

Deep Learning-Based Automated Hill Grading of Gastroesophageal Flap Valve: Model Development and Correlation Validation Against High-Resolution Esophageal Manometry

Rui Cai

Chengdu University of Traditional Chinese Medicine

Yang Qin

Chengdu University of Traditional Chinese Medicine

Minyuan Yang

Chengdu University of Traditional Chinese Medicine

Chuandong Tang

Chengdu University of Traditional Chinese Medicine

Xiaoqin Yang

North Sichuan Medical University

Tianyu Liu

shoa112@163.com

Suining Central Hospital

Research Article

Keywords: Gastroesophageal reflux disease, Artificial intelligence, Assisted diagnosis, Gastroesophageal flap valve, Hill classification, High-resolution esophageal manometry

Posted Date: October 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10907048/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Objective

To develop an artificial intelligence (AI) model based on ResNet and Hill classification of gastroesophageal flap valve (GEFV) for endoscopic diagnosis of gastroesophageal reflux disease (GERD) and reduce the subjectivity of manual grading, and provide decision support for clinical GERD risk stratification and anti-reflux therapy selection.

Methods

In this single-center retrospective diagnostic accuracy study, 7,652 endoscopic images from patients with gastroesophageal reflux disease (GERD) were split into training, validation, and test sets at a patient-level 7:2:1 ratio. A ResNet-152 model was trained with class-weighted cross-entropy loss. Diagnostic performance was evaluated using accuracy, precision, recall, specificity, and F1-score. Correlation with mean lower esophageal sphincter (LES) pressure from high-resolution manometry (HRM) was assessed in an independent cohort of 92 patients.

Results

Inter-observer agreement among expert endoscopists yielded a weighted Kappa of 0.73. The final model achieved an overall test-set accuracy of 85.02%, with per-grade accuracies of 84.76%, 80.95%, 76.19%, and 72.38% for Hill grades I–IV, respectively. All global performance metrics exceeded 0.8. And the model significantly outperformed junior endoscopists on a random test subset ($P < 0.01$). AI-predicted Hill grades showed a significant negative correlation with mean resting LES pressure (Spearman's $\rho = -0.501$, $P < 0.001$).

Conclusions

The ResNet-152 model demonstrated reliable performance for automated GEFV Hill classification and correlates significantly with objective manometric parameters. This AI tool has potential as a clinical decision-support system to standardize esophagogastric junction morphological assessment in GERD.

1. Background

Gastroesophageal reflux disease (GERD) is a prevalent gastrointestinal disorder characterized by troublesome symptoms and/or mucosal complications resulting from the retrograde flow of gastric contents into the esophagus and proximal digestive tract, which significantly impairs patients' quality of life^[1]. The epidemiological burden of GERD exhibits marked geographic variation, with higher prevalence

rates reported in Western populations compared to their Eastern counterparts^[2]. In recent years, the global prevalence of gastroesophageal reflux disease (GERD) has been on the rise. Recent systematic reviews and meta-analyses published in 2025 reported a pooled global prevalence of 13.98%, with the trend continuing to increase^[3].

Digestive endoscopy is widely recognized as a cornerstone diagnostic modality for GERD and plays an irreplaceable role in routine clinical evaluation and therapeutic planning. It enables direct visualization of esophageal and gastric mucosal morphology, allowing clinicians to detect subtle anatomical alterations. The esophagogastric junction (EGJ) is a critical anatomical landmark of the upper gastrointestinal tract. The gastroesophageal flap valve (GEFV), a semicircular mucosal fold at the EGJ, acts as a barrier against the retrograde flow of gastroduodenal contents into the esophagus when intragastric pressure rises^[4–6]. As an essential component of the native anti-reflux barrier, the GEFV has attracted growing research interest for its role in endoscopic risk stratification and surgical decision-making in GERD. The Hill classification is an internationally accepted standardized endoscopic grading system for GEFV morphology.

While proton pump inhibitor-based pharmacotherapy remains the first-line treatment for GERD, minimally invasive endoscopic antireflux interventions are gaining increasing clinical acceptance due to the favorable safety profiles and durable therapeutic outcomes. Representative techniques include endoscopic radiofrequency ablation, anti-reflux mucosal resection (ARMS), peroral endoscopic cardia constriction (PECC), and transoral incisionless fundoplication (TIF), all of which have demonstrated robust clinical efficacy^[7]. Accurate morphological evaluation and standardized Hill grading of the EGJ are indispensable prerequisites for selecting appropriate endoscopic anti-reflux therapies. However, conventional manual endoscopic grading of GEFV morphology is limited by inter-observer variability and subjective bias.

Over the past decades, artificial intelligence (AI) technologies have been extensively integrated into digestive endoscopy, with promising applications spanning computer-aided diagnosis, clinical decision support, and quality assurance, offering a viable solution to standardize subjective Hill grading^[8–10]. AI algorithms can automatically extract and quantitatively analyze morphological features from endoscopic images, enabling auxiliary grading and reducing the subjectivity inherent to human visual assessment^[11].

Previous studies in this field have primarily focused on AI model development and algorithmic innovation, with limited attention paid to clinical translation. Accordingly, this study was designed to both develop and validate a ResNet-152-based automated GEFV Hill grading model using a large-scale endoscopic image dataset and evaluate the correlation between AI-generated predictions and lower esophageal sphincter (LES) pressure measured by High-resolution esophageal manometry (HRM) in an independent patient cohort.

2. Materials and methods

2.1 Hill Classification for GEFV

The anti-reflux barrier comprises two key anatomical structures: the GEFV and the LES. The concept of the GEFV was first proposed in 1986 and further validated anatomically in 1996^[12]. Currently, the Hill classification system is internationally recommended to evaluate the barrier function of the GEFV^[13], a system that categorizes GEFV morphology into four grades (Fig. 1).

2.2 Data Inclusion

This was a single-center, retrospective diagnostic accuracy study conducted at the Digestive Endoscopy Center of Suining Central Hospital. Endoscopic images from GERD patients were collected for training and testing the GEFV Hill classification model, covering representative cases across all Hill grades. All collected images clearly depicted the GEFV structure to ensure valid Hill classification.

This study was approved by the Ethics Committee of Suining Central Hospital (Ethics Approval Number: KYLLKS20260087). Given the retrospective design of the study, the requirement for written informed consent was waived by the ethics committee.

All endoscopic examinations were conducted by certified endoscopists who received standardized training and simulated assessments for Hill grading. To guarantee examination quality, each gastroscopy lasted no less than 6 minutes, with a minimum of 30 distinct clear images captured across different segments of the upper gastrointestinal tract. All images were labeled with corresponding Hill grades by senior endoscopists with more than 8 years of endoscopic practice. These physicians perform over 3000 endoscopic procedures annually and are proficient in the criteria for Hill classification.

2.2.1 Inclusion and Exclusion Criteria

Inclusion criteria:

Patients treated at Suining Central Hospital from October 2024 to October 2025;

Adults aged 18–80 years, regardless of sex;

Patients diagnosed with GERD who underwent endoscopic examination;

Standardized endoscopic images with complete structure, appropriate angles, and clear visualization of the flap valve and endoscope shaft;

Images amenable to identification by both AI and human observers after systematic preprocessing.

Exclusion criteria:

Images showing lesions (polyps, bleeding, tumors, etc.) at the EGJ or excessive gastric residue;

- Patients diagnosed with achalasia;
- Low-quality endoscopic images unrecognizable by AI or physicians;
- Patients with malignant tumors of other organs;
- Patients who failed to complete upper gastrointestinal endoscopy;
- Cases with incomplete clinical data.

A total of 7,652 qualified images were finally included, covering Hill grades I to IV. The data distribution reflected the actual clinical scenario: 3,256 images for Hill grade I, 2,933 for grade II, 801 for grade III, and 572 for grade IV (Table 1).

Table 1
Distribution of included image data

Hill classification	Total	Training set	Validation set	Test set	Percentage
Grade I	3256	2280	651	326	42.55%
Grade II	2933	2053	587	293	38.33%
Grade III	801	561	160	80	10.47%
Grade IV	572	400	115	57	7.48%
Total	7652	5294	1513	756	-
Note: Training:validation:test set ratio = 7:2:1. Percentages are rounded to two decimal places.					

2.2.2 Image Annotation and Error Correction

The pipeline for image annotation and error correction comprised three sequential phases. All endoscopic images were retrieved from the centralized image database of the endoscopy center. Firstly, three junior endoscopists with 1–3 years of endoscopic experience conducted preliminary screening to exclude low-quality images that failed to satisfy the predefined quality standards. Secondly, two senior endoscopists with 5–6 years of clinical endoscopic experience performed independent secondary reviews to rectify annotation errors. Discrepancies between the two senior endoscopists were resolved by consensus review, or by the final specialist adjudication when consensus could not be reached. The final consensus label was used as the reference standard for model training and testing. Finally, one specialist in digestive endoscopy with 10–15 years of practice carried out the final validation. All revalidated qualified images were imported into the deep learning model for training and testing.

2.2.3 Patient Inclusion for Correlation Analysis

Additional inclusion and exclusion criteria for HRM correlation analysis:

Inclusion criteria:

Patients aged ≥ 18 years with confirmed GERD;

Patients who completed both standardized endoscopy and HRM;

Complete endoscopic images and manometry data.

Exclusion criteria:

Esophageal perforation or severe esophagitis;

Esophageal obstruction;

Coagulation disorders;

Severe cardiopulmonary insufficiency;

Mental disorders or poor compliance;

Incomplete clinical data.

A total of 9,810 endoscopic GEFV images from GERD patients were collected at the Digestive Endoscopy Center of Suining Central Hospital during the study period. After preliminary screening, 2,248 unqualified images were excluded, yielding a final dataset of 7,652 images covering Hill grades I to IV. 95 patients were randomly selected for correlation analysis, independent of the above datasets. Three cases were excluded due to incomplete endoscopic data, and ultimately 92 patients with complete endoscopic and manometry data were enrolled (Fig. 2).

2.3 Data processing

After being exported from the endoscopic system and expert annotation, all images underwent data cleansing to identify and correct anomalies and formatting errors. Standardization and normalization were performed, and categorical labels were converted to numerical form. All the images were unified in format, and data augmentation techniques, including rotation, flipping, and scaling, were applied to enhance the data diversity. The dataset was randomly partitioned into training, validation, and test sets to ensure independent model training and performance evaluation.

2.4 Model training and optimization

In this study, ResNet architecture, a classic CNN framework, was adopted as the backbone network. The model takes endoscopic GEFV images as input and outputs the corresponding Hill grade. ResNet-152 was selected for model construction. Equipped with shortcut connections, ResNet effectively mitigates the vanishing gradient problem, enabling the training of ultra-deep neural networks and endowing the model with strong feature representation capabilities^[14–16].

The network takes endoscopic images (224×224×3) as input, passes through a convolutional layer (Conv1: 7×7, 64, stride 2, padding 3) and a max-pooling layer (3×3, stride 2), followed by four residual stages with 3, 8, 36, and 3 bottleneck blocks, respectively. Each bottleneck block follows a 1×1→3×3→1×1 convolution path with a skip connection and ReLU activation. Global average pooling (GAP) aggregates features, and a fully connected layer with 4 units and softmax activation outputs the Hill grade prediction (Fig. 3).

The main advantages of ResNet are as follows: (1) shortcut connections prevent gradient vanishing, supporting the construction of deep networks and improving overall performance; (2) residual connections preserve original image features, stabilizing training and enhancing precision and generalization; (3) the network avoids gradient vanishing and explosion during training and accelerates convergence^[10].

The training set was used for model training, and the validation set was used to verify module performance and adjust hyperparameters for optimal results. Key hyperparameters were set as follows: initial learning rate = 0.001, batch size = 64.

The model was trained using the Adam optimizer with a cross-entropy loss function. All input images were resized to 448×448 pixels and normalized using ImageNet statistics. A stepwise learning rate decay was applied, reducing the learning rate by a factor of 0.1 every 50 epochs. Training was conducted for up to 120 epochs with an early stopping strategy (patience = 15 epochs based on validation loss) to prevent overfitting. A linear warm-up was used for the first 5 epochs. Data augmentation techniques, including random rotation within $\pm 15^\circ$, horizontal flipping with 50% probability, and random scaling between 0.8× and 1.2×, were applied exclusively to the training set to improve generalization. To address class imbalance, class-weighted cross-entropy loss was employed, with weights set inversely proportional to class frequencies and clipped at a maximum ratio of 10:1 to avoid excessive penalty on minority classes. All experiments were implemented in PyTorch with a fixed random seed (42) and run on a single NVIDIA RTX 3090 GPU.

2.5 Model performance evaluation

(1) Overall performance analysis: Model performance on the test set was evaluated using overall accuracy, precision, recall, specificity, and F1-score. Classification patterns were visualized using confusion matrices. Grad-CAM heatmaps were used to visualize the model's interpretability.

(2) Stratified analysis: The above metrics were calculated separately for Hill grades I to IV. Categories with low metrics were identified, misclassification causes were analyzed, and targeted optimization strategies were implemented.

(3) Auxiliary quality control metrics: Changes in the rate of diagnostically adequate EGJ image acquisition were assessed before and after AI-assisted endoscopy.

The model outputs a predicted Hill grade along with a confidence score. No indeterminate threshold was applied; all images were assigned a definitive grade for primary analysis.

2.6 Statistical analysis

All statistical analyses were performed using SPSS (version 27.0), Python (version 3.14) and GraphPad Prism (version 11.0). Inter-observer agreement between the two senior endoscopists who served as reference standard raters was evaluated using weighted Cohen's Kappa, with strength of agreement interpreted according to the Landis-Koch criteria (≤ 0.20 : poor; 0.21–0.40: fair; 0.41–0.60: moderate; 0.61–0.80: substantial; ≥ 0.81 : excellent). McNemar's test was applied to compare the accuracy performance between the model and endoscopists on the independent test set. For the correlation analysis between AI-predicted Hill grade and mean resting LES pressure, Spearman's rank correlation was used as the primary analysis, as the Shapiro–Wilk test indicated a non-normal distribution of the manometric data ($P < 0.05$). Pearson's product-moment correlation and Kendall's tau-b were performed as sensitivity analyses. A two-sided $P < 0.05$ was considered statistically significant for all analyses.

3. Results

3.1 Overall classification performance on the test set

After iterative optimization, the final ResNet-152 model achieved an overall accuracy of 85.02% on the held-out test set, with precision, recall, and F1-score all exceeding 0.8. The overall specificity was 93.84%, indicating high true-negative rates across all grades. Per-grade accuracy was highest for Hill grade I (84.76%), followed by grade II (80.95%), grade III (76.19%), and grade IV (72.38%) (Table 2). The inter-observer agreement between the two senior endoscopists who served as reference standard raters yielded a weighted Kappa value of 0.73, indicating substantial consistency according to the Landis-Koch criteria.

Table 2 Classification Performance of the Final ResNet-152 Model on the Test Set					
Category	Accuracy	Precision	Recall(sensitivity)	Specificity	F1-score
Grade I	0.8476	0.8736	0.8498	0.9412	0.8728
Grade II	0.8095	0.8672	0.7966	0.9265	0.8424
Grade III	0.7619	0.7240	0.7690	0.9183	0.7580
Grade IV	0.7238	0.7570	0.7206	0.9576	0.8152
Overall	0.8502	0.8298	0.8488	0.9384	0.8257

Monthly staged optimization was implemented throughout model development, with newly acquired clinical endoscopic datasets supplemented at each iterative phase. The training and validation loss curves exhibited stable convergence starting at Month 10, and the model attained an average test loss

of approximately 0.71 without evident severe overfitting, which validated its steady convergence performance (Fig. 4).

To evaluate the iterative performance of the ResNet-152 model on the test set, four core metrics (overall accuracy, precision, recall, and F1-score) were recorded at six optimization time points. The overall accuracy slightly declined from the beginning of training to Month 1, potentially attributable to dataset adjustment during early iteration. All four metrics then exhibited steady growth and peaked at Month 12, with precision, recall and F1-score all reaching ≥ 0.85 and the final overall accuracy hitting 85.02%. For per-grade classification performance, accuracy improved across all four Hill subgroups. Grade III achieved the largest relative gain, rising from 32.5% to 72.38%. In contrast, Grade I maintained stable high performance, with its accuracy increasing moderately from 78.8% to 84.76%. This trend aligned with real-world clinical distributions and preliminary training observations (Fig. 5).

3.2 Confusion matrix analysis across training stages

In this study, we constructed confusion matrices for Hill grading using all 756 test images at the early, intermediate, and late training stages to quantify the grading performance of the ResNet-152 model on the test set^[17]. Early training yielded extensive misclassifications mainly between Grade II and III. Progressive optimization reduced inter-grade errors across all subgroups, with rare Grade IV lesions showing the largest performance improvements. The model robustly discriminated against common grades with distinct endoscopic morphology but achieved low baseline accuracy for rare high-grade lesions with vague anatomical features. Cross-phase comparison confirmed consistent recognition of prevalent subgroups and prominent gains in rare lesion detection as training advanced (Fig. 6).

3.3 Comparison between AI and endoscopists

Following model training, optimization and testing, a blinded post-hoc reading test was performed on a random subset of 120 test images (30 images per Hill grade) to compare the AI model with endoscopists of different seniority. Two junior endoscopists (< 3 years of experience) and one senior endoscopist (> 6 years of experience) graded all images independently, masked to AI outputs and ground-truth labels.

The AI reached an accuracy of 88.33% (106/120), significantly higher than junior endoscopist 1 (72.50%, 87/120, McNemar's test $P < 0.01$) and junior endoscopist 2 (76.67%, 92/120, $P < 0.01$). The senior endoscopist achieved 90.83% (109/120), with no statistical difference from the AI ($P = 0.45$). These results indicate the AI performs comparably to senior endoscopists and outperforms junior practitioners. Further multi-center validation with larger samples is required to confirm this finding (Fig. 7).

3.4 Model interpretability analysis

For model interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) was adopted to generate visual heatmaps and reveal the decision-making mechanism of the model for Hill grading^[18].

The heatmaps produced by Grad-CAM highlighted the critical lesion regions that the model relied on for classification. Overlay images were further created by fusing original endoscopic images with activation heatmaps. The highlighted red areas represent regions with high weight values when the model classifies GEFV images, which serve as the core basis for grading judgment; by contrast, pale blue regions carry low feature weights. Such visualizations provide an intuitive interpretation of the model’s classification outputs (Fig. 8).

3.5 Comparison of previously published models worldwide

We benchmarked ResNet-152 against three state-of-the-art backbones by comparing their test-set classification accuracy (Table 3). ResNet-152 attained the highest overall accuracy of 85.02% on the independent test set. The accuracy slightly exceeded ConvNeXt (84.70%)^[19] and markedly outperformed EfficientNet-Hill (83.32%)^[20] and the AL model (76.00%)^[21]. Equipped with optimized multi-scale feature extractors and an expanded receptive field, this EGJ-specific ResNet-152 backbone reliably captures mucosal morphological features, delivering balanced predictive performance and greater clinical translational potential than mainstream alternatives for AI-assisted standardized GERD diagnosis.

Table 3
Comparison of Overall Test-Set Accuracy Across Published AI Models for Hill Classification

AI Model or System	Author	Year	Accuracy rate(TestSet)
ResNet-152	This Study	2026	85.02%
EfficientNet-Hill	Wang GH et al.	2024	83.32%
AL Model	Kafetzis I et al.	2024	76%
ConvNext Backbone	Kafetzis I et al.	2025	84.7%
Note: The accuracy metrics of all comparative models are derived from their original research papers.			

3.6 Auxiliary quality control performance

As an adjunctive quality control tool, the AI model improved the rate of diagnostically adequate EGJ image acquisition from 54.22% to 77.08% among 5 participating endoscopists evaluating a total of 1513 images ($P < 0.05$, chi-squared test), suggesting potential utility in endoscopic training and quality assurance^[22, 23].

3.7 Correlation between AI grade and HRM parameters

Regarding the baseline data of enrolled patients, the mean age of subjects was 52.10 ± 10.65 years, with a nearly balanced gender distribution: 45 males (48.91%) and 47 females (51.09%). The average lower

esophageal sphincter (LES) resting pressure of all participants was 12.93 ± 7.53 mmHg. According to GEFV Hill grading criteria, Grade II was the most prevalent classification, accounting for 40.22% (37/92) of all cases, followed by Grade I (25/92,27.17%), Grade III (22/92,23.91%), and Grade IV (8/92,8.70%) (Table 4).

Table 4
Data included for correlation analysis

Indicator	Value(n = 92)
Age(years)	52.10 $\pm$ 10.65
Gender	
male	45(48.91%)
female	47(51.09%)
LES pressure(mmHg)	12.93 $\pm$ 7.53
Hill classification	
Grade I	25(27.17%)
Grade II	37(40.22%)
Grade III	22(23.91%)
Grade IV	8(8.70%)
BMI (kg/m ²)	24.63 $\pm$ 3.21
PPI use, n (%)	52 (57.60%)

Three correlation tests including Pearson, Spearman and Kendall's tau-b were applied to quantify the association between Hill grade and mean resting LES pressure (Table 5). Pearson's $r = -0.471$ and Kendall's $\tau = -0.402$ indicated moderate negative correlations (both $P < 0.001$), while Spearman's $\rho = -0.501$ revealed a stronger inverse association ($P < 0.001$). All three analyses reached statistical significance.

Table 5
Results of Three Correlation Analyses

Analysis Method	Correlation Coefficient	Correlation Strength	Significance
Pearson Correlation	$r = -0.471$	Moderate negative	$P < 0.001$
Spearman Rank Correlation (primary)	$\rho = -0.501$	Moderate negative	$P < 0.001$
Kendall Rank Correlation	$\tau = -0.402$	Moderate negative	$P < 0.001$
Multivariable linear regression (adjusted for age, sex, BMI, PPI use)	$\beta = -0.42$	Independent negative	-

The LES pressure declined continuously as the Hill grade increased from grade I to grade IV. Given the skewed distribution of manometric data (Shapiro–Wilk test, $P < 0.05$), Spearman’s rank correlation was adopted as the primary outcome, and a scatter trend plot was generated accordingly. The figure illustrated a gradual decline in LES pressure with ascending Hill grades, demonstrating that higher Hill classifications corresponded to lower resting LES values (Fig. 9).

4. Discussion

In this study, we developed a ResNet-152–based deep learning model for automated GEFV Hill grading using 7,652 endoscopic images from a single center. The final model achieved an overall accuracy of 85.02% on the test set, with all global performance metrics (precision, recall, F1-score) exceeding 0.8. Compared with recently published AI models for Hill classification, our ResNet-152 framework demonstrated competitive performance, suggesting that deeper residual architectures may capture subtle mucosal morphological features more effectively for EGJ grading tasks^[24, 25].

Per-grade analysis revealed that classification accuracy was highest for Hill grade I (84.76%) and lowest for grade IV (72.38%). This gradient is largely attributable to class imbalance and to the morphological overlap between adjacent grades. Grade IV accounted for only 7.48% of the training dataset, while grades II and III exhibited the most substantial morphological overlap, which was also the most common misclassification pattern in the confusion matrix. Notably, the model outperformed junior endoscopists in grading efficiency, requiring only approximately 5 seconds per image, which supports its potential utility as a real-time adjunct during endoscopy^[26, 27].

A key strength of this study is the correlation validation against HRM, an objective functional gold standard for esophageal motility. We found a significant negative correlation between AI-predicted Hill grade and LES pressure. This magnitude of correlation is clinically meaningful. It confirms that the morphological features captured by the AI model correspond directly to anti-reflux barrier function, rather than merely learning superficial visual patterns. The stepwise decline in LES pressure across Hill grades

I–IV further validates the pathophysiological basis of the Hill classification system and supports its role as a non-invasive surrogate marker of EGJ competence^[28].

From a clinical decision-making perspective, this AI model could serve multiple roles along the GERD care pathway^[29]. For junior endoscopists, real-time AI grading could provide immediate feedback and reduce inter-observer variability during examinations. For treatment planning, automated Hill grading could help stratify patients by anti-reflux barrier competence—identifying those with higher-grade GEFV insufficiency who may benefit from earlier consideration of endoscopic or surgical anti-reflux therapy^[30, 31]. Furthermore, standardized AI grading could facilitate quality benchmarking across endoscopy centers and support large-scale epidemiological research on EGJ morphology.

However, manual Hill grading suffers from inherent inter-observer variability, with reported Kappa values ranging from 0.71 to 0.78 among experienced endoscopists^[17]. In our cohort, expert raters achieved a weighted Kappa of 0.73, consistent with published literature, confirming that subjectivity remains a fundamental limitation of the Hill classification system—a limitation that also imposes a theoretical ceiling on AI model performance. This study has several other limitations that should be acknowledged. First, as a single-center retrospective study with a convenience sampling design, the generalizability of the model to other populations and endoscopic systems remains uncertain. Distribution shifts arising from different equipment manufacturers, imaging angles, and staining protocols may further degrade performance in external settings^[13, 30]. Second, the sample size for Hill grade IV was small ($n = 572$ images, 8 patients in the HRM cohort), which limits statistical robustness for this subgroup. Third, the model was trained on static images only; temporal dynamics and real-time video grading were not evaluated. Fourth, verification bias is minimized as all included images underwent the same reference standard evaluation. Measurement bias is partially mitigated by the multi-stage annotation pipeline and blinded grading.

Future more advances in AI-assisted EGJ grading are anticipated. Beyond HRM parameters used in this work, additional morphological and functional markers can be integrated for comprehensive GERD patient evaluation to guide individualized therapeutic regimens^[32, 33]. Compared with the Hill classification, the 2022 American Foregut Society (AFS) grading system incorporates quantitative metrics of hiatal hernia length^[13, 34], hiatal diameter and flap valve function, yielding superior inter-observer reproducibility and serving as a more robust target for future AI models. Subsequent research directions include multimodal combined analysis, multicenter prospective validation, video-based dynamic endoscopic assessment, and head-to-head comparative analyses between AFS and Hill grading frameworks^[35].

5. Conclusion

In summary, this study demonstrates that a ResNet-152-based deep learning model achieves reliable automated classification of GEFV Hill grades, with an overall test-set accuracy of 85.02% and grading

performance comparable to that of experienced endoscopists. The significant correlation between AI-predicted grades and HRM-derived resting LES pressure provides objective pathophysiological validation for the model's predictions. The AI system also improved endoscopic quality control metrics, suggesting its potential as a training adjunct. Nevertheless, these preliminary findings are constrained by the single-center retrospective design and limited representation of high-grade lesions. External multicenter prospective validation is mandatory prior to any routine clinical adoption. Future work should also incorporate real-time video analysis, uncertainty quantification, and the newer AFS classification criteria to further enhance clinical utility.

Abbreviations

AI

Artificial intelligence

GERD

Gastroesophageal reflux disease

GEFV

Gastroesophageal flap valve

EGJ

Esophagogastric junction

HRM

High-resolution esophageal manometry

LES

Lower esophageal sphincter

Declarations

Ethics approval and consent to participate

The present retrospective single-center study was approved by the Ethics Committee of Suining Central Hospital (approval No. KYLLKS20260087). The requirement for written informed consent was waived by the ethics committee owing to the retrospective nature of the study. All procedures were performed in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable. All endoscopic images included in this study were fully deidentified without any individual patient identifiable information. No identifiable personal data was published.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or non-profit sectors.

Author Contribution

T.L. (Tianyu Liu) and R.C. (Rui Cai) conceived and designed the study. R.C. and X.Y. (Xiaoqin Yang) performed data curation. C.T. (Chuangdong Tang) developed the software. R.C. and M.Y. (Minyuan Yang) conducted formal analysis and prepared the visualizations. R.C. and Y.Q. (Yang Qin) wrote the original draft of the manuscript. T.L. performed manuscript review and editing, and provided supervision. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

The raw endoscopic image dataset cannot be publicly released due to hospital patient-privacy restrictions. De-identified aggregated numerical data generated and analysed during this study are available from the corresponding author upon reasonable written request. The python code for ResNet-152 model training, Grad-CAM visualization and statistical analysis is available at available from corresponding author on reasonable request.

References

1. Gyawali CP, Yadlapati R, Fass R, Katzka D, Pandolfino J, Savarino E, et al. Updates to the modern diagnosis of GERD: Lyon consensus 2.0. *Gut*. 2024;73(2):361–71. 10.1136/gutjnl-2023-330616.
2. El-Serag HB, Sweet S, Winchester CC, Dent J. Update on the epidemiology of gastro-oesophageal reflux disease: a systematic review. *Gut*. 2014;63(6):871–80. 10.1136/gutjnl-2012. PMID:23853213.
3. Wickramasinghe N, Devanarayana NM. Insight into global burden of gastroesophageal reflux disease: understanding its reach and impact. *World J Gastrointest Pharmacol Ther*. 2025;16(1):97918. 10.4292/wjgpt.v16.i1.97918.
4. Guo Z, Wu Y, Zhan Y, Zhang C. Correlation between gastroesophageal flap valve abnormality and novel parameters in patients with gastroesophageal reflux disease symptoms by the lyon consensus. *Sci Rep*. 2021;11(1):15076. 10.1038/s41598-021.
5. Cable J, Giuli R, Fox MR, Rosen RL, Valdovinos-García LR, Siboni S, et al. Gatherings in esophagology: innovations and future directions in the diagnosis and management of reflux

- disease. *Ann N Y Acad Sci.* 2026;1557(1):e70225. 10.1111/nyas.70225.
6. Vespa E, Barchi A, Fanizza J, Barà R, Fasulo E, Bencardino S, et al. Mean basal esophagogastric junction pressure is a better high-resolution manometry predictor of gastroesophageal reflux disease than esophagogastric junction contractile integral. *J Neurogastroenterol Motil.* 2026;32(1):35–45. 10.5056/jnm25029.
 7. Brännström L, Werner M, Wallner B, Franklin KA, Karling P. What is the significance of the Hill classification? *Dis Esophagus.* 2023;36(9):doad004. 10.1093/dote/doad004.
 8. Visaggi P, Barberio B, Gregori D, Azzolina D, Martinato M, Hassan C, et al. Systematic review with meta-analysis: artificial intelligence in the diagnosis of oesophageal diseases. *Aliment Pharmacol Ther.* 2022;55(5):528–40. 10.1111/apt.16778.
 9. Popovic D, Glisic T, Milosavljevic T, Panic N, Marjanovic-Haljilji M, Mijac D, et al. The importance of artificial intelligence in upper gastrointestinal endoscopy. *Diagnostics (Basel).* 2023;13(18):2862. 10.3390/diagnostics13182862.
 10. Quek SXZ, Ho KY. Artificial intelligence in upper gastrointestinal diagnosis. *Korean J Helicobacter Up Gastrointest Res.* 2025;25(3):251–60. 10.7704/kjhugr.2025.0024.
 11. Abe S, Kitagawa Y, Hatta W, Maekita T, Kato M, Nagahara A, et al. A multicenter pivotal study on the artificial intelligence system for neoplastic lesions detection in upper gastrointestinal endoscopy. *Dig Endosc.* 2025;37(12):1306–14. 10.1111/den.70015.
 12. Hill LD, Kozarek RA, Kraemer SJ, Aye RW, Mercer CD, Low DE, Pope CE 2. The gastroesophageal flap valve: in vitro and in vivo observations. *Gastrointest Endosc.* 1996;44(5):541–7. 10.1016/s0016-5107(96)70006-8.
 13. Nguyen NT, Thosani NC, Canto MI, Chang K, Lipham J, Abu Dayyeh B, et al. The American Foregut Society white paper on the endoscopic classification of the esophagogastric junction integrity. *Foregut.* 2022;2(4):241–52. 10.1177/26345161221126961.
 14. Hu Y, Xu J, Huang L, Zheng Z, Zhao J, Chen T, et al. Artificial intelligence-assisted endoscopic diagnosis system for diagnosing *Helicobacter pylori* infection: a multicenter study. *BMC Med.* 2025;23(1):540. 10.1186/s12916-025-04379-2.
 15. Wang CC, Chiu YC, Chen WL, Yang TW, Tsai MC, Tseng MH. A deep learning model for classification of endoscopic gastroesophageal reflux disease. *Int J Environ Res Public Health.* 2021;18(5):2428. 10.3390/ijerph18052428.
 16. Kommuru S, Adekunle F, Niño S, Arefin S, Thalvayapati SP, Kuriakose D, et al. Role of artificial intelligence in the diagnosis of gastroesophageal reflux disease. *Cureus.* 2024;16(6):e62206. 10.7759/cureus.62206.
 17. Ge Z, Wang B, Chang J, Yu Z, Zhou Z, Zhang J, et al. Using deep learning and explainable artificial intelligence to assess the severity of gastroesophageal reflux disease according to the Los Angeles classification system. *Scand J Gastroenterol.* 2023 Jun;58(6):596–604. 10.1080/00365521.2022.2163185.

18. Yen HH, Tsai HY, Wang CC, Tsai MC, Tseng MH. An improved end-to-end deep learning model for classification of endoscopic gastroesophageal reflux disease using deep learning integrated machine learning. *Diagnostics (Basel)*. 2022;12(11):2827. 10.3390/diagnostics1211-2827.
19. Kafetzis I, Sodmann P, Herghelegiu BE, Brand M, Zoller WG, Seyfried F, et al. Prospective evaluation of real-time artificial intelligence for the Hill classification of the gastroesophageal junction. *United Eur Gastroenterol J*. 2025;13(2):240–6. 10.1002/ueg2.12721.
20. Chen J, Wang G, Xia K, Wang Z, Liu L, Xu X. Constructing an artificial intelligence-assisted system for the assessment of gastroesophageal valve function based on the hill classification (with video). *BMC Med Inf Decis Mak*. 2025;25(1):144. 10.1186/s12911-025-02973-1.
21. Kafetzis I, Fuchs KH, Sodmann P, Troya J, Zoller W, Meining A, et al. Efficient artificial intelligence-based assessment of the gastroesophageal valve with Hill classification through active learning. *Sci Rep*. 2024;14(1):18825. 10.1038/s41598-.
22. Jamil M, Li Q. Artificial intelligence in gastrointestinal endoscopy: current evidence and future directions. *Ther Adv Gastrointest Endosc*. 2025;18:26317745251398945. 10.1177/26317745251398945.
23. Niikura R, Aoki T, Shichijo S, Yamada A, Kawahara T, Kato Y, et al. Artificial intelligence versus expert endoscopists for diagnosis of gastric cancer in patients who have undergone upper gastrointestinal endoscopy. *Endoscopy*. 2022;54(8):780–4. 10.1055/a-1660.
24. Osman A, Albashir MM, Nandipati K, Walters RW, Chandra S. Esophagogastric junction morphology on Hill's classification predicts gastroesophageal reflux with good accuracy and consistency. *Dig Dis Sci*. 2021;66(1):151–9. 10.1007/s10620-020-.
25. Tokat M, van Tilburg L, Koch AD, Spaander MCW. Artificial intelligence in upper gastrointestinal endoscopy. *Dig Dis*. 2022;40(4):395–408. 10.1159/000518232.
26. Massironi S. Advancements in Barrett's esophagus detection: the role of artificial intelligence and its implications. *World J Gastroenterol*. 2024;30(11):1494–6. 10.3748/wjg.v30.i11.1494.
27. Guo F, Meng H. Application of artificial intelligence in gastrointestinal endoscopy. *Arab J Gastroenterol*. 2024;25(2):93–6. 10.1016/j.ajg.2023.12.010.
28. Li XR, Kong MW, Guan XF, Gao Y. Revolutionizing upper gastrointestinal disease diagnosis: the transformative role of artificial intelligence in endoscopy. *World J Gastrointest Endosc*. 2025;17(7):108293. 10.4253/wjge.v17.i7.108293.
29. Meng QQ, Gao Y, Lin H, Wang TJ, Zhang YR, Feng J, et al. Application of an artificial intelligence system for endoscopic diagnosis of superficial esophageal squamous cell carcinoma. *World J Gastroenterol*. 2022;28(37):5483–93. 10.3748/wjg.v28.i37.5483.
30. Gong EJ, Woo J, Lee JJ, Bang CS. Role of artificial intelligence in gastric diseases. *World J Gastroenterol*. 2025;31(37):111327. 10.3748/wjg.v31.i37.111327.
31. Adams MA, Feld A. Ethical and legal considerations for responsible use of artificial intelligence in clinical gastroenterology. *Am J Gastroenterol*. 2026;121(4):855–62. 10.14309/ajg.0000000000003897.

32. Chue KM, Goh DWX, Chua CME, Toh BC, Ong LWL, Wong WK, et al. The Hill's classification is useful to predict the development of postoperative gastroesophageal reflux disease and erosive esophagitis after laparoscopic sleeve gastrectomy. *J Gastrointest Surg.* 2022;26(6):1162–70. 10.1007/s11605-022-05324-x.
33. Alvarez R, Ward BL, Xiao T, Zadeh J, Sarode A, Khaitan L, et al. Independent association of preoperative Hill grade with gastroesophageal reflux disease 2 years after sleeve gastrectomy. *Surg Obes Relat Dis.* 2023;19(6):563–75. 10.1016/j.soard.2022.12.013.
34. Siboni S, Sozzi M, Lovece A, Milito P, Visaggi P, Ghisa M, et al. The AFS endoscopic classification of esophago–gastric junction integrity is superior to the Hill classification in terms of interobserver variability. *J Clin Gastroenterol.* 2025 Nov. 10.1097/MCG.0000000000002288.
35. Maity R, Raja Sankari VM, Salvador USNAR. Explainable AI–based automated segmentation and multi–stage classification of gastroesophageal reflux using machine learning techniques. *Biomed Phys Eng Express.* 2024;10(4). 10.1088/2057-1976/ad5a14.

Figures

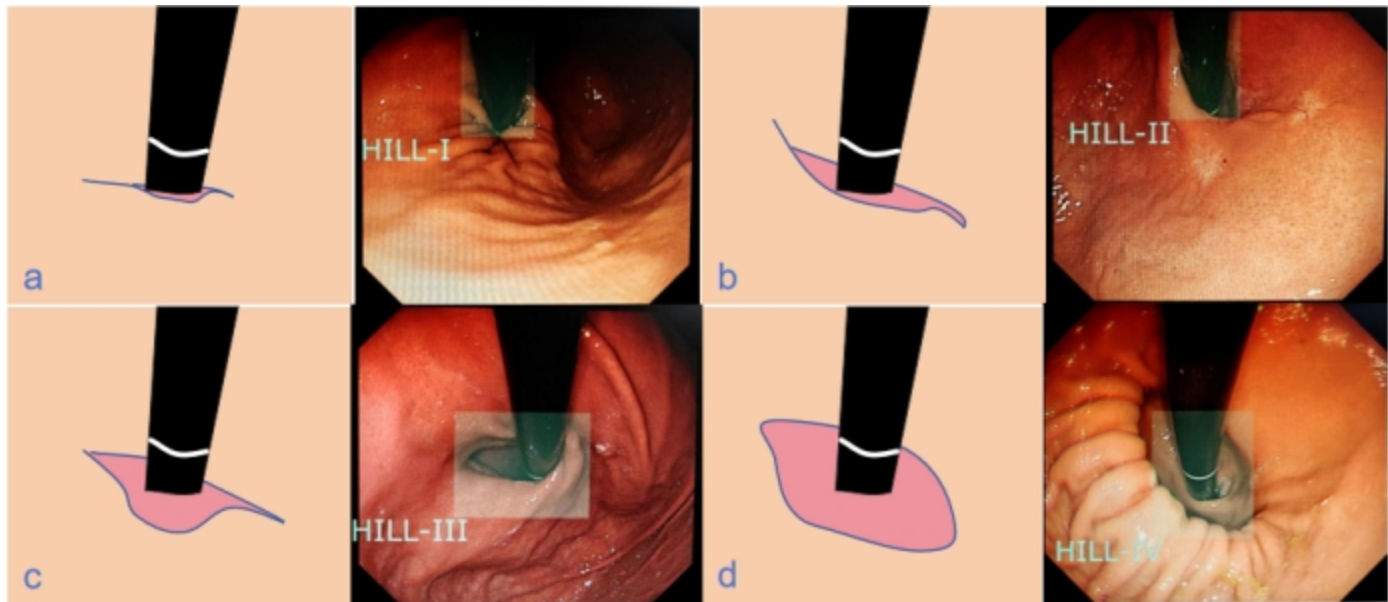


Figure 1

Endoscopic reference images of Hill grading for GEFV (original drawings by the authors).

a. Hill grade I: Folds tightly encircle the endoscope. **b.** Hill grade II: Folds are lax with intermittent opening and closing. **c.** Hill grade III: Folds are indistinct. **d.** Hill grade IV: Folds are absent, accompanied by hiatal hernia.

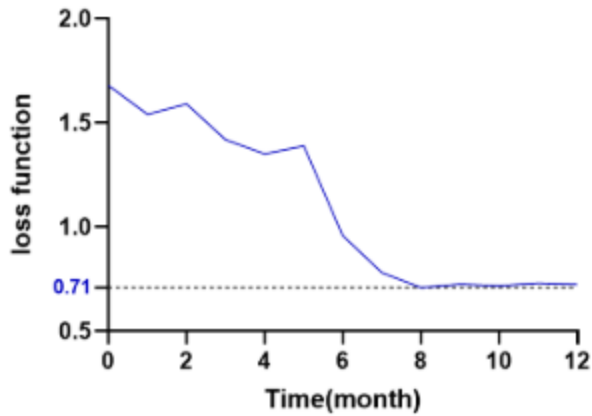


Figure 4

Change in the average loss function on the test set.

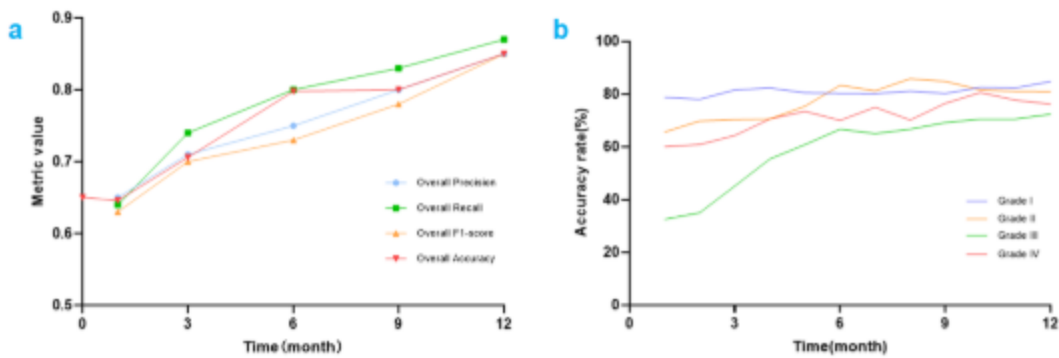


Figure 5

Dynamic trends of model performance throughout iterative optimization.

- a.** Trends in overall precision, recall, F1-score, and accuracy across 12 months of model optimization. **b.** Per-grade classification accuracy for Hill grades I-IV at each optimization stage.

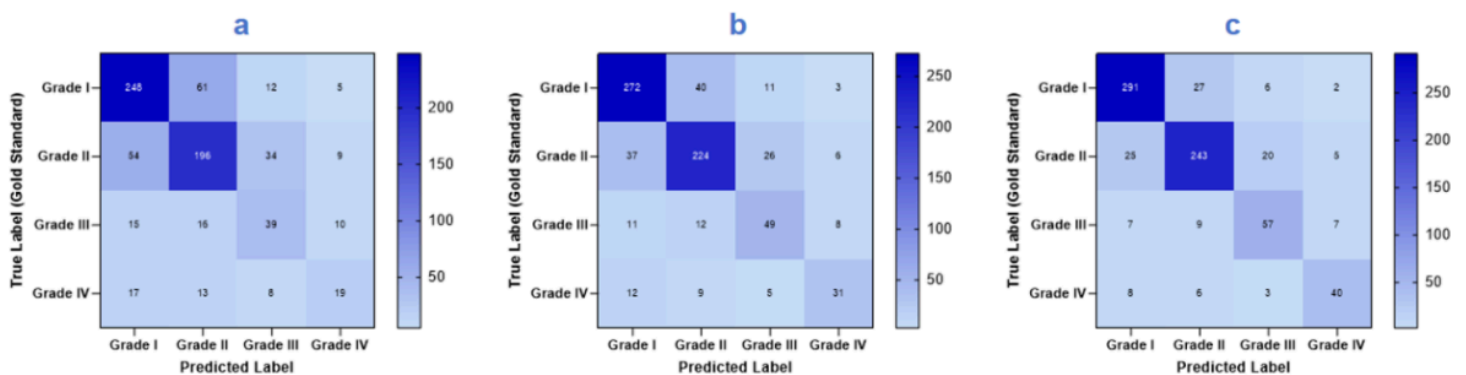


Figure 6

Confusion Matrix Heatmap.

a. Early training stage. **b.** Mid-training stage. **c.** Late training stage. (Note: Numbers may not sum exactly to the total test set size ($n = 756$) due to rounding in the heatmap visualization.)

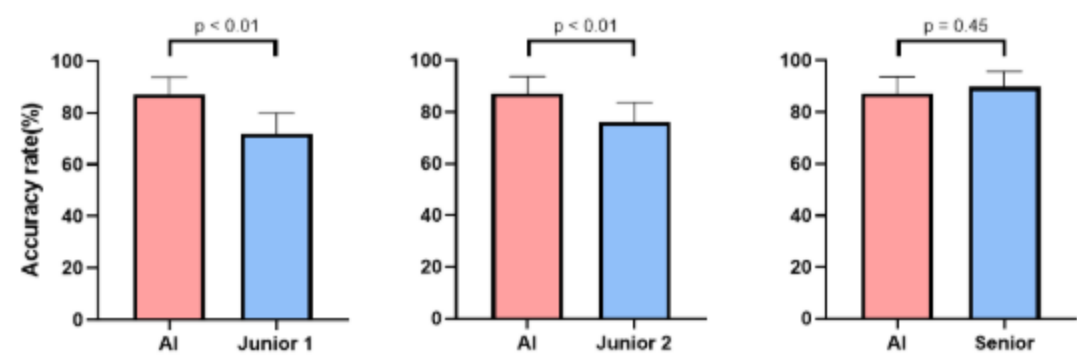


Figure 7

Comparison of accuracy between AI and endoscopist assessments for Hill grading on a random subset.

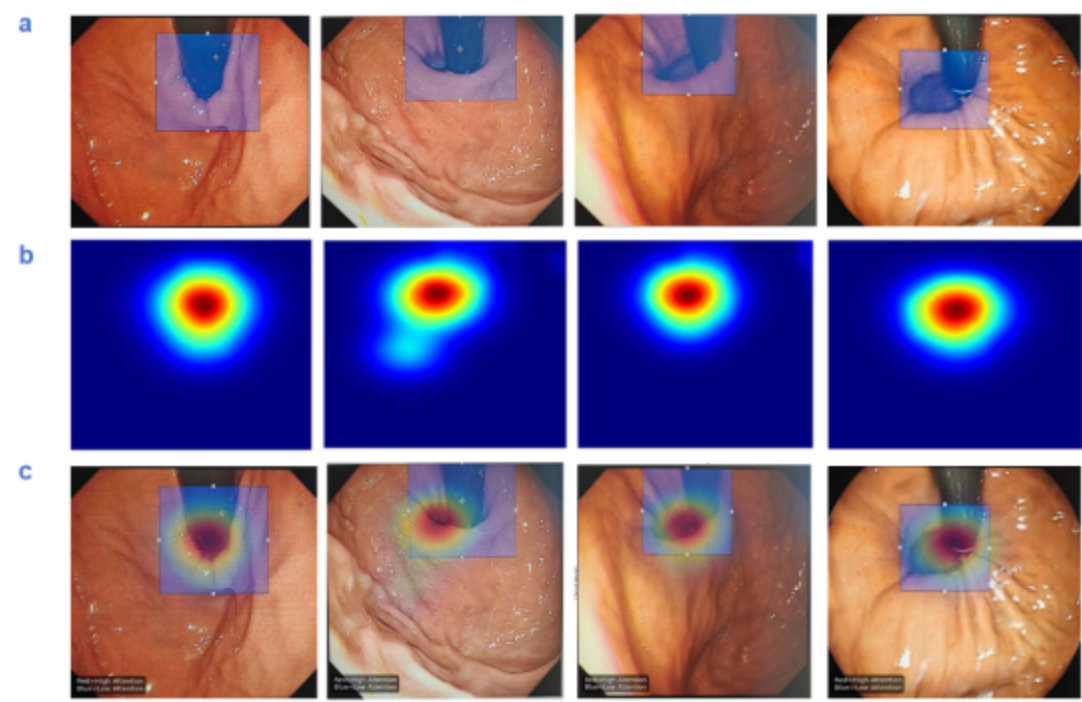


Figure 8

Interpretability of the ResNet-152 Hill grading model.

a. The original endoscopic images. **b.** Pixel-wise activation heatmaps generated using the Grad-CAM technique. **c.** Overlay images combining the original endoscopy with the activation heatmaps. Warmer colors (red) indicate higher contribution to the predicted class.

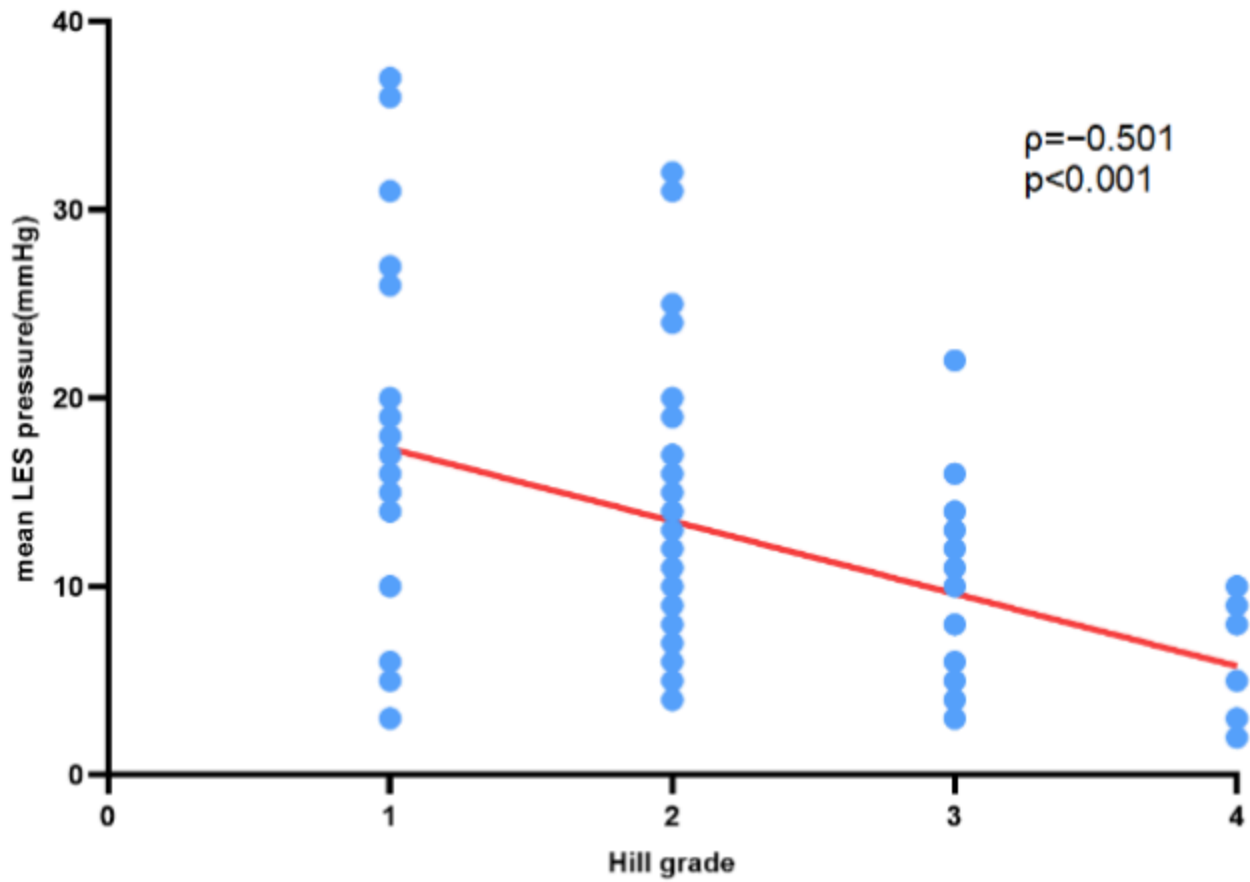


Figure 9

Scatter Plot with Trend Line: mean resting LES pressure decreased significantly with increasing Hill grade. Spearman correlation showed a moderate negative correlation ($\rho = -0.501$, $P < 0.001$) between LES pressure and Hill grade.