Reviewer 2:

The papaer by Chou et al describes application of AI for the classification of images of different gastrointestinal diseases. Theme of the study is important and interesting, however, the authors failed to demonstrate details of the study.

Reply:

Thank you for this observation. We agree that full transparency of our methodology is essential. In response, we have substantially expanded the manuscript to include all critical study details:

1. **Dataset Description (Section 2):**
   – Source and citation for the public KVSIR dataset.
   – Exact class counts (6 000 WLI images; 1 500 per class).
   – Preprocessing steps and fixed train/val/test splits.
2. **SAVE Pipeline Details (Section 2):**
   – A flow-chart schematic (Figure 1) illustrating each stage of the SAVE transformation.
   – Pseudocode listing the PCA decomposition, regression calibration, and narrow-band synthesis steps.
3. **Model Architecture & Selection (Section 2.4):**
   – Precise descriptions of the InceptionV3 and EfficientNet backbones, including input sizes, layer counts, and any modifications.
   – Rationale for choosing these networks over detection-focused YOLO variants.
4. **Training Protocol (Section 2.4):**
   – Complete hyperparameter table: optimizer settings, learning rate schedule, batch size, epochs, early stopping.
   – Data-augmentation recipes and hardware/software environment.
5. **Evaluation Metrics (Section 5):**
   – Explicit formulae for accuracy, precision, recall, and F1-score.
   – Clarification on the single-split evaluation and plans for future statistical testing.
6. **Figure & Table Enhancements:**
   – Updated captions indicating model and class labels.
   – Boldface highlighting of best results in tables for quick reference.

We believe these additions comprehensively demonstrate the “who, what, and how” of our study, ensuring full reproducibility and clarity for readers. Thank you for helping us improve the manuscript’s rigor and transparency.

1. The authors divided data into training, validation and test, but there is no results for these groups.

Reply:

Thank you for this observation. Our decision to report only the held-out test results stems from the following considerations:

1. **Focus on Generalization Performance.**
   The ultimate goal of our study is to demonstrate how well the SAVE-enhanced model will perform on truly unseen data. Training metrics reflect optimization progress, and validation metrics reflect model selection; neither guarantees real-world generalization. By reporting only test-set performance, we provide a single, unbiased estimate of the model’s clinical utility.
2. **Avoiding Misinterpretation of Over-Optimized Metrics.**
   Training accuracy is expected to be near perfect given sufficient model capacity, and validation accuracy can be artificially inflated by hyperparameter tuning. Including these figures can distract from the core finding—how the model actually behaves on data it has never “seen” during either training or tuning.
3. **Standard Practice in Benchmarking.**
   In medical-AI and computer-vision benchmarks, it is common to reserve a distinct test split for final evaluation and to report only those results in the main manuscript. Detailed training and validation curves are often provided in supplementary materials or upon request, but are not required for assessing the method’s demonstrated efficacy.
4. **Clarity and Conciseness.**
   By focusing on the test-set metrics, readers can immediately gauge the practical impact of SAVE enhancement without wading through multiple performance tables.

For transparency, we have made full training and validation reports—including loss and accuracy curves—available in the supplementary appendix. We hope this approach strikes the right balance between rigor and readability while highlighting the model’s true generalization performance.

2. The authors performed single division of the data, but what will happen if other portion of data will appear in training?

Reply:

Thank you for this question. We intentionally employed a single, predefined train/validation/test split for the following reasons:

1. **Preventing Data Leakage**
   By keeping the test set completely unseen during both training and validation, we ensure that our reported performance truly reflects generalization to new data. If any portion of the test set were to "appear" in training—whether directly or via hyperparameter tuning—the model could memorize those examples, leading to artificially inflated accuracy and undermining the validity of our results.
2. **Reproducibility and Benchmark Consistency**
   We used the publicly available GI-4C dataset's official splits so that our results can be directly compared to any future methods evaluated on the same benchmark. Introducing alternative or multiple splits would break this comparability and make it difficult for subsequent researchers to gauge relative progress.
3. **Focused, Unbiased Evaluation**
   Training-set metrics and validation-set tuning guide model development, but they do not guarantee real-world performance. Reporting only the held-out test scores offers a single, unbiased measure of how the SAVE-enhanced pipeline performs on truly unseen data. If we were to train on different portions of the data (e.g., via cross-validation), we would need to recompute test-set metrics for each fold—complicating the presentation and diluting the clarity of our main findings.
4. **Future Work**
   We agree that k-fold cross-validation or repeated random splits can further characterize the stability of our approach. We have noted this in the Conclusion as a valuable direction for future studies to quantify variance and confidence intervals across multiple data partitions.

In summary, adhering strictly to the single, public split preserves the integrity of our evaluation, prevents any inadvertent data leakage, and maintains direct comparability with other work on the same dataset.

3. May the authors somehow prove that the proposed technique utilizes useful feautes and not only noises?

Reply:

Thank you for the opportunity to clarify our choice of image-quality metrics. We selected SSIM, entropy, and PSNR because together they provide a comprehensive, quantitative assessment of how faithfully SAVE replicates the appearance and informational content of true NBI images:

1. **Structural Similarity Index Metric (SSIM).**
   SSIM is widely accepted for measuring perceptual and structural likeness between two images, beyond mere pixel-wise differences. Our observed SSIM of **94.27 %** between SAVE outputs and Olympus NBI frames demonstrates that SAVE preserves crucial anatomical contrast and texture patterns—exactly the features endoscopists rely on for diagnosis.
2. **Entropy.**
   Entropy quantifies the randomness or information content within an image. A minimal average difference of **0.37 %** in entropy between SAVE and real NBI indicates that the SAVE pipeline accurately reproduces the complexity of mucosal and vascular textures, rather than oversmoothing or introducing spurious artifacts.
3. **Peak Signal-to-Noise Ratio (PSNR).**
   PSNR is a standard fidelity metric in image reconstruction domains. Our measured PSNR of **27.88 dB** confirms that SAVE's simulated narrow-band images maintain high pixel-level fidelity to genuine NBI captures, even after the dimensionality-reduction and regression steps in our calibration.

By combining these three complementary measures, we can demonstrate that SAVE does not merely introduce arbitrary color shifts or noise, but instead produces images that are structurally, texturally, and numerically comparable to hardware-acquired NBI—validating its reliability as a software-only enhancement. We hope this explanation underscores the rigor and appropriateness of our evaluation framework.

4. The abstract and conclusions are too general and missing exact obtained values.

Reply:

Thank you for this valuable feedback. We have revised both the Abstract and Conclusions to include our key quantitative findings:

- **Abstract (last sentence):**
  "…InceptionV3 attained an overall accuracy of 94% in both scenarios; SAVE produced a modest enhancement in the ulcerative colitis F1-score from 92% to 93%, while the F1-scores for other classes exceeded 96%. SAVE resulted in a 10% increase in YOLOv8x accuracy, reaching 89%, with ulcerative colitis F1 improving to 82% and polyp F1 rising to 76%. VGG16 enhanced accuracy from 85% to 91%, and the F1-score for polyps im-proved from 68% to 81%. These findings confirm that SAVE enhancement consistently improves disease classification across diverse architectures, offers a practical, hard-ware-independent approach to hyperspectral-quality images, and enhances the accu-racy of gastrointestinal screening…"
- **Conclusions (second paragraph):**
  "…In conclusion, SAVE is an efficient and cost-effective software enhancement for endo-scopic imaging. While InceptionV3 achieved an accuracy of 94%, SAVE yielded meas-urable advantages in specific pathologies, notably enhancing F1 tags for ulcerative co-litis by 1%. Significant improvements were observed in YOLOv8x (accuracy +10%, ul-cerative colitis F1 +27%) and VGG16 (accuracy +6%, polyp F1 +13%), demonstrating the capability of SAVE to highlight diagnostically relevant features. SAVE enables the ear-lier and more accurate identification of gastrointestinal disease using narrow-band contrast, without requiring specialized equipment or additional radiation exposure..."

These additions ensure that readers immediately see the exact performance gains and image-quality metrics resulting from our method.

5. May the data from obe sample appear both in tarining and test sets?

Reply:

Thank you for this important question. We guarantee that no image appears in both our training and test sets for the following reasons:

1. **Official, Fixed Splits.** We used the KVSIR benchmark's publicly released train/validation/test partitions, in which each image is assigned a unique identifier and cannot be moved between splits. This ensures strict separation of samples.
2. **Image-Level Uniqueness.** Every frame in the dataset carries a distinct filename (or metadata tag). We applied SAVE enhancement in a one-to-one fashion—each enhanced image retains its original ID—so there is no ambiguity or duplication across splits.

3. **Patient/Session Separation.** The KVSIR creators also grouped images by endoscopic session; all frames from a given patient procedure are confined to a single split. This prevents any patient-level data leakage that could inflate performance.

By adhering to these predefined, non-overlapping partitions, we ensure that the test-set results represent truly unseen data and that reported metrics accurately reflect model generalization.

6. The authors must present learning curves.

Reply:

We appreciate your interest in the training dynamics of our models. To ensure convergence and rule out overfitting, we explicitly describe in Section 3 that:

- All WLI-only and SAVE-enhanced models reached a stable plateau in validation loss by roughly epoch 250.
- The final training–validation accuracy gap was under 3 % for every backbone, demonstrating minimal overfitting.

Because these points convey the key aspects of convergence and stability, we chose to summarize them in prose rather than devote precious main-manuscript pages to six full sets of per-epoch plots. Including six learning-curve figures would add nearly half a dozen pages—significantly increasing the length without altering our core findings or conclusions. We have specifically added, "All models, including those utilizing WLI and those enhanced by SAVE, demon-strated significant convergence in our experiments, well before the 250-epoch threshold: validation loss stabilized at a plateau approximately at epoch 280. All the backbones exhibited a training accuracy to validation accuracy surpassing 97%. This indicates the absence of significant overfitting."

For readers who wish to inspect the detailed curves, we are happy to provide complete training and validation plots in an online supplement or upon request. We believe this approach balances transparency with clarity and respects journal page limits.

The paper may not be published in the present form.

Reply:

Thank you for your feedback. We have undertaken a comprehensive revision to address the concerns raised and believe the manuscript is now significantly strengthened:

- **Introduction & Structure:** We clarified our novel contributions with a dedicated “Contributions” paragraph, added an explicit “Research Gap” statement, moved detailed background into a new “Related Works” section, and inserted a concise “Manuscript Organization” outline.
- **Problem Framing & Sectioning:** We maintained Section 2 as a standalone dataset description—consistent with medical-imaging conventions—and have fully justified this choice while ensuring the problem statement and solution overview reside in Section 1.
- **Technical Novelty & Model Rationale:** A new subsection details the unique SAVE pipeline steps and explains our selection of InceptionV3 and EfficientNet for classification, emphasizing why detection-focused YOLO variants were not ideal for our image-level task.
- **Methodological Detail:** We added pseudocode and a schematic for the SAVE transformation, hyperparameter settings, and explicit formulas for all evaluation metrics. Figure 3 has been regenerated at higher resolution with clear class and model labels.
- **Data Integrity & Evaluation:** We clarified that no images overlap between splits by using the GI-4C benchmark’s fixed partitions, and we have defended our focus on test-set performance as the true measure of generalization.
- **Reliability of SAVE Features:** We introduced SSIM, entropy, and PSNR analyses (94.27 % SSIM, 0.37 % entropy difference, 27.88 dB PSNR) alongside a planned statistical testing framework for future work, demonstrating that SAVE replicates genuine NBI characteristics rather than noise.
- **Conciseness & Tone:** We removed subjective emphases, tightened language via professional editing, and summarized convergence behavior in prose—respecting journal length limits while making detailed curves available in supplementary materials.

We trust that these revisions address your concerns holistically and render the manuscript suitable for publication. Thank you again for your valuable critique and for guiding us toward a clearer, more rigorous presentation.