as opposed to detailed textures, particularly when there is high intra-class heterogeneity. Such results indicate that augmenting the performance of classification will require more than just increasing the size of the sample. Intra-class variability is an important aspect that should be dealt with by means of pre-processing methods, such as normalization of illumination, picture alignment, and focused augmentation, in general, focusing on minority classes. Moreover, frequency-domain analysis offers quantitative method of feature selection that reduces dimensions and helps to improve computing efficiency with no significant loss of discriminative power.To sum up, it is stressed in the study that a delicate balance between inter-class differentiation, intra-class variability, and sample distribution is the key to the successful classification of GI images. These lessons will provide fundamental guidance to the creation of sustainable, interpretable, and effective automated medical imaging.

This research forms the basis of building the custom classification models of the hardest classes that are detected by the intra and inter class similarity analysis. Using the low-frequency features based on the wavelets, future studies are aimed to provide increased accuracy and strength of such complicated classes and enable more effective automated processing of the gastrointestinal images.

An extension of the present work to include cross-validation and optimization of hyperparameters could be applied to a specific classification model once a dedicated classification model is created relying on the results of this exploratory analysis. Although the Haar wavelet was used in the present research because of its simplicity and success in decomposing the low and high-frequency components, research can be done in the future using the more advanced wavelet families, including Daubechies and Symlets. Incorporating these wavelets may possibly provide better representation of features, intra-

class discrimination and increase classification accuracy, especially on difficult classes found in this work.

References

[1] Sivari, E., Bostanci, E., Guzel, M. S., Acici, K., Asuroglu, T., and Ercelebi Ayyildiz, T., 2023, A New Approach for Gastrointestinal Tract Findings Detection and Classification: Deep Learning-Based Hybrid Stacking Ensemble Models, *Diagnostics*, 13, doi: 10.3390/diagnostics13040720.

[2] Aljanabi, M. A., Hussain, Z. M., and Lu, S. F., An Entropy-Histogram Approach for Image Similarity and Face Recognition, *Mathematical Problems in Engineering*, vol. 2018, 2018, doi: 10.1155/2018/9801308.

[3] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., 2004, Image quality assessment: From error visibility to structural similarity, *IEEE Transactions on Image Processing*, 13, 1–14, doi: 10.1109/TIP.2003.819861.

[4] Zhang, L., Zhang, L., Mou, X., and Zhang, D., 2011, FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, 20, 2378–2386, doi: 10.1109/TIP.2011.2109730.

[5] Liu, X., Song, L., Liu, S., and Zhang, Y., 2021, A review of deep-learning-based medical image segmentation methods," *Sustainability (Switzerland)*, 13, 1–29, doi: 10.3390/su13031224.

[6] Wang, Z.,. Hu, M, Zeng, M., and Wang, G., 2022, Intelligent Image Diagnosis of Pneumoconiosis Based on Wavelet Transform-Derived Texture Features," *Computational and Mathematical Methods in Medicine*, 2022, doi: 10.1155/2022/2037019.

[7] Kiran, S. K. S., and Areeckal, A. S., 2025, Classification of Osteoporotic X-ray Images using Wavelet Texture Analysis and Machine Learning, *International Journal of Computing and Digital Systems*, 17, 1–14, doi: 10.12785/ijcds/1570996365.

[8] Al Shafi, A., Ahmed, M., Rahman, M. S., Hossain, M. S., and Uddin, M. F., 2024,Deep Learning for Imbalanced Gastrointestinal Image Classification: A Comparative Study of Architectural Choices, 741–746, doi: 10.1145/3723178.3723276.

[9] Jha, D.*et al.*, 2024, GastroVision: A Multi-class Endoscopy Image Dataset for Computer Aided Gastrointestinal Disease Detection, *Lecture Notes in*

Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14315, 125–140, doi: 10.1007/978-3-031-47679-2_10.

[10] Latif, I. H., Abdulredha, S. H., and Hassan, S. K. A., 2024, Discrete Wavelet Transform-Based Image Processing: A Review, *Al-Nahrain Journal of Science*, 27, 109–125, doi: 10.22401/ANJS.27.3.13.

[11] Kingsbury, N., and Magarey, J., Wavelet Transforms in Image Processing, 27–46, 1998, doi: 10.1007/978-1-4612-1768-8_2.

[12] Güven, S. A., Şahin, E., and Talu, M. F., 2024, Image-to-Image Translation with CNN Based Perceptual Similarity Metrics, *Journal of Computer Science*, 9, 84–98,

[13] M. Arabboev, S. Begmatov, M. Rikhsivoev, K. Nosirov, and S. Saydiakbarov, 2024, A comprehensive review of image super-resolution metrics: classical and AI-based approaches, *Acta IMEKO*, 13, 1–8, doi: 10.21014/ACTAIMEKO.V13I1.1679.

[14] Raigonda, M. R., and Shweta, 2024, Signature Verification System Using SSIM In Image Processing," *Journal of Scientific Research and Technology*, 2, 5–11, doi: 10.61808/jsrt79.

[15] Cambay, V. Y., *et al.*, 2024, Automated Detection of Gastrointestinal Diseases Using Resnet50*-Based Explainable Deep Feature Engineering Model with Endoscopy Images, *Sensors*, 24, 23. doi: 10.3390/s24237710.