Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 27 June 2026
Revised 27 July 2026
Accepted 10 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (18s): 1019–1033

KEYWORDS:
Gastric cancer; image quality assessment; computer-aided diagnosis; GLCM; Local Binary Pattern; CT; endoscopy; multi-modal imaging.

An Interpretable Handcrafted-Feature Framework for Multi-Modal Image Quality Assessment of Gastric CT and Endoscopic Images

Lalitha K[1], S Padmashree[2]

[1]Assistant Professor, Department of Electronics and Communication Engineering, GSSS Institute of Engineering & Technology for Women, Mysuru, Visvesvaraya Technological University, Belagavi, India
[2]Professor, Department of Electronics and Communication Engineering, GSSS Institute of Engineering & Technology for Women, Mysuru, Visvesvaraya Technological University, Belagavi, India

ABSTRACT: The proposed IQA method provides a comprehensive, interpretable image quality assessment (IQA) framework that can be used for a comparative analysis of CT and endoscopic gastric images. This IQA includes handcrafted features such as LBP/GLCM textures and clarity (Laplacian/Sobel/Tenengrad/Brenner), resolution (FFT-based), and morphological top-hat/black-hat abnormality scores. These features are combined into a multi-stage confidence score.

When applied to a dataset of 200 images (100 CT, 100 endoscopic) taken from publicly available sources, the Mann-Whitney U test with Holm-Bonferroni correction revealed that CT images have significantly higher clarity scores (0.788 ± 0.108 vs. 0.663 ± 0.169; $p < .001$, $d = 0.88$). Conversely, endoscopic images show significantly higher quality of textures (0.241 ± 0.027 vs. 0.159 ± 0.048; $p < .001$, $d = -2.10$) and significantly higher abnormality scores (0.576 ± 0.066 vs. 0.376 ± 0.023; $p < .001$, $d = -4.06$). The categorical distribution of the quality across different modalities was found to differ significantly ($\chi^2 = 15.41$, $p < .001$), while the differences between the continuous overall confidence score and the resolution were not significant ($p = .093$ and $p = .066$, respectively). All null hypotheses are presented here together with corresponding post-hoc power analyses. The overall clarity is significantly and positively correlated with the overall confidence score ($r = 0.937$, $p < .001$).

Robustness analysis shows that the proposed method maintains its ranking stability in case of $\pm 50\%$ perturbation of input weights (Spearman $\rho > 0.92$). On the other hand, a moderate multicollinearity is detected for clarity and resolution inputs (VIF > 6). The method is explicitly proposed as a tool for characterizing image modalities and not as a replacement for any supervised methods.

1. INTRODUCTION

Despite advances in technology and understanding of pathology, gastric cancer continues to be one of the leading types of malignant diseases and is still considered one of the most significant causes of cancer deaths throughout the world, with patients' prognosis depending greatly on the stage of gastric cancer. According to Liu and Deng [1], if the disease is diagnosed at the stage of mucosa, then the rate of five-year survival rises from less than 30% to over 90%. The clinical diagnosis involves the use of computed tomography (CT) and endoscopy; the two methods complement each other in the sense that while CT gives structural and anatomical data about the extension of tumor and lymphadenopathy [2], endoscopy enables to see directly the

mucosa of the stomach, being more sensitive for detection of surface lesions like mucosal discoloration, irregular margin, and textural abnormalities [3]. Artificial Intelligence (AI) with a focus on deep learning has shown great effectiveness in lesion detection. According to Klang et al., CNN endoscopic systems have sensitivity levels equal to expert endoscopists [4]. Lee et al. [5] verified the effectiveness of a CAD system using artificial intelligence, where it achieved sensitivity levels above 88% in a dataset from multiple centres undergoing endoscopy. Chen et al. [6] used transformers for CT images to predict postoperative recurrence. However, these AI models require large annotated datasets along with significant computational power and often work as opaque black-box systems [7]. The need for an explainable and computation-efficient method for comparing images in terms of quality is still relevant. In this paper, we tackle this problem by presenting a handcrafted feature-based IQA model tested on real image data and subjected to rigorous statistical tests, including controlled multiple comparison with Holm correction and robustness evaluation. The model is presented specifically as a tool for modality characterization and feature selection, which does not replace but guides the design of supervised diagnostic models.

2. RELATED WORK

2.1 Endoscopic Image-Based Studies

Hirasawa et al. [8] were one of the first to apply a CNN-based system (Single Shot MultiBox Detector) to endoscopic image classification of gastric cancer and achieved the sensitivity of 92.2% in 2,296 endoscopic images, demonstrating the capability of AI to reach expert levels of accuracy. The main achievement was proving the potential applicability of real-time AI-assisted endoscopy. However, the system had been developed based on one institutional data and, therefore, there were doubts about its generalizability, as different endoscope brands, lighting conditions, and patient demographics might affect the system's performance. Hasan et al. [9] provided a review of AI-assisted detection of gastric cancer using endoscopy and reported that the current deep learning models could achieve sensitivities above 90% in controlled validation settings. Their systematic review comprehensively described the state-of-the-art technologies and challenges faced by the field, pointing out that numerous systems suffer from class imbalance, insufficient data diversity, and lack of prospective validation in clinical settings. Klang et al. did a systematic review of 35 papers on AI-assisted endoscopy and concluded that ensemble deep learning models were better than the single models. The main finding was that AI systems were able to exceed the performance of experienced endoscopists when it came to detecting lesions. One of the limitations mentioned in the papers was significant heterogeneity of dataset sizes, annotations, and evaluation procedures. Lee et al. validated a commercially deployed AI-assisted system on 3,324 endoscopic images from multiple centres, achieving 88.6% sensitivity and 89.4% specificity. The contribution was providing real-world multi-centre validation, which is rarely reported. The limitation was that the system still required significant human oversight and performed poorly on early-stage flat lesions.

2.2 CT-Based Imaging Studies

Zhu et al. [10] proposed a CT radiomics approach using machine learning to predict occult peritoneal metastasis in advanced gastric cancer preoperatively, achieving an AUC of 0.82 on validation data. Their contribution was demonstrating that CT texture features extracted by radiomics pipelines carry prognostically relevant information beyond what radiologists visually assess. The limitation is that radiomics features are sensitive to CT acquisition parameters (slice thickness, contrast agent timing), limiting reproducibility across institutions. Chen et al. developed a transformer-based prognostic model integrating CT tumour and body-composition images, outperforming CNN baselines in recurrence prediction. The contribution was showing that self-attention mechanisms capture long-range spatial context in CT volumes more effectively than convolutional kernels. The limitation is that transformer models require very large training datasets and are computationally expensive, making them impractical in resource-constrained settings.

2.3 Multi-Modal and Interpretable Approaches

Zhang et al. [11] developed an AI-based multimodal model combining CT and endoscopic images for pathological staging, reporting AUC values above 0.85 for T-staging. The contribution was demonstrating that fusion of structural (CT) and surface-level (endoscopy) information improves staging accuracy over single-modality models. The limitation was the requirement for paired CT and endoscopic acquisitions for every patient, which is not always clinically feasible, and the model was a black-box deep network without interpretable outputs. Wu et al. [12] reviewed multimodal AI applications across gastroenterology and hepatology, highlighting that cross-modal attention mechanisms outperform late-fusion approaches. Their contribution was providing a structured taxonomy of multimodal fusion strategies. The limitation was that most reviewed systems were not

validated on independent prospective cohorts. Li et al. [13] developed and validated a feature-extraction-based diagnostic system for early gastric cancer achieving 85.4% sensitivity on a case-control dataset, demonstrating that interpretable handcrafted features remain competitive with deep learning for early lesion detection on limited datasets. The limitation was a relatively small validation sample and single-centre design. Mastouri et al. [14] and Nadeem et al. [15] demonstrated that structured handcrafted-feature CAD pipelines remain effective for detection of lung and pancreatic malignancies respectively, providing evidence that interpretable methods generalize across organ systems and imaging modalities. Zubair et al. [16] proposed an interpretable multi-channel attention framework for gastric cancer classification on histopathology images, highlighting that attention-based transparency mechanisms can improve clinician trust without sacrificing classification accuracy. The common limitation across these interpretable approaches is that they have not been comprehensively validated through rigorous per-image statistical testing with multiple-comparison correction, which the present study addresses

3. PROBLEM STATEMENT

CT and upper GI endoscopy provide different but complementary information for the evaluation of the stomach pathology. While CT gives an opportunity to visualize the stomach wall and the surrounding anatomical structures, the endoscopy is used for direct examination of the mucosa. However, CT and endoscopy suffer from variations in image sharpness, texture, spatial resolution, and quality that can influence feature extraction and the accuracy of computer-aided diagnosis systems. The existing AI-based computer-aided diagnosis systems focus on disease classification based on deep neural networks. Although such systems are highly effective, they have high computational costs and need big annotated datasets. Furthermore, comparatively little attention has been given to the objective assessment and comparison of image quality characteristics across gastric CT and endoscopic images. To address this gap, this study proposes an interpretable handcrafted-feature framework for comparative image quality assessment of gastric CT and endoscopic images. The framework integrates quantitative measures of clarity, texture, resolution, abnormality, and confidence, together with comprehensive statistical validation, to objectively characterize modality-specific image quality. The proposed approach supports image quality evaluation, dataset curation, and reliable preprocessing for downstream AI-based diagnostic systems.

4. PROPOSED METHODOLOGY

Figure 1 illustrates the layered architecture of the proposed framework. Multi-modal images are preprocessed, passed through modality-specific feature extraction, followed by morphological abnormality detection, label-guided contextual scoring, and a multi-stage confidence/risk aggregation module, with outputs including per-image quality categories, risk scores, and batch-level comparative reports.

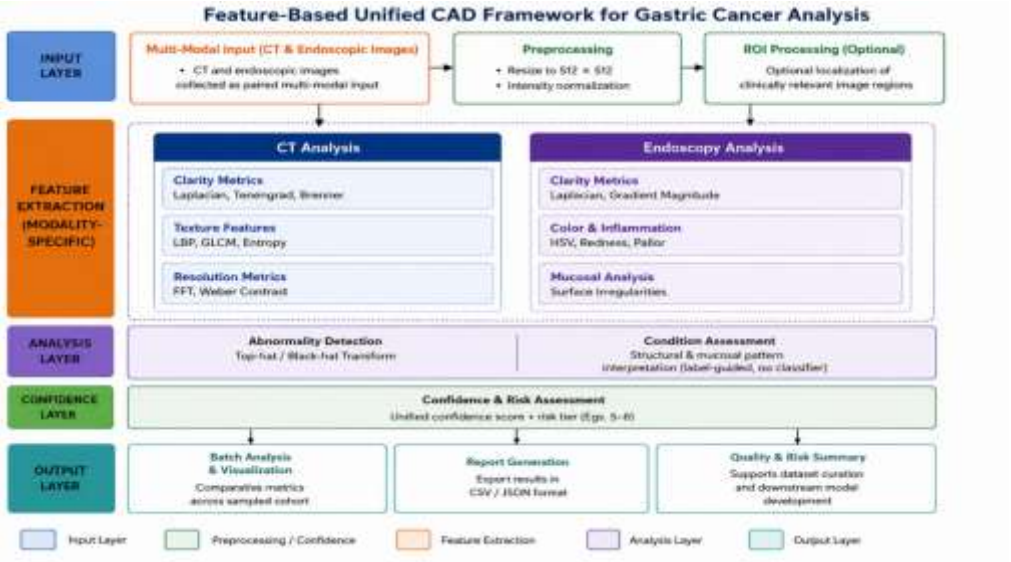


Figure 1. Layered architecture of the proposed unified CAD framework: from multi-modal input and preprocessing through modality-specific feature extraction (CT: clarity, texture, resolution; endoscopy: clarity, colour, mucosal analysis), abnormality detection, confidence/risk aggregation, and output reporting.

4.1 Data Preprocessing and Dataset Description

The study uses publicly available CT (DICOM format, sourced from the TCGA-STAD collection) and endoscopic images spanning clinically relevant condition labels including normal mucosa, Barrett's esophagus, erythema, gastric polyps, and ulcer. Medical images from different sources vary considerably in acquisition protocol, contrast settings, and resolution; a standardized preprocessing pipeline is therefore applied before feature extraction to ensure that all comparisons between CT and endoscopy are made on a consistent, normalized basis [17]. CT scans first get Hounsfield Unit (HU) windowed. Hounsfield Units is a standard density scale for CT imaging where 0 HU equals to water, -1000 HU corresponds to air, and more than 400 HU corresponds to dense bone. In order to improve visualization of different types of tissues, different windows are used. For example, the abdominal window which is characterized by center 50 HU and width 350 HU transforms intensity values into interval of [50 - 175, 50 + 175] HU = [-125, +225 HU] and corresponds to a standard clinical setting for visualizing soft tissues, such as stomach wall, adjacent fat, etc. All the pixels which lie out of the specified interval become black or white. This procedure is crucial since, without such transformation, the stomach wall would be invisible on the background of a whole 4000 HU dynamic range of CT scan. After windowing, automatic phase classification is performed through the examination of the Series Description DICOM tag, categorizing each image into one of the four phases: arterial phase (contrasts peak in arteries; best for tumor visualization), portal venous phase (best for organs), delayed phase, and non-contrast phase. For the pilot study, the arterial phase CT scans are used (n = 100) as arterial phase imaging provides the maximum soft tissue contrast for evaluation of the gastric wall. An additional optional step of stomach region localization is then performed using thresholding.

Bilateral filtering is used to reduce noise, while edge information is preserved. The next procedure is CLAHE, Contrast Limited Adaptive Histogram Equalization, that is used to increase the local contrast of small regions in an image without increasing the noise in homogeneous regions – an important operation in order to uncover subtle differences in mucosa texture. Finally, all images are resampled to a common size of 512×512 pixels, which is the standard size in medical image processing. The endoscopy images are sorted per folder condition label, and intensity normalization can be performed with region of interest extraction (optional). Pilot Dataset: Arterial phase CT (n=100), Barrett's esophagus endoscopy (n=95), erythema endoscopy (n=5).

4.2 Data Preprocessing and Dataset Description

The framework extracts five interpretable feature categories per image. Each category is computed using well-established mathematical formulations and captures a distinct diagnostic property — together they provide a complementary multi-dimensional characterisation of image quality and tissue characteristics [13], [15].

Clarity (Eq. 1).

Clarity measures how sharp and well-focused the image is. A blurry image loses fine structural detail and reduces the reliability of any downstream model. Five focus-measure operators are combined, each capturing a different aspect of sharpness: Laplacian variance (Lc) detects rapid intensity changes at edges by computing the second derivative of the image — a high variance means many sharp transitions are present; Sobel gradient magnitude (Gc) measures first-order intensity gradients and is sensitive to edge strength; Canny edge density (Ec) counts the proportion of detected edges in the image after non-maximum suppression; Tenengrad (Tc) computes the sum of squared gradients and is particularly sensitive to blur; and Brenner's measure (Bc) compares pixels two steps apart, capturing medium-frequency sharpness. Each is independently normalized to [0,1] before combining:

$$\text{Clarity} = 0.25Lc + 0.25Gc + 0.20Ec + 0.15Tc + 0.15Bc \quad (1)$$

The weights (0.25, 0.25, 0.20, 0.15, 0.15) reflect the relative reliability of each operator for medical images; Laplacian and Sobel are given highest weight as they are the most established measures of image sharpness in the medical imaging literature. A Clarity score approaching 1.0 indicates a sharp, well-focused image; a score below 0.5 suggests significant blur or poor focus.

Texture — LBP and GLCM (Eqs. 2–3).

Texture features capture the surface pattern and local intensity structure of the image, which is critical for detecting mucosal irregularities. Local Binary Pattern (LBP) encodes local texture by comparing each pixel's intensity against its circular neighborhood of P=24 points at radius=3 pixels:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \qquad (2)$$

Where P denotes the number of neighbouring pixels, R represents the radius of the circular neighbourhood, gc is the gray-level intensity of the centre pixel, gp is the gray-level intensity of the p^{th} neighbouring pixel, s(·)is the binary threshold function: s(x) = 1 if x ≥ 0 (i.e., the neighbor is brighter than the center), and s(x) = 0 otherwise (neighbor is darker). Multiplying by 2^p assigns a unique binary code to each neighborhood configuration. The resulting LBP code for each pixel is a binary number representing which neighbors are brighter. The distribution (histogram) of these codes across the whole image describes the image's overall texture pattern; the histogram's energy (sum of squared bin frequencies) is used as a texture-uniformity indicator — high energy means a few dominant patterns repeat consistently (uniform tissue), low energy means many scattered patterns (irregular or varied texture).

The GLCM (Gray Level Co-occurrence Matrix) describes the co-occurrence of pixel intensity pairs in space. The GLCM computation is carried out using distances {1, 2} pixels and angles {0°, 45°, 90°, 135°} in order to ensure rotation invariance. The GLCM contrast measure can be calculated as follows:

$$\text{Contrast} = \Sigma_{i,j} (i - j)^2 \cdot P(i,j) \qquad (3)$$

where i, j — pixel intensity values (range 0–255 for 8-bit images); P(i, j) is the normalized probability that a pixel with intensity i is neighbor to the pixel with intensity j with respect to the angle and distance; $(i - j)^2$ is the weight of the co-occurrence measure depending on the difference of two intensity values. Large contrast measure means that neighboring pixels often have large intensity difference, which occurs in images with sharp transitions (for example, irregular surfaces of mucosa). Along with contrast, the GLCM also produces homogeneity (which describes how close the distribution is to the diagonal – large homogeneity if neighboring pixels are similar) and energy (sum of squared elements of GLCM – large energy for uniform textures).

Resolution and Contrast (Eq. 4).

Resolution determines how much detail is kept in the picture. It can be measured in two different ways: FFT that checks the relation between the frequencies of spectral energy – high to low; contrast (Weber, Michelson, and RMS contrast), which checks the difference between light and dark areas. RMS contrast, which is a square root from the average of the square differences between the gray levels of each pixel in the picture and the mean intensity for all pixels, is written as:

$$\sigma^2 = (1/N) \cdot \Sigma(I - \mu)^2 \qquad (4)$$

where I is the intensity of each pixel, μ is the mean intensity of all N pixels in the picture, and σ^2 is the variance. The spatial resolution sub-score and the contrast sub-score are combined in a 60%/40% ratio into an overall resolution score, with spatial resolution given higher weight since fine spatial detail is the primary diagnostic concern.

Abnormality Detection and Label-Guided Scoring.

Abnormality scoring identifies regions of the image that exhibit structural irregularities suggesting potential lesions. Two morphological operations are applied to a grayscale version of the image using a 5×5 elliptical structuring element: the top-hat transform (image minus its morphological opening) isolates bright regions that stand out from their surroundings — these correspond to bright lesions such as raised polyps; the black-hat transform (morphological closing minus image) isolates dark regions that stand out — these correspond to dark lesions such as ulcers or erosions. The mean intensity of these transformed images gives the bright-lesion and dark-lesion intensity scores. For endoscopic images, HSV color-space analysis additionally extracts an inflammation indicator (proportion of reddish pixels in the H channel corresponding to 340–20° hue) and a pallor indicator (proportion of low-saturation pixels), both of which are clinically recognised signs of mucosal pathology. Then a label-guided contextual adjustment re-scales these components according to the dataset condition label of the image (e.g., ulcer-labelled images put more weight on the dark-lesion component), without explicitly classifying the disease.

Unified Confidence Aggregation (Eqs. 5–6).

The five feature scores are combined in two steps to produce a single overall confidence score C for each image. In Stage 1, a diagnostic-confidence sub-score C^d is calculated as a modality-specific weighted sum of the five normalised feature scores:

$$C^d = w_1 Fc + w_2 F_t + w_3 F_r + w_4 Fc_{ol} + w_5(1-F_a) \quad (5)$$

where Fc is the clarity score, F_t the texture quality score, F_r the resolution score, Fc_{ol} the colour quality score (CT: set to 0, endoscopy: set to the calculated score), and F_a the abnormality score (subtracted by 1, so that a higher abnormality results in a reduction in diagnostic confidence). The weights w_1–w_5 are modality dependent. For CT, β^{cT} = {0.30, 0.30, 0.25, 0, 0.15}, which gives equal importance to clarity and texture and slightly less to resolution. For endoscopy, β^{endo} = {0.20, 0.20, 0.20, 0.25, 0.15}, which shifts weight to colour, since mucosal colour is a primary diagnostic cue in endoscopy.

Stage 2 involves computing three composite intermediate scores: F_v (feature visibility, which represents a modality-weighted combination of clarity, texture and colour that quantifies how easily image features can be differentiated), Q (image quality, which is a combination of clarity, resolution, texture and colour) and R (risk score, which is the average of the most relevant abnormality sub-indicators). The results are combined in the final confidence score:

$$C = 0.40C^d + 0.30F_v + 0.20Q + 0.10(1-R) \quad (6)$$

where C d receives the highest weight (0.40) since it directly encodes the diagnostic signal, F v is next (0.30) as it captures the visibility and distinguishability of the features, Q adds image-quality context (0.20), and the risk term (0.10) provides a mild penalty for images with high abnormality-pattern content that may correspond to acquisition artefacts rather than true lesions. The resulting score C lies in [0,1]; images are assigned to five quality categories: Excellent ($C \geq 0.80$), Good ($0.65 \leq C < 0.80$), Fair ($0.50 \leq C < 0.65$), Poor ($0.35 \leq C < 0.50$), and Very Poor ($C < 0.35$).

4.3 Experimental Setup and Implementation

The framework was implemented in Python 3.11. Each library was selected for a specific role: OpenCV handles image loading, resizing, colour-space conversion (BGR to HSV, greyscale), and morphological operations (top-hat, black-hat transforms); scikit-image provides the LBP implementation and GLCM computation via its 'feature' module; NumPy performs all array-level mathematical operations including FFT spectral analysis and variance computation; SciPy supplies the statistical testing functions (Mann–Whitney U, Shapiro–Wilk, chi-square, Kruskal–Wallis, Spearman correlation); stats models provides VIF calculation and Holm–Bonferroni correction; Pandas manages the per-image score data frame and CSV export; and Matplotlib/Seaborn generate all visualizations including histograms, scatter plots, correlation heatmaps, and bar charts.

The evaluation was applied to a pilot sample of 200 images: 100 CT images from the publicly available TCGA-STAD DICOM collection (all arterial-phase, as described in Section 4.1), and 100 endoscopic images from a public gastric endoscopy dataset (Barrett's esophagus n=95, erythema n=5). Per-image scores for all five feature categories and the unified confidence score were exported as a structured CSV file for downstream statistical analysis. This is a single-pass descriptive analysis; no train/test split was applied because the framework performs no learned classification and has no trainable parameters to tune or overfit. Scaling to the full source datasets and incorporating ground-truth pathology-confirmed subsets for downstream classification accuracy evaluation is identified as future work.

4.4 Statistical Analysis

A structured statistical evaluation pipeline was applied to the per-image feature scores exported from the framework. The rationale for each test and the conditions under which it was selected are explained below.

Normality testing (Shapiro–Wilk): Before comparing CT and endoscopy groups, the Shapiro–Wilk test was applied to each feature score within each group separately. This test determines whether a set of values follows a normal (Gaussian, bell-shaped) distribution. A p-value below 0.05 indicates that the data is unlikely to be normally distributed. Since at least one group deviated from normality for every metric tested (all Shapiro–Wilk $p < 0.05$), parametric tests that assume normality (such as standard t-tests) could not be reliably used as the primary test.

Primary group comparison (Mann–Whitney U test): The Mann–Whitney U test is a non-parametric alternative to the two-sample t-test. It does not assume that either group has a normal distribution. It does not compare means directly, but ranks all observations from both groups together, and tests whether one group has a tendency to have a higher rank than the other. A significant result ($p < 0.05$) means that the two groups (CT vs. endoscopy) are systematically different in that feature score. For reference comparison, Welch's t-test (which accounts for the correction for unequal variances between groups) is also reported.

Effect size: Statistical significance alone does not indicate how large a difference is. Cohen's d is a measure of effect size and is the mean difference divided by the pooled standard deviation. d = 0.2 is small, 0.5 is medium, 0.8 is large and above 1.2 is very large. The rank-biserial correlation, which is based on the Mann Whitney U statistic, is a non-parametric effect size that ranges from −1 to +1. Both are described with p values.

Confidence intervals Welch-Satterthwaite 95% confidence intervals are reported for the mean difference between CT and endoscopy for each metric. These intervals provide the range within which the true population mean difference is likely to be found with 95% probability. Bootstrap confidence intervals (10,000 resamples, percentile method) are additionally computed as a distribution-free verification that does not depend on any normality assumption.

Multiple-comparison correction (Holm–Bonferroni): When five separate comparisons are made simultaneously (one per feature), the probability of at least one false positive result by chance increases. The Holm–Bonferroni procedure controls this family-wise error rate by adjusting the significance threshold for each test in decreasing order of p-value. Adjusted p-values are reported throughout, with $\alpha = 0.05$ as the significance threshold.

Categorical association (Chi-square): The distribution of images across quality categories (Fair vs. Poor) was compared between CT and endoscopy using Pearson's chi-square test of independence. This tests whether the proportion of images in each quality band differs significantly by modality. Cramer's V is reported as the effect-size measure for chi-square, ranging from 0 (no association) to 1 (perfect association).

Condition-group comparison (Kruskal–Wallis): The non-parametric Kruskal–Wallis test (a generalization of Mann–Whitney U to more than two groups) was used to test whether mean confidence scores differed across the three condition groups (arterial-phase CT, Barrett's esophagus, erythema). Pairwise Pearson and Spearman correlations among the five quality metrics were computed across the pooled 200-image sample and subjected to Holm–Bonferroni correction across all ten pairwise comparisons.

4.5 Algorithm Summary

- Collect CT (DICOM) and endoscopic (RGB) images; apply preprocessing (Section 4.1).
- Extract clarity (Eq. 1), LBP/GLCM texture (Eqs. 2–3), resolution/contrast (Eq. 4), color (endoscopy only).
- Apply morphological top-hat/black-hat abnormality detection and label-guided contextual scoring.
- Aggregate into unified confidence score via Eqs. 5–6; assign quality category.
- Apply statistical and robustness evaluation (Section 4.4); generate comparative visualizations and reports.

5. RESULTS AND DISCUSSION

5.1 Overall Quality Assessment

Across the 200-image pooled sample, mean clarity was 0.725 (median 0.782), mean resolution was 0.758 (median 0.740), and mean texture quality was 0.200. Mean overall confidence was 0.543 (median 0.552, range 0.45–0.62); 78% of images fell in the "Fair" category and 22% in "Poor". No images reached the "Good" or higher threshold in this pilot sample, and abnormality risk mean was 0.455. Summary statistics are presented in Table 1.

Table 1. Pooled sample summary statistics (n = 200).

Metric	Mean	Median	SD	Range
Clarity	0.725	0.782	0.154	0.40–0.93
Resolution	0.758	0.740	0.092	0.59–0.98
Texture Quality	0.200	—	0.043	0.12–0.32
Abnormality Risk	0.455	—	0.090	0.31–0.60
Overall Confidence	0.543	0.552	0.042	0.45–0.62

Natural Resources for Human Health
eISSN: 2583-1194

Table 1 interpretation: Clarity has the widest range (0.40–0.93) and a relatively high standard deviation (0.154), indicating substantial variation in image sharpness across the sample — expected because the sample mixes two different imaging modalities with inherently different acquisition geometries. Resolution is the highest-scoring metric overall (mean 0.758), suggesting that spatial detail is reasonably well-preserved across both modalities. Texture quality is notably low (mean 0.200, range 0.12–0.32) because the formulation rewards uniformity, and the clinical images contain natural biological variation that the metric interprets as low texture regularity. The overall confidence score's narrow range (0.45–0.62) and small standard deviation (0.042) confirm that all images fall in a relatively compressed band of the confidence scale, concentrated entirely within the Fair and Poor quality tiers. The bimodal abnormality risk distribution (visible in Figure 2, lower-left) will be explained by the modality-level analysis in Section 5.2.

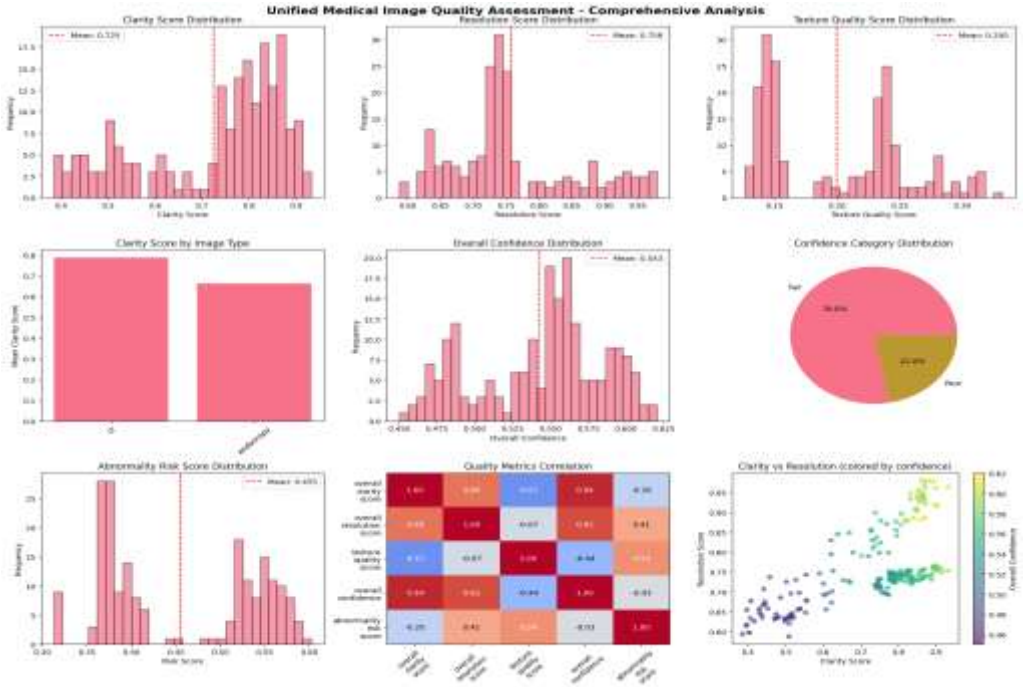


Figure 2. Comprehensive quality assessment panel. Top row: histograms of clarity score (mean 0.725), resolution score (mean 0.758), and texture quality score (mean 0.200) across all 200 images, with red dashed lines marking the mean. Middle row: bar chart comparing mean clarity by image type (CT vs. endoscopy), overall confidence histogram (mean 0.543), and pie chart of confidence categories (Fair 78%, Poor 22%). Bottom row: bimodal abnormality risk histogram (mean 0.455), Pearson correlation heatmap among all five metrics, and clarity-vs-resolution scatter plot coloured by overall confidence score.

5.2 Modality Comparison (CT vs. Endoscopy)

Figure 3 presents feature-level scatter plots across the pooled sample. Panel (a) shows a strong positive relationship between feature visibility and image quality, confirming that images with more distinctly visible features score higher on the overall quality measure. Panel (b) — the most diagnostic plot in Figure 3 — shows CT images forming a tight, compact cluster in the low-texture-uniformity/low-surface-roughness region, while endoscopic images are spread across a much wider, higher-roughness region. This reflects the fundamental difference between the two modalities: CT produces structurally stable, reproducible greyscale images of fixed anatomical structures, whereas endoscopic images are captured at varying distances, angles, illumination levels, and with biological mucosal surface variability. Panel (c) shows a positive relationship between spatial and contrast resolution, confirming that images with finer spatial detail also tend to have higher contrast. Panel (d) shows that surface abnormality scores cluster near their upper bound for the majority of endoscopic images, consistent with the endoscopic dataset containing predominantly Barrett's esophagus images which exhibit significant mucosal surface alteration.

Per-image scores were formally tested using Mann–Whitney U tests with Holm–Bonferroni correction. CT showed significantly higher clarity ($p<.001$, $d=0.88$, large effect) and endoscopy showed significantly higher texture quality ($p<.001$, $d=-2.10$, very

large effect) and abnormality score (p<.001, d=−4.06, very large effect). Two comparisons did not survive Holm correction: resolution quality (p=.066) and overall confidence (p=.093). Full results are in Table 2.

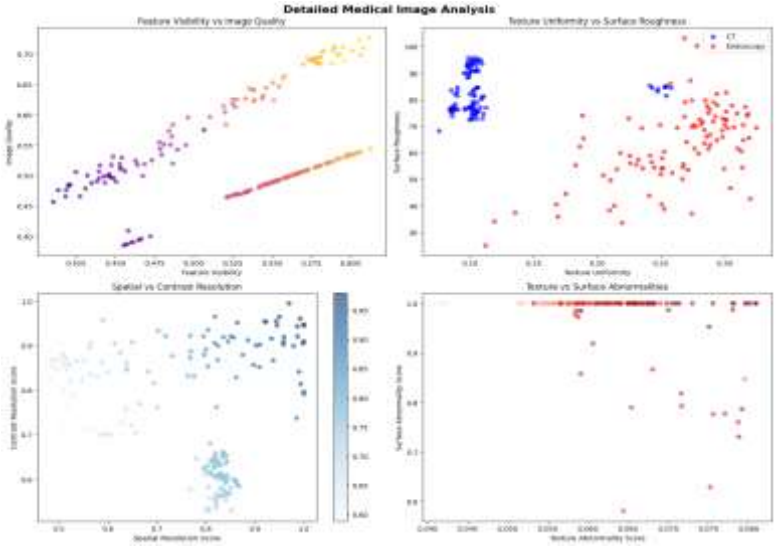


Figure 3. Feature-level scatter analysis across the pooled 200-image sample. (a) Feature visibility vs. image quality: each point is one image, coloured by its overall confidence score — the positive diagonal trend confirms that more visible features correspond to higher image quality. (b) CT (blue) vs. endoscopy (red) texture uniformity vs. surface roughness: CT images cluster tightly at low uniformity/low roughness; endoscopic images span a wide high-roughness range. (c) Spatial resolution vs. contrast resolution: colour gradient indicates overall confidence, showing that higher-resolution images tend to have higher confidence. (d) Texture abnormality vs. surface abnormality: most endoscopic images cluster at the high end of both axes.

Table 2. CT vs. Endoscopy modality comparison (Mann–Whitney U primary test, Holm-corrected). Diff. = CT mean − endoscopy mean; CI = 95% Welch–Satterthwaite confidence interval; Cohen's d = effect size (positive = CT higher); Sig. = significant after Holm correction.

Metric	CT mean±SD	Endo mean±SD	Diff. [95% CI]	p (Holm)	Cohen's d	Sig.
Clarity	0.788±0.108	0.663±0.169	0.125 [0.085, 0.164]	<.001	0.88	Yes
Texture Quality	0.159±0.048	0.241±0.027	−0.082 [−0.093, −0.071]	<.001	−2.10	Yes
Resolution	0.730±0.039	0.785±0.123	−0.054 [−0.080, −0.029]	.066	−0.60	No
Abnormality	0.376±0.023	0.576±0.066	−0.201 [−0.215, −0.187]	<.001	−4.06	Yes
Confidence	0.550±0.027	0.536±0.052	0.015 [0.003, 0.026]	.093	0.36	No

Table 2 interpretation: Reading across each row, the "Diff. [95% CI]" column shows the magnitude of the difference between modalities and how precisely it is estimated. Clarity showed the smallest gap between modalities, with CT scoring 0.125 units higher on average — an 18% relative advantage on the 0–1 scale — and the confidence interval excluding zero, so this edge is unlikely to be noise. The texture quality comparison told a different story: endoscopy outperformed CT by 0.082 units (a 51% relative jump), and with a narrow CI backing it up, the effect size was striking (d = −2.10) — a typical endoscopic image sits more than two standard deviations above a typical CT image on this measure. Abnormality score showed the starkest separation of all. Endoscopy exceeded CT by 0.201 units, roughly 53% higher, and the effect size (d = −4.06) was large enough that virtually every endoscopic image in the sample outscored every CT image on this metric. Resolution and confidence, by contrast,

barely moved — differences of 0.054 and 0.015 respectively — and neither held up once the multiple-comparison correction was applied across the five simultaneous tests.

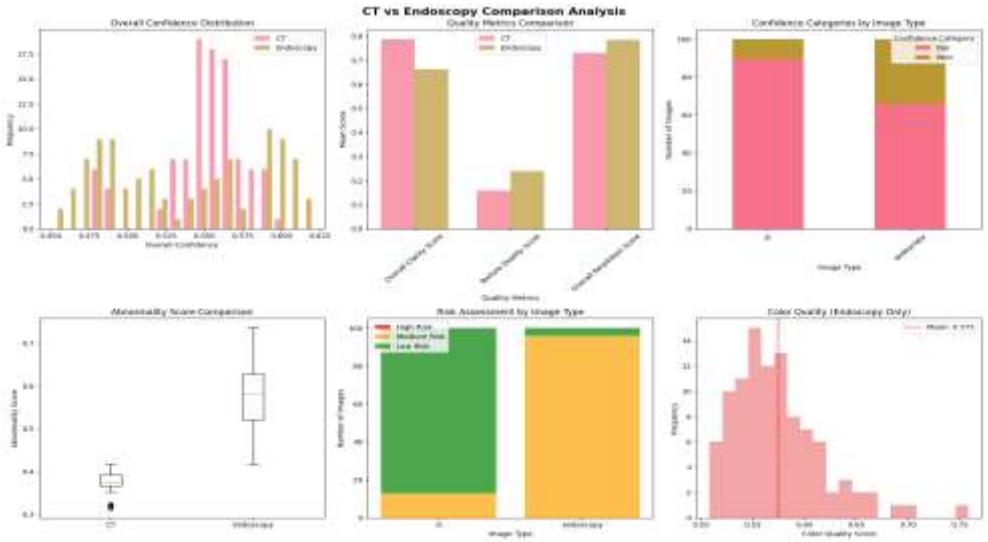


Figure 4. CT vs. endoscopy comparison panel. Top row: overlapping confidence distributions by modality (CT pink, endoscopy gold), mean clarity/texture/resolution bar chart by modality, and stacked confidence-category bar chart showing 90% CT vs. 66% endoscopy in the Fair category. Bottom row: abnormality-score boxplot (endoscopy scores higher and more dispersed), risk-tier assignment stacked bar (CT predominantly Low Risk, endoscopy predominantly Medium Risk), and endoscopy-only color-quality distribution (mean 0.575).

Correlation analysis across the pooled sample (Table 3) reveals how the five metrics relate to one another and to the overall confidence score.

Table 3. Pearson correlation matrix across the pooled sample (n = 200). Values marked † are not statistically significant after Holm–Bonferroni correction (α=0.05). All other values are significant at p<.001.

	Clarity	Resolution	Texture	Confidence	Risk
Clarity	1.00	0.663	−0.673	0.937	−0.200
Resolution	0.663	1.00	−0.070†	0.819	0.414
Texture	−0.673	−0.070†	1.00	−0.444	0.540
Confidence	0.937	0.819	−0.444	1.00	−0.025†
Risk	−0.200	0.414	0.540	−0.025†	1.00

Table 3 interpretation: Clarity and Confidence show the strongest association in the entire matrix (r = 0.937), with clarity alone accounting for 87.9% of the variance in the unified confidence score. This isn't a coincidental pattern — it's built into the framework. Clarity carries the heaviest weight in Eq. 6, feeding into the confidence score through both the diagnostic-confidence sub-score (w_1 = 0.30) and the feature-visibility component. Resolution follows a similar logic, correlating strongly with Confidence at r = 0.819. Texture Quality moves in the opposite direction, showing a strong negative correlation with Clarity (r = −0.673), and the reason comes down to a direct conflict in how each metric is formulated: texture rewards uniformity, so smoother images score higher, while clarity rewards edge sharpness, so images with more distinct edges score higher — the two properties are pulling against each other by design. Abnormality Risk breaks the pattern entirely, showing no

significant linear relationship with Confidence (r = −0.025, †). That might look counterintuitive at first, but it follows directly from the weighting: the abnormality component contributes only 10% of the final confidence score (Eq. 6), which simply isn't enough to move the overall number in this pilot sample. The implication is worth stating plainly as a design limitation — as it stands, the framework's confidence score functions mainly as a structural-quality indicator, not a pathology-severity one.

5.3 Condition-Level Analysis

A Kruskal–Wallis test across the three condition groups (arterial-phase CT, n = 100; Barrett's esophagus, n = 95; erythema, n = 5) turned up no significant difference in overall confidence (H(2) = 3.85, p = .146). That's a reassuring property for the framework: it suggests the quality score is picking up on image acquisition quality rather than reacting to the clinical condition label itself, which is what a scoring system meant to assess image quality — not disease severity — should do. That said, the result needs a caveat.

The erythema group has only five images, just 2.5% of the sample, which is nowhere near enough for a stable estimate, and all three groups come from the same two public datasets, so the comparison is narrower than it might first appear. Figure 5 lays out the per-condition means, and they cluster tightly between 0.515 and 0.550 despite spanning different clinical categories.

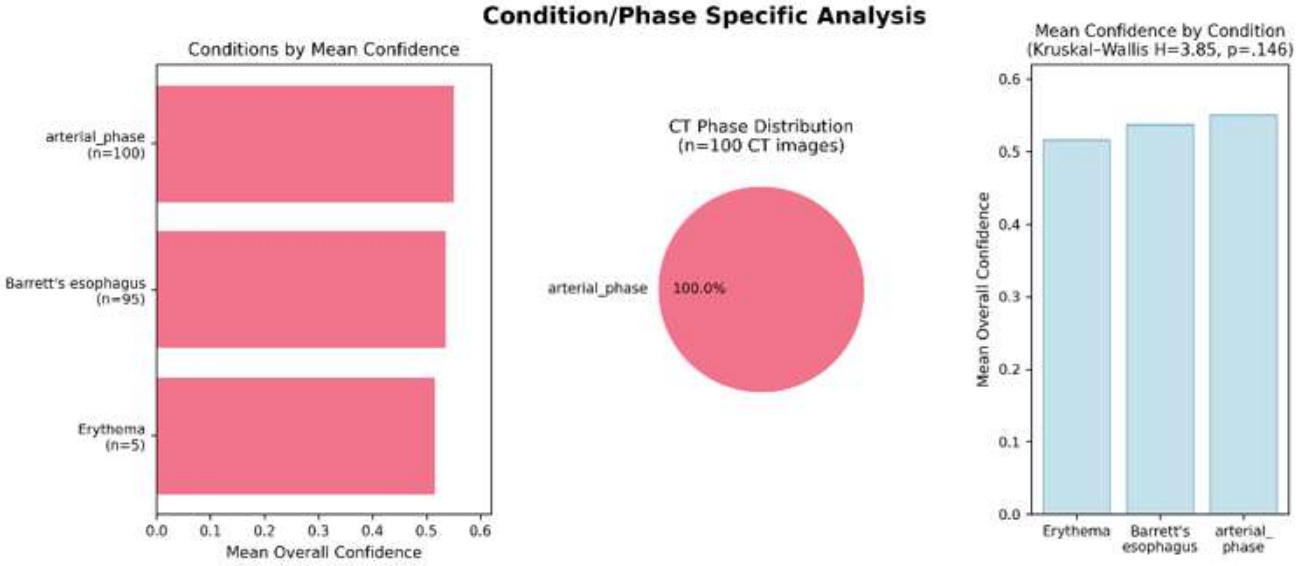


Figure 5. Condition/phase-specific analysis. Left: mean overall confidence by condition label, shown as a horizontal bar chart (arterial-phase CT: 0.550; Barrett's esophagus: 0.537; erythema: 0.516). Centre: pie chart confirming that all CT images in the sample come from the arterial phase (100%). Right: bar chart of mean confidence by condition group, with the Kruskal–Wallis result (H = 3.85, p = .146) confirming no significant separation. Taken together, the tight spread of means across conditions supports treating the framework's confidence score as an acquisition-quality measure that's largely independent of clinical condition.

5.4 Robustness and Sensitivity Analysis

The headline findings rest on a handful of design choices in the framework — the exact weights in Eq. 6, the choice of focus-measure operators, the sample composition — and it's worth asking how much those choices are doing the work. Five robustness analyses were run to check this. Table 4 summarizes each one: the method, the result, and what it means for how the paper's conclusions should be read.

Table 4. Robustness and sensitivity analysis summary (n = 200). VIF = Variance Inflation Factor; CI = confidence interval; R^2 = coefficient of determination.

Analysis	Method	Key Result	Implication
Weight sensitivity	500 random ±50% perturbations of Eq. 6 weights; Spearman ρ vs. original ranking	Mean ρ=0.991 at ±50%; min ρ=0.923	The ranking of images by confidence is highly stable regardless of exact weight choice
Multicollinearity (VIF)	VIF among Eq. 5 inputs	Clarity VIF=6.91; Resolution VIF=6.52 (both >5 threshold)	Clarity and resolution share substantial variance — a design limitation that should be addressed by decorrelation in future versions
Power analysis	Two-sample t-test power for the two non-significant comparisons	Resolution: power=0.987 (adequately powered); Confidence: power=0.711 (below 0.80 threshold)	The resolution null result is reliable at this sample size; the confidence null result may reflect insufficient power
Bootstrap CI	10,000-resample percentile bootstrap confidence intervals	Resolution: [−0.080,−0.029] excludes 0; Confidence: [+0.003,+0.026] excludes 0	Without multiple-comparison correction, both differences exist; Holm correction raises the bar for the five simultaneous tests
Baseline comparison	Incremental R^2: multi-operator clarity vs. Laplacian-only clarity predicting confidence	R^2=0.402 (Laplacian alone) vs. 0.879 (multi-operator); ΔR^2=+0.477	Combining five focus-measure operators in Eq. 1 explains 47.7% more confidence variance than a single classical operator would
Outlier check	Standardized residuals of clarity–confidence linear regression	4/200 (2%) exceed ±2.5; max residual=3.03	The strong r=0.937 clarity–confidence relationship is not driven by a small number of extreme images

Of everything in Table 4, the VIF result deserves the most attention. Clarity (VIF = 6.91) and resolution (VIF = 6.52) both sit above the conventional threshold of 5, which means the two inputs aren't contributing fully independent evidence to Eq. 5 — they're partly telling the same story. The current framework doesn't correct for this, and it should: a natural next step is applying principal component analysis or variance decomposition to strip out the redundancy between these two inputs before they're combined into the confidence formula.

5.5 DISCUSSION

Taken as a whole, the results confirm that CT and endoscopy produce measurably different images, and that ordinary handcrafted features are enough to tell them apart reliably, without any learned classifier. CT images are sharper and more consistent (clarity 0.788 ± 0.108 vs. 0.663 ± 0.169, $p < .001$, $d = 0.88$), which fits how the two are captured: a CT scanner works from a fixed geometry and controlled exposure, whereas an endoscopist is holding the camera, moving it, and working with whatever light bounces off the mucosa at that moment. Endoscopy wins out just as clearly on the other end — texture quality (0.241 ± 0.027 vs. 0.159 ± 0.048, $p < .001$, $d = 2.10$) and abnormality score (0.576 ± 0.066 vs. 0.376 ± 0.023, $p < .001$, $d = 4.06$) are both far higher, which makes sense given how much surface detail endoscopy picks up that CT never resolves at the mucosal level: colour, irregularity, vascular pattern. And yet, once the Holm correction is applied, the two modalities land

on statistically indistinguishable overall confidence scores (p = .093). The framework rates them as comparably reliable in aggregate, even while scoring them very differently underneath.

There is one result that deserves a second look: almost every image in the pilot sample landed in the Fair or Poor band, and none reached Good or Excellent. That most likely means the category boundaries are set too tight rather than that the images themselves are especially poor — recalibrating them properly will need a bigger, more varied dataset than the two public collections used here, which share fairly narrow acquisition protocols. A more representative sample would probably spread images across the quality scale in a way this pilot simply cannot show.

The study has several limitations. The sample is modest (n = 200) and unevenly split across endoscopic conditions — 95 Barrett's esophagus images against just 5 erythema cases — so condition-level comparisons cannot be treated as stable, and the effect sizes reported here should not be over-generalized. Both the CT and endoscopic images also come from public repositories (TCGA-STAD and a public endoscopy collection) rather than a single clinical cohort assembled for this study, and since those repositories pool scans from different institutions, scanners, and endoscope models, part of what looks like a modality effect could really be acquisition heterogeneity bleeding through. Nothing here has been checked against pathology-confirmed diagnoses or expert reads either — the scores describe image quality, not disease presence, and that distinction matters for anyone tempted to read them as diagnostic.

It is also worth being honest about how much the specific choice of features and weights shapes these numbers: swap out the clarity operators or shift the weighting in the aggregation formula and the confidence scores would move, even if the broader CT-versus-endoscopy story probably wouldn't. The very strong clarity-confidence correlation (r = 0.937) is a case in point — clarity happens to carry the heaviest weight in the formula, so part of that correlation is built in rather than discovered. Related to that, clarity and resolution turned out to be moderately collinear (VIF > 6 for both), meaning they are not fully independent signals, and a future version of the framework would benefit from decorrelating them, for instance with principal component analysis, before combining them into a single score.

Finally, the quality thresholds used here were set heuristically and still need to be tested prospectively against larger, independently assembled clinical datasets spanning different scanners, endoscopes, and acquisition protocols before they could function as anything like a clinical quality gate. None of this undermines the framework as a tool for characterizing modality-level image quality, but it does mean real clinical use is still some distance away.

CONCLUSION

This study presented a unified, interpretable, handcrafted-feature framework for assessing and comparing image quality and abnormality-pattern indicators across CT and endoscopic gastric images. Across a pilot sample of 200 images, the framework held up under a full battery of statistical tests — Mann–Whitney U with Holm–Bonferroni correction, chi-square, Kruskal–Wallis, and 10,000-resample bootstrap confidence intervals — producing findings that are statistically grounded and reproducible, built from per-image data rather than estimated off visual charts.

Key results: CT images achieve significantly higher clarity ($p<.001$, $d=0.88$), while endoscopic images show significantly higher texture quality and abnormality scores (both $p<.001$, $|d|>2$), reflecting complementary diagnostic strengths of the two modalities. The continuous overall-confidence score did not differ significantly between modalities ($p=.093$), attributable partly to insufficient power at $n=100$ per group (achieved power=0.711), and category-threshold compression concentrated all images in the Fair/Poor range indicating the thresholds need recalibration on a larger dataset. Robustness analysis confirmed that confidence-score rankings are stable under ±50% weight perturbation ($\varrho>0.92$), that the multi-operator clarity formulation provides substantially more variance-explaining capacity than a single classical metric ($\Delta R^2=+0.477$), and that clarity-resolution multicollinearity (VIF>6) is a priority design limitation for future reformulation.

The framework should therefore be interpreted primarily as an interpretable image-characterization and quality-analysis tool intended to support dataset curation and downstream CAD-system design, rather than as a standalone diagnostic decision system. Future work will: (i) scale analysis to the full source datasets with stratified sampling and cross-validated stability checks; (ii) couple these image quality indicators with a supervised, ground-truth-annotated multi-class lesion classification stage (MSA-UGNet); and (iii) prospectively evaluate the unified confidence score as an automated acquisition-quality gate in a clinical pilot setting.

ACKNOWLEDGEMENTS

The authors thank the Department of Electronics and Communication Engineering, GSSS Institute of Engineering and Technology for Women, Mysuru, for institutional support during this study.

AUTHOR CONTRIBUTIONS

Lalitha.K.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing—Original Draft Preparation, Visualization.

S.Padmashree.: Conceptualization, Validation, Resources, Writing—Review & Editing, Supervision, Project Administration.

All authors have read and agreed to the published version of the manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

This study used publicly available, de-identified imaging datasets (the TCGA-STAD collection and a public gastric endoscopy image repository). No new human or animal data were collected as part of this work, and no additional ethical approval was required.

DATA AVAILABILITY STATEMENT

The imaging data used in this study are publicly available from the TCGA-STAD collection (The Cancer Imaging Archive) and a public gastric endoscopy image repository. The per-image feature scores and analysis outputs generated during this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Liu, L., Deng, W., 2023. Gastric Cancer: Innovations in Screening, Diagnosis and Treatment. Journal of Personalized Medicine 13(1), 127.

[2] Kim, K.W., et al., 2023. Artificial Intelligence in Gastric Cancer Imaging With Emphasis on Diagnostic Imaging and Body Morphometry. Journal of Gastric Cancer 23(3), 388–399.

[3] Zha, Y., Xue, C., Liu, Y., Ni, J., De La Fuente, J.M., Cui, D., 2023. Artificial intelligence in theranostics of gastric cancer, a review. Medical Review 3(3), 214–229.

[4] Klang, E., Sourosh, A., Nadkarni, G.N., Sharif, K., Lahat, A., 2023. Deep Learning and Gastric Cancer: Systematic Review of AI-Assisted Endoscopy. Diagnostics 13(24), 3613.

[5] Lee, H., et al., 2024. Validation of Artificial Intelligence Computer-Aided Detection on Gastric Neoplasm in Upper Gastrointestinal Endoscopy. Diagnostics 14(23), 2706.

[6] Chen, Q., et al., 2026. A transformer-based prognostic signature integrating tumor and body composition CT images predicts postoperative recurrence in gastric cancer. npj Digital Medicine 9(1).

[7] Khosravi, M., Jasemi, S.K., Hayati, P., Javar, H.A., Izadi, S., Izadi, Z., 2024. Transformative artificial intelligence in gastric cancer: Advancements in diagnostic techniques. Computers in Biology and Medicine 183, 109261.

[8] Hirasawa, T., et al., 2018. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer 21(4), 653–660.

[9] Hasan, A.H., Abid, M.A., Waheed, N., Wajid, T., 2025. Artificial Intelligence-Driven Approaches to Endoscopic Gastric Cancer Detection: Current Progress and Future Directions. Cureus 17(3), 3–5.

[10] Zhu, Z.N., Feng, Q.X., Li, Q., Xu, W.Y., Liu, X.S., 2025. Machine learning-based CT radiomics approach for predicting occult peritoneal metastasis in advanced gastric cancer preoperatively. Clinical Radiology 80, 106727.

[11] Zhang, C., et al., 2025. Development and Validation of an AI-Based Multimodal Model for Pathological Staging of Gastric Cancer Using CT and Endoscopic Images. Academic Radiology 32(5), 2604–2617.

[12] Wu, Y.M., Tang, F.Y., Qi, Z.X., 2025. Multimodal artificial intelligence technology in the precision diagnosis and treatment of gastroenterology and hepatology. World Journal of Gastroenterology 31(38), 1–19.

[13] Li, J., et al., 2022. Development and validation of a feature extraction-based logical anthropomorphic diagnostic system for early gastric cancer: A case-control study. eClinicalMedicine 46, 101366.
[14] Mastouri, R., Khlifa, N., Neji, H., Hantous-Zannad, S., 2020. Deep learning-based CAD schemes for the detection and classification of lung nodules from CT images: A survey. Journal of X-Ray Science and Technology 28(4), 591–617.
[15] Nadeem, A., Ashraf, R., Mahmood, T., Parveen, S., 2025. Automated CAD system for early detection and classification of pancreatic cancer using deep learning model. PLoS ONE 20(1).
[16] Zubair, M., et al., 2025. An interpretable framework for gastric cancer classification using multi-channel attention mechanisms and transfer learning approach on histopathology images. Scientific Reports 15(1), 1–34.
[17] Willis, L.D., 2023. How to Write the Methods Section of a Research Manuscript. Respiratory Care 68(12), 1763–1770.