ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Application of an Intelligent Analysis System to Therapeutic Efficacy Evaluation of Wenyujin in Precancerous Gastric Lesions

Kehan Zhang[1] *and Haifeng Jin*[1,2,*]

[1] Zhejiang Chinese Medical University, Hangzhou, Zhejiang, 310053, China
[2] The First Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, Zhejiang, 310006, China

SUMMARY: *Aiming at the problems that only depending on pathological conclusions to carry out efficacy explanation after intervention for gastric precancerous lesions, endoscopic descriptions do not keep consistent, and multi-source information is hard to read in unified way, this paper builds a multimodal intelligent analysis system which is made specially for the intervention situation of Curcuma wenyujin. This system is designed for the identification of efficacy responses at the patient level and the assessment of stage migration. This research has in it 228 patients whose gastric precancerous lesions got pathologically proved, there are 114 cases in the Curcuma wenyujin intervention group and 114 cases in the control group. Longitudinal following data on Baseline, Month 3, and Month 6 were arranged to make a structured sample which includes 5,472 endoscopic images, 1,824 pathological site labels, and 26 clinical serum variables. Methodologically, the system consists of an endoscopic image branch, a pathological branch, and a clinical-serum branch. It achieves patient-level feature alignment through an attention-weighted cross-modal fusion mechanism and simultaneously outputs efficacy response results and OLGA/OLGIM stage migration results. The results show that the response rate, pathological remission rate, and OLGIM/OLGA downstage ratio in the Curcuma wenyujin group are higher than those in the control group at 6 months; specifically, the response rate is 58.8% compared to 37.7% in the control group, and the pathological remission rates are 53.5% and 31.6%, respectively. Additionally, the Curcuma wenyujin group exhibits more consistent improvement directions in indicators such as pathological load, mucosal load, PGI, PGR, G-17, IL-6, and TNF-α. Model comparison results indicate that the multimodal system achieves in the patient-level effect identification, we have obtained an AUC value 0.923, an F1-score value 0.854, and a Brier score value 0.108, hence it has better performance than the control models which only use clinical data, only use endoscopy data, only use pathology data, and dual-modal control models. The ablation experiment outcomes further make known that the pathological branch possesses the strongest pulling force for patient-level explanation, the endoscopic branch offers important shape-related increases, and the clinical-serum branch greatly enhances probability adjustment and subgroup stability. The analysis of errors indicates that the errors of the model are mainly concentrated in the situations of boundary stage transfer, multiple focal lesions, low-quality endoscopic images, and situations where pathological changes and endoscopic changes are not synchronous. The outcomes of this research prove that the system has the ability to integrate pathological alterations, mucosal appearance characteristics, and clinical serum targets into one united assessment frame, hence allowing combined explanation on the patient level after the intervention is done. This approach can be utilized to support review arrangements and risk reassessment, and provides a scalable methodological foundation for the intelligent efficacy*

*jinhaifeng0908@163.com
https://doi.org/10.65102/is20261069
Published: 10 June, 2026

evaluation of traditional Chinese medicine interventions in precancerous lesions of the stomach.

KEYWORDS: *Wenyujin; Precancerous Gastric Lesions; Multimodal Intelligent Analysis; Therapeutic Efficacy Evaluation; Risk Stratification*

1 Introduction

In the face of the problems that only depend on pathological results to assess the effect of interventions for gastric precancerous lesions, endoscopic descriptions are not consistent, and the difficulty in uniformly interpreting multi-source information, gastric precancerous lesions are located in a critical interval where chronic damage to the gastric mucosa progresses towards malignant transformation. Clinically, common conditions such as atrophy, intestinal metaplasia, and low-grade intraepithelial neoplasia do not correspond to the same rate of progression, but all directly affect follow-up frequency, biopsy strategy, and intervention timing. The latest MAPS III guidelines have placed high-quality endoscopy, virtual staining, zoned biopsy, and risk stratification based on OLGA/OLGIM or endoscopic grading lies in the core of management, therefore it shows that the identification and stratification of this stage have been shifted from empirical judgment to standardized and traceable evidence-based organization [1]. From the perspective of actual diagnostic and treatment scenarios, there is still a significant gap in the efficacy evaluation of gastric precancerous lesions. Pathological results provide a basis for grading, but their conclusions are influenced by sampling site, lesion heterogeneity, and observer variability, making it difficult to fully reflect the continuous changes of gastric mucosa over time. Endoscopy can provide more intuitive information on surface structure and vascular texture, but image interpretation is highly dependent on operator experience, and the descriptive standards vary between different centers. Recent reviews on the management of gastric intestinal metaplasia point out that although current follow-up strategies have gradually become more unified, clinical practice at present still confronts problems including inconsistent layering standards, sampling depth, and risk threshold values, thus it is difficult to depend on a single index to judge "whether improvement is needed, to what degree the improvement needs to reach, and when to adjust strategies."[2, 3]. In specific practice, the new art form semantics and performance style approach how to apply still need to be more perfect and iterative, in which most of the designers are affected by the traditional concept of the speed of innovation and development is slow, the effect of the degree of creativity is mainly dependent on the degree of their own understanding of the way to get transformed through communication between designers, brainstorming, etc., the existence of the works of the degree of creativity and lack of expressive power, design efficiency and other low Problems[4]. Endoscopic studies have also demonstrated that artificial intelligence systems can identify atrophy and perform risk stratification under white light endoscopy [5]. Furthermore, systematic reviews and meta-analyses of AI-assisted endoscopy in the identification of gastric intestinal metaplasia indicate that this approach has achieved high diagnostic performance, but it primarily serves diagnostic scenarios rather than assessing post-intervention efficacy.[6]Machine learning research aimed at high-risk screening for gastric precancerous lesions and stratified prediction using ME-NBI has also advanced the issue to the level of high-risk identification and staging assessment[7, 8].

In contrast, research on pharmacological intervention for gastric precancerous lesions is accumulating more detailed clinical evidence. Randomized controlled trials have shown that traditional Chinese medicine compounds can be used to reverse gastric precancerous lesions, especially with clinical significance in improving dysplasia.[9]. From the perspective of pharmacological basis, studies on Curcuma wenyujin, which belongs to Curcumae Rhizoma,

suggest that its active components can affect cell proliferation, migration, apoptosis, autophagy, and tumor-related microenvironment; while recent reviews on chronic atrophic gastritis also indicate that evidence for traditional Chinese medicine intervention has gradually shifted from empirical efficacy observations to the collation of mechanisms and targets[10, 11].

In the work that already exists, there are still many blank spaces that have not been filled. First of all, the already published research works have for the most part put emphasis on single-modality input data, among which pathology, endoscopy and clinical indicators are usually modeled each by itself, hence making it hard to make patient-level information align inside the identical framework; The single one data source already cannot anymore satisfy the support for stable judgment that is at patient level. Secondly, many models make use of static cross-sectional data, which are more good at evaluating the present state, but their explanation ability for how lesion burden moves before and after intervention still has limitation. Thirdly, the researches about the intervention of gastric precancerous lesions by Curcuma wenyujin mainly put focus on effect observation and mechanism deduction, hence they lack an intelligent analysis framework which can together explain endoscopic phenotypes, pathological changes and clinical indicators in the same coordinate system.

According to the above-mentioned gaps, this thesis builds a multimodal intelligent analysis system that takes the efficacy assessment of Curcuma wenyujin in intervening gastric precancerous lesions as the center. It makes the combination of endoscopic images, pathological labels and clinical serum indexes on the patient level, therefore it facilitates the explanation of curative effect on patient level, the judgment of lesion stage transfer and the tracking of key characteristics. The core stand of this thesis is not placed in the repeated verification of which direction a single index changes toward, but is placed in the construction of a quantifiable, examinable, and explicable effect assessment framework. This framework lets the intervention effect form an evidence chain which mutually confirms among histology, mucosal phenotype, and clinical characteristics, and thus can be utilized to assist review arrangements and intervention adjustments.

2 Methods

2.1 Cohort Definition, Wenyujin Intervention, and Multisource Data Acquisition

The main analysis cohort was set to include 228 patients with gastric precancerous lesions confirmed by gastroscopy and biopsy pathology, with 114 cases in the intervention group of Wenyu Jin and 114 cases in the control group. The inclusion criteria included chronic atrophy, intestinal metaplasia, and low-grade intraepithelial neoplasia; cases with previous history of gastric cancer or gastric resection, those who had received eradication therapy or other strong intervention schemes within the past 3 months, severe liver and kidney dysfunction, missing follow-up data, and cases where key images or pathological sites could not be aligned were excluded. All cases were organized based on the patient as the smallest analysis unit, with three observation nodes set at Baseline, Month 3, and Month 6 to ensure that endoscopic, pathological, and clinical indicators before and after intervention fell within the same time frame. The biopsy sampling and risk stratification criteria were implemented in accordance with the updated recommendations of MAPS III for high-quality gastroscopy, zoned sampling, and OLGA/OLGIM stratification[1].

The group of Wenyu Jin accepted the intervention from Wenyu Jin on the foundation of routine treatment, the period of treatment is set as 6 months, and follow-up checks were carried out at the 3rd Month and the 6th Month. For the guarantee that the reproducibility of treatment

response labels can be achieved, this study did not take a single "overall response rate" to be the primary judging standard. Instead of that, pathological changes, endoscopic mucosal performance manifestations, and serum gastric function index items were together taken as entry points for effect evaluation. To speak concretely, the endpoints at patient level are made of four parts: firstly, a reduction of pathological load, including atrophy, intestinal metaplasia, and decreased low-grade intraepithelial neoplasia; secondly, a lowering in OLGA/OLGIM staging or keeping at a low-risk grade; third, a lowering of the overall score of endoscope mucous membrane texture abnormality, blood vessel exposure, and partial irregular zones; Fourth, the ascension of PGI, PGR, G-17, and inflammation-related index indicators. When no less than three of the above-mentioned dimensions attained the pre-set improvement threshold, this situation is defined as a response, hence the other situation is recorded as non-response. This setting let post-treatment changes not rely on single-point pathological conclusions any more, hence turned them into traceable labels on patient level.

Multi-source data collection revolves around the principle of "same patient, same time point, same labeling system". For the endoscopy part, qualified images from white light gastroscopy and NBI/magnifying endoscopy were retained, with a total of 5,472 images set for modeling analysis. Before image inclusion, de-identification, blurry frame removal, duplicate frame removal, and site verification were completed, and site labeling was performed according to the distribution of gastric antrum, gastric body, gastric angle, and multiple foci. For the pathology part, it corresponds to four-point biopsy at Baseline and Month 6, resulting in a total of 1,824 pathological site labels. All slides were reviewed by two gastrointestinal pathologists and uniformly mapped to atrophy grade, intestinal metaplasia grade, low-grade intraepithelial neoplasia, OLGA stage, and OLGIM stage. For the clinical part, 26 variables including age, gender, H. pylori status, smoking and drinking habits, symptom score, as well as PGI, PGII, PGR, G-17, CRP, IL-6, TNF-α, etc., were collected. After completing the aforementioned processing, all information was vertically merged based on patient-level primary keys to form structured samples that can be directly used for subsequent multimodal modeling.

In order to make the feature extraction of visual and pathological encoders keep stable, and to reduce the constraints that the small sample size from one single center brings to the initial representation capability of the model, this research has utilized public accessible auxiliary data that are outside the main analysis cohort for pre-training, structure adjustment, and external shape reference. On the endoscope observing aspect, HyperKvasir has been chosen as a general gastrointestinal endoscope vision pre-training resource, and GastroHUN is added as gastroscopy assistance data which is customized for systematic stomach examination workflows. On the pathological aspect, a fully marked pathological slide data set of early gastric cancer and precancerous lesions, which was publicly issued in 2025, was brought in to supply morphological reference materials for atrophy, intestinal metaplasia, and the boundary between precancerous and early cancer. Public data was only employed for model initialization, morphological prior restriction, and external structure comparison, and did not take part in the calculation of treatment effects in the main analysis cohort, hence avoiding other explanations for the intervention outcomes of Curcuma aromatica that are caused by public sample distribution[12-14]. For guaranteeing the correspondence among intervention objects, follow-up points and multi-source labels, this research firstly finished case inclusion, sample rebuilding and patient-level data matching, which is shown in Fig 1.

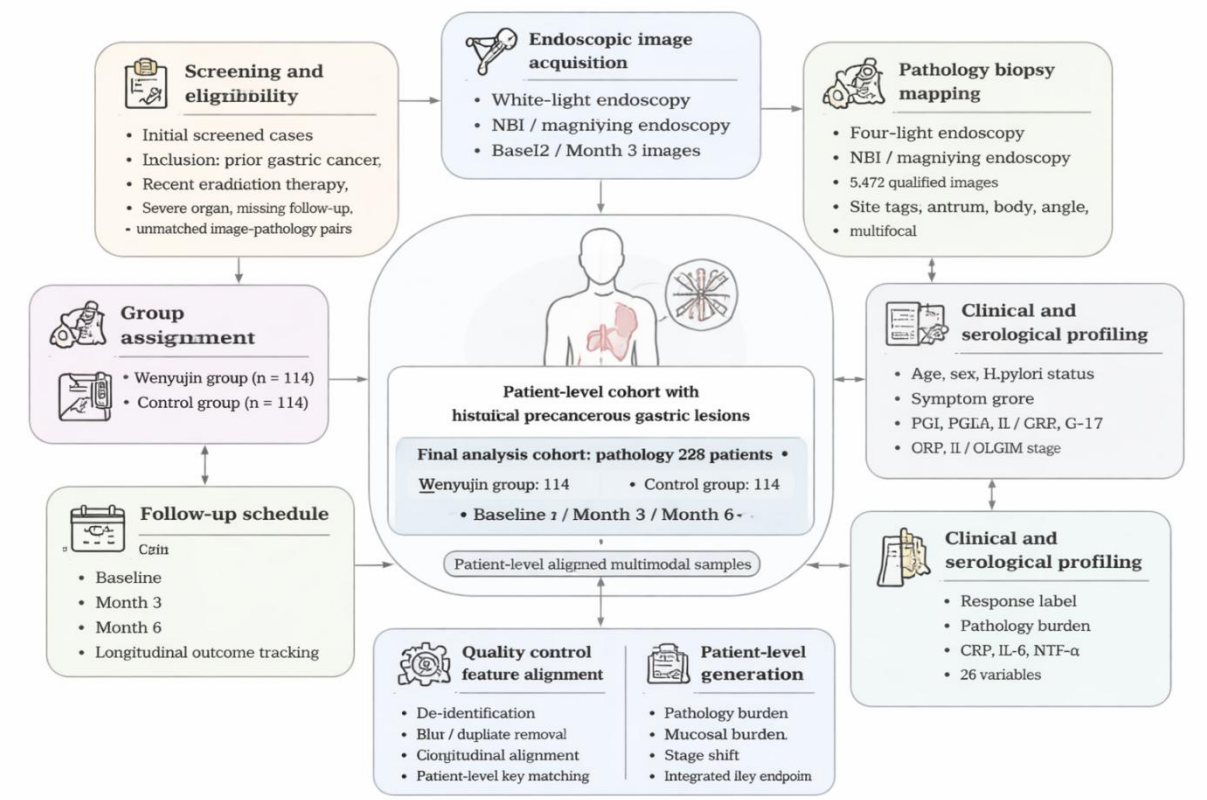

Figure 1: Data organization and cohort construction for therapeutic efficacy evaluation of Wenyujin in precancerous gastric lesions.

2.2 Construction of the Multimodal Intelligent Analysis System

For this study, the efficacy evaluation task is defined as a patient-level multimodal discrimination problem, with inputs consisting of three types of information: the endoscopic image branch receives white light gastroscopy and NBI/magnifying endoscopy images at Baseline, Month 3, and Month 6; the pathological branch receives site-level pathological labels and stage information corresponding to follow-up nodes; the clinical-serological branch receives age, gender, H. pylori status, lesion location, symptom scores, as well as gastric function and inflammation indicators. The reason for adopting a three-branch structure is that recent studies on gastric precancerous lesions have demonstrated that pathological images can stably support the grading of atrophy and intestinal metaplasia, white light gastroscopy can support the identification and risk stratification of atrophy, while clinical and serological variables can be used to supplement the risk assessment of gastric precancerous lesions or intestinal metaplasia. Organizing the three types of information side by side is more suitable for addressing the stage migration and efficacy response issues after intervention with Wenyu Jin [4, 5, 7, 8].

The main body of the model is constituted by one representation encoding layer, one cross-modal fusion layer, and one dual-task output layer. The branch of endoscopic image uses a visual encoder which has shared parameters to extract high-dimensional expressions of multi-time point mucous membrane texture, blood vessel exposure, gland edge, and focal abnormal regions. Multiple images that belong to the same patient at different time points are firstly got together inside the branch, and then are summed up into patient-level visual features. The pathology department uses position-level marks as the basic unit, maps atrophy level, intestinal metaplasia level, low-grade epithelial neoplasia, OLGA stage, and OLGIM stage to structured expressions, and therefore produces patient-level tissue characteristics through position-

weighted summation. The clinical-serological branch carries out standardization for continuous variables, carries out embedding for categorical variables, and produces low-dimensional clinical feature vectors by means of a fully connected representation layer. The three branch structures keep their own information density and time-related properties, thus avoiding the covering of weak signals by strong modalities when direct connection is done at the input side.

To enable the three types of features to participate in efficacy decision-making within the same discriminant space, this study introduces an attention-weighted cross-modal alignment mechanism at the fusion layer. Assuming that the endoscopic, pathological, and clinical features of patient i are denoted as $\mathbf{z}_i^{(e)}$, $\mathbf{z}_i^{(p)}$, and $\mathbf{z}_i^{(c)}$, respectively, the fused representation hi is defined as shown in formula (1).

$$\mathbf{h}_i = \alpha_i^{(e)}\mathbf{z}_i^{(e)} + \alpha_i^{(p)}\mathbf{z}_i^{(p)} + \alpha_i^{(c)}\mathbf{z}_i^{(c)} \tag{1a}$$

$$\alpha_i^{(m)} = \frac{exp(s_i^{(m)})}{\sum_{m\in\{e,p,c\}} exp\,(s_i^{(m)})} \tag{1b}$$

where $\alpha i(m)$ represents the adaptive weight of the m-th modality on patient i, and $si(m)$ denotes the confidence score output by the gated network for the corresponding modality. Specifically, $m \in e,p,c$ corresponds to the endoscopic, pathological, and clinical-serological branches, respectively. This design does not simply stack features but enables the model to dynamically adjust the participation intensity of each modality based on lesion morphology clarity, pathological label completeness, and clinical indicator stability.

At the output end, a therapeutic response head and a stage transition head are formed. The former is used to determine whether the patient achieves a response by Month 6, while the latter is used to predict whether the OLGA/OLGIM stage decreases, remains unchanged, or increases. Let $\hat{y}_i$ be the predicted probability of therapeutic response and $\hat{\mathbf{g}}_i$ be the class probability vector of stage transition. The output format is shown in formula (2).

$$\hat{y}_i = \sigma(\mathbf{W}_r\mathbf{h}_i + b_r) \tag{2a}$$

$$\hat{\mathbf{g}}_i = \mathrm{Softmax}(\mathbf{W}_s\mathbf{h}_i + \mathbf{b}_s) \tag{2b}$$

where $\sigma(\cdot)$ is the Sigmoid function, $\mathbf{W}_r$ and b_r are the efficacy response head parameters, and $\mathbf{W}_s$ and $\mathbf{b}_s$ are the stage transition head parameters.

To balance the two tasks of efficacy discrimination and stage transfer, this study adopts a weighted joint loss function to train the model, as shown in formula (3)

$$\mathcal{L} = \lambda_1\mathcal{L}_{\mathrm{resp}} + \lambda_2\mathcal{L}_{\mathrm{stage}} + \lambda_3\mathcal{L}_{\mathrm{cal}} \tag{3}$$

where $\mathcal{L}_{\mathrm{resp}}$ represents the binary cross-entropy loss for response/non-response, $\mathcal{L}_{\mathrm{stage}}$ denotes the multi-class cross-entropy loss for stage transfer tasks, $\mathcal{L}_{\mathrm{cal}}$ signifies the calibration constraint term based on temperature scaling of the validation set, and λ_1, λ_2, and λ_3 are the weight coefficients for the three loss terms.

On the level of explanation, this research keeps two explanation interfaces which are on the image side and the phenotype side. On the image aspect, Grad-CAM is utilized to trace back high-response regions inside endoscopic branches, marking regions of mucosal texture disturbance, focal edges, and abnormal vessel exposure that the model gives more attention to. On the phenotype aspect, SHAP is utilized to compute the marginal contribution values of clinical-serological variables for patient-level prediction outcomes, thus recognizing the

relative weight values of variables like PGI, PGR, G-17, IL-6, TNF-α, H. pylori's condition, and the position of lesion in the differentiation of treatment effect.

After completing the design of three-branch encoding, cross-modal fusion, and dual-task output, this study further developed a multimodal intelligent analysis system for efficacy recognition and risk output, as shown in Fig 2.

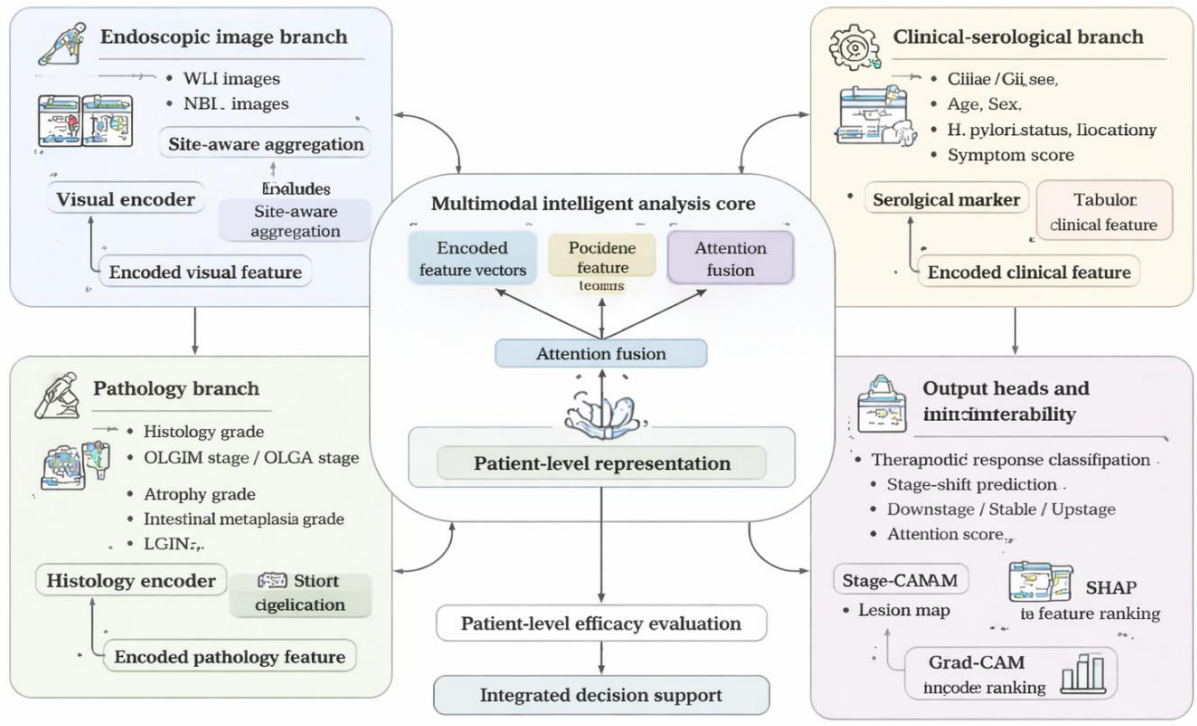

Figure 2: Architecture of the multimodal intelligent analysis system for efficacy evaluation

2.3 Experimental Protocol, Endpoint Definition, and Statistical Evaluation

This study completed data partitioning, endpoint reading, and performance reporting according to the preset protocol. Case partitioning, label interpretation, missing value handling, model comparison relationships, and result presentation criteria were all fixed before training to avoid reversely adjusting variable entry methods or evaluation indicators based on test results. The relevant reporting framework was implemented in accordance with the updated requirements of TRIPOD+AI for the development and validation of clinical predictive models, to ensure the reproducibility of sample organization, prediction objectives, and validation processes[15].

To assess the true performance of the model in the efficacy scenario, the experimental protocol was stratified at the patient level. All 228 cases were stratified and sampled based on the response/non-response ratio, Baseline OLGIM stage, and primary lesion site, and divided into training, validation, and test sets in a 6:2:2 ratio. Endoscopic images, pathological sites, and clinical records from the same patient were strictly locked in the same subset to avoid cross-set leakage. During the training phase, 5-fold cross-validation was implemented within the training set for hyperparameter search, modality weight adjustment, and early stopping judgment; the validation set was only used for model selection, temperature scaling, and threshold determination; the test set was opened once after all parameters were frozen and used to generate the final performance results. In addition to the multimodal main model, four control groups were simultaneously set up: clinical-only, endoscopy-only, pathology-only, and dual-modal.

The main research end point centers on "if the treatment effect reaches steady enhancement among all different methods". The main end point is the patient-level pathological remission mark at Month 6, which indicates that no less than two of atrophy, intestinal metaplasia, and low-grade intraepithelial neoplasia have lowered by one grade when compared with Baseline, and there is no rise in OLGIM or OLGA. The co-primary end point is the result of stage migration, which is parted into three sorts: downstage, stable, upstage, it is utilized to know the change of lesion stage. Secondary observation end points include the reduction of endoscopic mucous load score, the restoration of PGI and PGR, the lowering of G-17, the reduction of IL-6 and TNF-α, and the enhancement of overall symptom score. In order to evade the effect that one single indicator brings to endpoint judgment, this research maps the aforementioned information onto composite labels at the patient level and in a unified way records the variable origin, collection time point, coding method, and location inside the model; the mapping relationship between corresponding variable and endpoint is displayed in Table 2.

Table 2. Definition of multimodal variables and endpoint mapping used in the intelligent analysis system

Variable	Modality	Time point	Encoding	Model position	Endpoint
WLI image set	Endoscopy	Baseline, Month 3, Month 6	Image tensor + site tag	Endoscopic image branch	Mucosal burden, response
NBI / magnifying image set	Endoscopy	Baseline, Month 3, Month 6	Image tensor + site tag	Endoscopic image branch	Visual phenotype, response
Atrophy grade	Pathology	Baseline, Month 6	Ordinal encoding	Pathology branch	Remission, stage-shift
Intestinal metaplasia grade	Pathology	Baseline, Month 6	Ordinal encoding	Pathology branch	Remission, stage-shift
Low-grade intraepithelial neoplasia	Pathology	Baseline, Month 6	Binary / ordinal encoding	Pathology branch	Remission, stage-shift
OLGA stage	Pathology	Baseline, Month 6	Ordinal encoding	Pathology branch	Stage-shift
OLGIM stage	Pathology	Baseline, Month 6	Ordinal encoding	Pathology branch	Stage-shift
Age	Clinical	Baseline	Z-score normalization	Clinical-serological branch	Response adjustment
Sex	Clinical	Baseline	One-hot encoding	Clinical-serological branch	Subgroup analysis
H. pylori status	Clinical	Baseline, Month 6	Binary encoding	Clinical-serological branch	Response, subgroup robustness
Lesion location	Clinical/ Endoscopy	Baseline	One-hot encoding	Clinical-serological branch	Subgroup robustness
Symptom score	Clinical	Baseline, Month 3, Month 6	Z-score normalization	Clinical-serological branch	Response
PGI	Serology	Baseline, Month 3, Month 6	Z-score normalization	Clinical-serological branch	Response
PGII	Serology	Baseline, Month 3, Month 6	Z-score normalization	Clinical-serological branch	Supportive marker
PGR	Serology	Baseline, Month 3, Month 6	Ratio + normalization	Clinical-serological branch	Response
G-17	Serology	Baseline, Month 3, Month 6	Log transform + normalization	Clinical-serological branch	Response
CRP	Serology	Baseline, Month 3, Month 6	Log transform + normalization	Clinical-serological branch	Inflammatory change
IL-6	Serology	Baseline, Month 3, Month 6	Log transform + normalization	Clinical-serological branch	Inflammatory change, response
TNF-α	Serology	Baseline, Month 3, Month 6	Log transform + normalization	Clinical-serological branch	Inflammatory change, response
Composite pathology burden score	Derived	Baseline, Month 6	Weighted summation	Output label generation	Response, stage-shift
Endoscopic mucosal burden score	Derived	Baseline, Month 3, Month 6	Rule-based scoring	Output label generation	Response

Statistical evaluation is divided into four aspects: discriminatory performance, calibration performance, classification performance, and clinical net benefit. The primary indicators of the model are set to AUC and F1-score, which are used to read the overall discriminatory ability and the balanced performance of positive and negative classes, respectively. At the same time, sensitivity, specificity, accuracy, and precision are reported. To avoid representing all performance solely with AUC, this study further calculates the Brier score, calibration intercept, calibration slope, and draws a calibration curve. The clinical net benefit is read through decision curve analysis. The relevant evaluation protocol is implemented with reference to recent methodological studies on external validation of predictive models, selection of performance metrics, and model implementation and updating[16-18].

In the inter-group baseline comparison, continuous variables were first tested for normality. For those following a normal distribution, they were expressed as mean ± SD and tested using an independent samples t-test. For those not following a normal distribution, they were expressed as median (IQR) and tested using the Mann–Whitney U test. Categorical variables were tested using χ^2 test or Fisher's exact test. Repeated measures at three time points were processed using linear mixed effects models or generalized estimating equations, incorporating group, time, and their interaction terms. The 95% confidence intervals of model performance were calculated through 1,000 bootstrap simulations. Subgroup analyses were conducted based on H. pylori status, baseline OLGIM stage, lesion location, age stratification, and gender; error analyses were summarized around five categories: false positive, false negative, insufficient image quality, unclear pathological boundaries, and cross-modal information conflict. To verify the system's discriminatory ability, stability, and deployability in the efficacy evaluation scenario, this study further established a unified comparison and evaluation protocol, as shown in Figure 3.

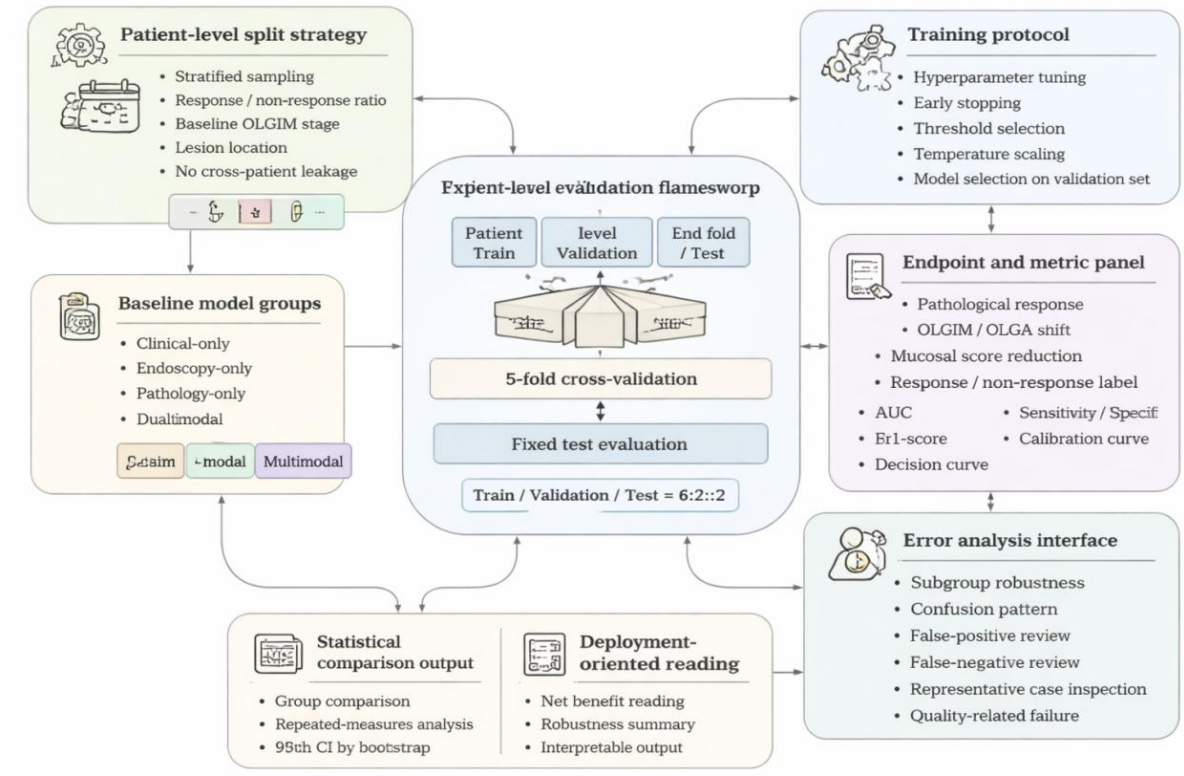

Figure 3: Experimental protocol, comparison strategy, and evaluation framework

3 Results and Discussion

3.1 Baseline Characteristics and Response of Wenyujin in Precancerous Gastric Lesions

This present section mainly talks about two questions: first, whether the two groups of patients have comparability before they enter the intervention; secondly, whether after the intervention using Wenyu Jin, there existed a simultaneous enhancement in pathology, endoscopy, and serum related indicators. The comparability of baseline conditions is what decides whether later differences can be attributed to the intervention itself, just like what is shown in Table 1. Two groups all possessed similar beginning conditions before they participated in the intervention. No statistic differences were gotten between the two groups on the side of age, gender, H. pylori's condition of infection, histories of smoking and drinking, distributing of lesion positions, and main pathological and serum indexes (all $P > 0.05$).

Table 1: Baseline clinicopathological and serological characteristics of the study cohort

Variable	Wenyujin group (n = 114)	Control group (n = 114)	P value
Age, years	52.8 ± 10.7	53.4 ± 11.1	0.684
Male, n (%)	61 (53.5)	59 (51.8)	0.792
H. pylori positive, n (%)	77 (67.5)	75 (65.8)	0.782
Current smoking, n (%)	33 (28.9)	35 (30.7)	0.768
Alcohol use, n (%)	28 (24.6)	30 (26.3)	0.765
Multifocal lesion, n (%)	43 (37.7)	41 (36.0)	0.789
Moderate-to-severe atrophy, n (%)	74 (64.9)	72 (63.2)	0.789
Moderate-to-severe intestinal metaplasia, n (%)	69 (60.5)	67 (58.8)	0.793
Low-grade intraepithelial neoplasia, n (%)	21 (18.4)	20 (17.5)	0.858
OLGIM stage III–IV, n (%)	36 (31.6)	34 (29.8)	0.769
PGI, ng/mL	63.8 ± 17.4	64.1 ± 18.0	0.903
PGII, ng/mL	13.7 ± 4.6	13.5 ± 4.4	0.739
PGR	4.68 ± 1.23	4.71 ± 1.19	0.851
G-17, pmol/L	12.4 ± 3.9	12.1 ± 3.7	0.548
IL-6, pg/mL	7.9 ± 2.4	7.7 ± 2.3	0.516
TNF-α, pg/mL	18.2 ± 5.0	17.9 ± 4.8	0.642

After we have made confirmation of the baseline equilibrium, the efficacy change tracks of the two groups in a 6-month time length were further done comparison. For the purpose of answering the question whether Wenyujin is able to obtain continuous improvements on pathological aspects, endoscopy, and serum dimensions, a joint comparison of the main efficacy trajectories of the two groups was conducted, as shown in Figure 4. Figure 4(a) indicates that the composite pathology burden score in the Wenyujin group decreased from 5.21 ± 1.37 at Baseline to 4.42 ± 1.30 at Month 3, and further to 3.74 ± 1.21 at Month 6; in the control group, it decreased from 5.18 ± 1.34 to 4.83 ± 1.31 and 4.58 ± 1.28. The group × time interaction term reached a significant level ($P < 0.001$), suggesting that the two groups did not merely experience a natural decline in the same direction, but rather the Wenyujin group showed a greater decrease. Figure 4(b) further demonstrates that the endoscopic mucosal burden score at Month 6 was significantly positively correlated with the pathology burden score ($r = 0.71$, $P < 0.001$), and the data points that belong to the Wenyujin group have a higher degree of gathering in the

quadrant which is low mucosal load-low pathology load. The OLGIM stage transition heatmap that is presented in Figure 4(c) also provides evidence in the same direction: the downstage proportion in the Wenyujin group was 43.0%, compared to 22.8% in the control group; conversely, the upstage proportion was 9.6% in the Wenyujin group and 21.1% in the control group. Figure 4(d) shows that PGI and PGR began to rebound from Month 3, while the decrease in G-17 was more pronounced at Month 6, suggesting that multidimensional indicators moved in the same direction after treatment

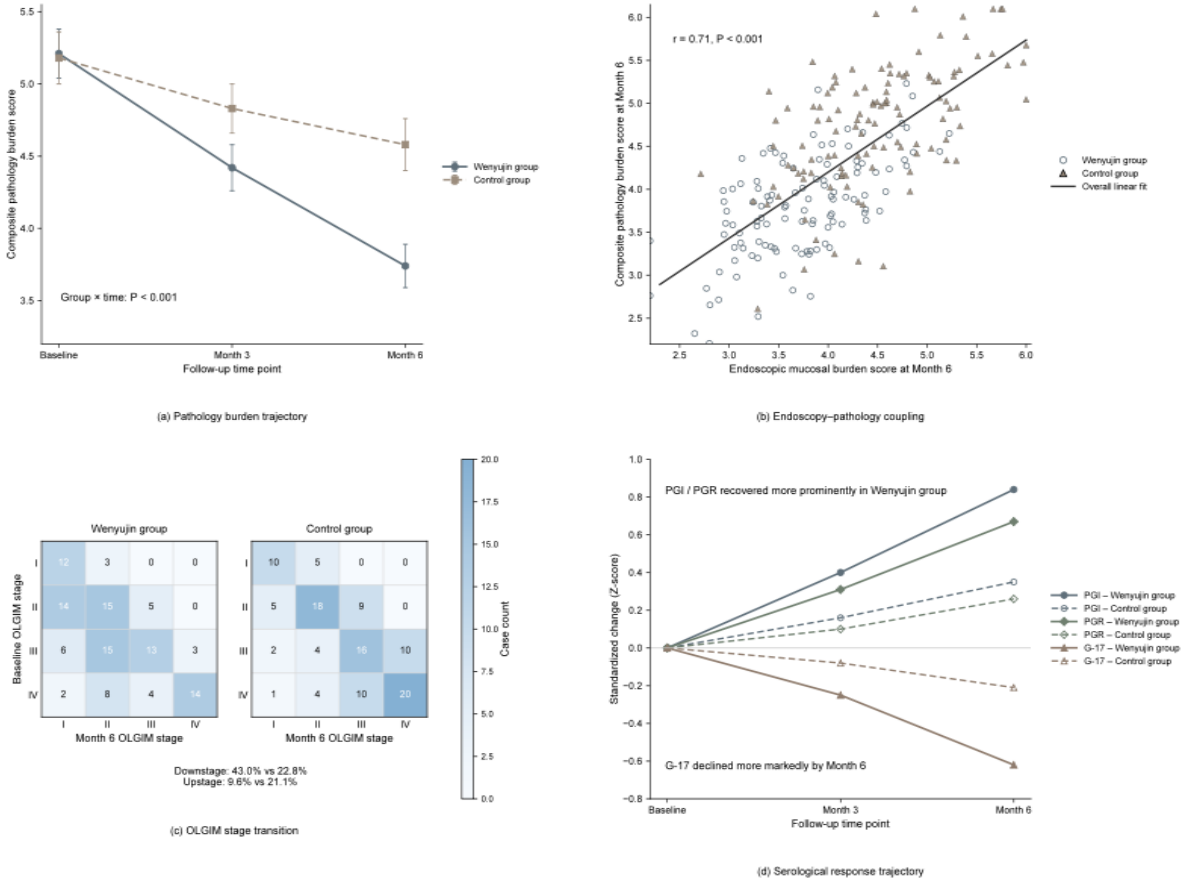

Figure 4: Therapeutic response trajectories of Wenyujin across pathology, endoscopy, and serum indicators

When we concentrate on the whole tendency that is shown in Figure 4, further splitting of each end point gives the results which are shown in Table 3. When it came to the sixth month, the response rate on patient level in the group of Wenyu Jin was 58.8 percent, which was obviously higher than 37.7 percent of the control group (P = 0.001); the pathological alleviation rates were 53.5 percent and 31.6 percent, in turn (P = 0.001). On the level of stage migration, the proportion of downstaging of OLGIM in the group of Wenyu Jin is 43.0%, and the proportion of downstaging of OLGA is 40.4%, hence both of them are higher than 22.8% and 21.1% that in the control group. Correspondingly, in the sixth month after the start, the comprehensive pathology load score of the Wenyu Jin group, compared with the starting baseline, has a reduction of 1.47 ± 0.94 points, while the control group has a reduction of 0.60 ± 0.82 points; the score of endoscopic mucosal load reduced by 1.67 ± 0.88 points and 0.79 ± 0.85 points, separately, hence there are obvious differences between the two groups (both P < 0.001). The change direction of serum markers was consistent with aforementioned results: in

Wenyu Jin group, PGI elevated by 11.2 ± 9.4 ng/mL and PGR elevated by 1.12 ± 0.94, hence G-17, IL-6, and TNF-α reduced by 3.08 ± 2.41 pmol/L, 2.41 ± 1.96 pg/mL, and 4.34 ± 3.52 pg/mL, respectively, therefore their improvement ranges were better than those that are in the control group. The total symptom score additionally further was reduced from the starting baseline.

Table 3: Quantitative changes in therapeutic endpoints after Wenyujin intervention

Endpoint	Wenyujin group (n = 114)	Control group (n = 114)	P value
Response at Month 6, n (%)	67 (58.8)	43 (37.7)	0.001
Pathological remission, n (%)	61 (53.5)	36 (31.6)	0.001
OLGIM downstage, n (%)	49 (43.0)	26 (22.8)	0.001
OLGA downstage, n (%)	46 (40.4)	24 (21.1)	0.002
Upstage cases, n (%)	11 (9.6)	24 (21.1)	0.016
Composite pathology burden score, Baseline	5.21 ± 1.37	5.18 ± 1.34	0.861
Composite pathology burden score, Month 3	4.42 ± 1.30	4.83 ± 1.31	0.018
Composite pathology burden score, Month 6	3.74 ± 1.21	4.58 ± 1.28	<0.001
Change in pathology burden (Month 6 − Baseline)	-1.47 ± 0.94	-0.60 ± 0.82	<0.001
Endoscopic mucosal burden score, Baseline	4.86 ± 1.08	4.81 ± 1.12	0.721
Endoscopic mucosal burden score, Month 3	3.98 ± 1.01	4.36 ± 1.03	0.006
Endoscopic mucosal burden score, Month 6	3.19 ± 0.96	4.02 ± 1.05	<0.001
Change in mucosal burden (Month 6 − Baseline)	-1.67 ± 0.88	-0.79 ± 0.85	<0.001
Change in PGI, ng/mL	+11.2 ± 9.4	+4.3 ± 8.8	<0.001
Change in PGR	+1.12 ± 0.94	+0.39 ± 0.82	<0.001
Change in G-17, pmol/L	-3.08 ± 2.41	-1.21 ± 2.18	<0.001
Change in IL-6, pg/mL	-2.41 ± 1.96	-0.97 ± 1.74	<0.001
Change in TNF-α, pg/mL	-4.34 ± 3.52	-1.86 ± 3.11	<0.001
Change in symptom score	-3.76 ± 2.22	-1.95 ± 2.08	<0.001

According to the outcomes shown in this part, the modifications caused by Wenyu Jin, which is a kind of Chinese traditional medicine, not only ease individual symptoms. On the contrary, they show a mutually promoting developing track of advancement in pathological grading, stage shifting, mucous membrane phenotype, and inflammation-connected indices. Recent overviews about traditional Chinese medicine for chronic atrophic gastritis also put forward that effect evaluations should include mucosal repair, inflammation reduction, and gastric function rebuilding, instead of depending only on one single pathological result. At the same time, the latest summaries which concentrate on gastric intestinal metaplasia and OLGIM/OLGA stratification lay stress that these hierarchical pieces of information still keep being crucial entry points for long-term risk identification.

3.2 Predictive Performance, Ablation Results, and Robustness of the Intelligent Analysis System

This section mainly discusses two problems: in what degree the distinguishing capacity of multi-mode systems has got promotion in the identification of curative effect on the patient-level; and from which study modules does this promotion mainly come, and whether it still keeps steady among different small groups. After we have done the comparison of the primary

endpoint, we then further compared the ability of discrimination and the net clinical benefit that different models have in the identification of patient-level response, just as it is shown in Figure 5. Figure 5(a) shows that the ROC curve of the multimodal model on the test set on the whole is located at the outermost position, with an AUC achieving 0.923 (95% CI: 0.879–0.960), which exceeds the dual-modal model's 0.892, the pathology-only model's 0.861, the endoscopy-only model's 0.838, and the clinical-only model's 0.742. The PR curve that is in Figure 5(b) gives the same ranking, hence the AUPRC of the multimodal model is 0.917. Figure 5(c) further tells us that the calibration line of the multimodal model is the nearest to the ideal diagonal, therefore it shows a Brier score of 0.108 and a calibration slope of 1.01. The decision curve that is drawn in Figure 5(d) also shows that within the main decision range of threshold probabilities which are from 0.20 to 0.75, the multi-mode model always keeps the highest net benefit.

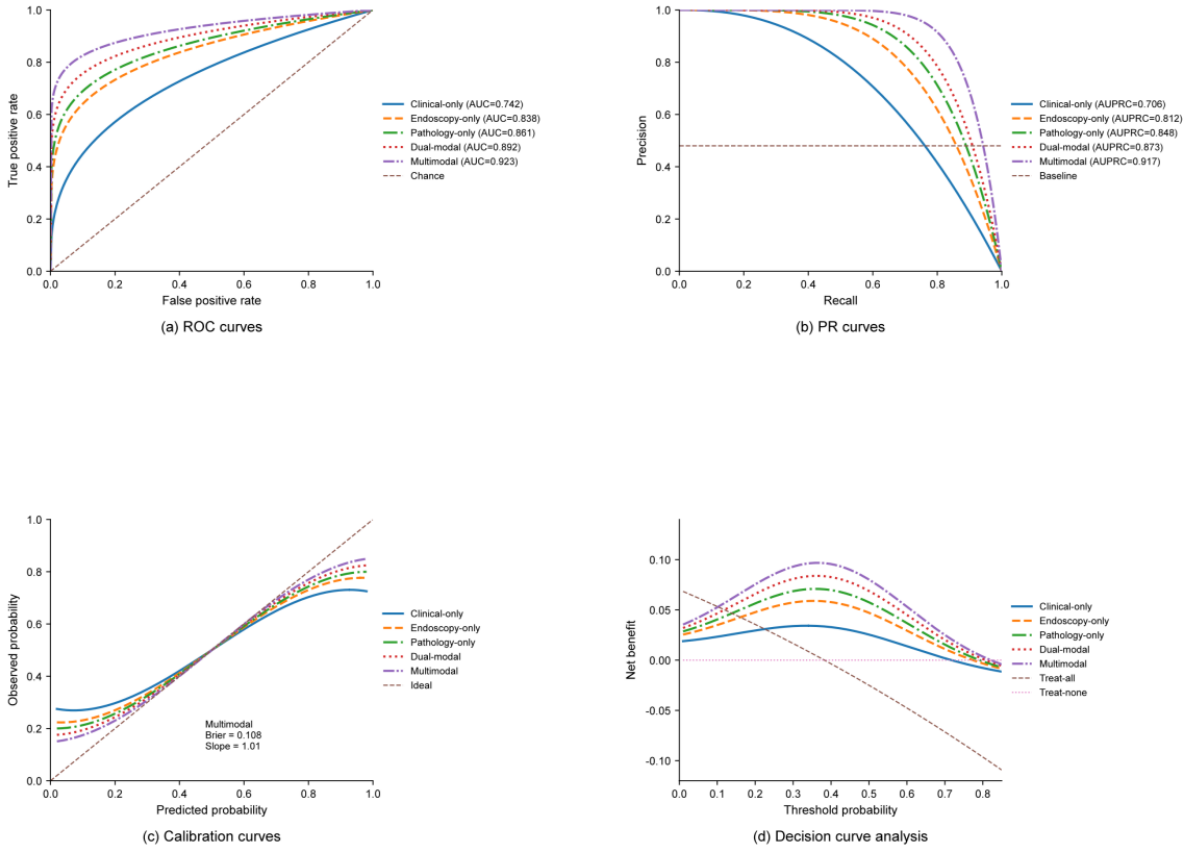

Figure 5: Predictive performance and clinical benefit of the intelligent analysis system.

As for the total ordering in Figure 5, the main performance measuring indexes are further elaborated in Table 4. The accuracy value, F1-score, sensitivity value, and specificity value of the multimodal model are 0.855, 0.854, 0.868, and 0.842, respectively, all these values are the highest among all contrasted models. The model which only uses pathology data has the best performance among all single-modality models, hence it obtains an AUC of 0.861; the model that uses only endoscopy is placed in the following position; although the model which only uses clinical data can give some baseline risk information, it is still not enough when it is used alone to explain the curative effect. It need be pointed out that although the dual-modal model has promoted the AUC to 0.892, it still falls behind the three-modal combined model with respect to F1-score, Brier score, and calibration slope, hence this indicates that the clinical-serological branch does not give marginal supplementation but plays a large role in probability calibration and patient-level explanation stability.

Table 4: Quantitative comparison of single-modal, dual-modal, and multimodal models

Model	AUC	Accuracy	F1-score	Sensitivity	Specificity	Brier score	Calibration slope
Clinical-only	0.742	0.711	0.674	0.651	0.709	0.187	0.81
Endoscopy-only	0.838	0.776	0.764	0.781	0.754	0.151	0.89
Pathology-only	0.861	0.798	0.789	0.814	0.763	0.139	0.93
Dual-modal	0.892	0.823	0.821	0.833	0.807	0.124	0.97
Multimodal	0.923	0.855	0.854	0.868	0.842	0.108	1.01

After we have confirmed that the main model has overall superiority, we have further carried out analysis on the origins of its performance and its deployable capability, which is shown in Figure 6. After we take apart each branch one by one, the ablation outcome that is shown in Figure 6(a) has told us that when the pathology branch is taken away, this causes the AUC of the model to get a decrease from 0.923 to 0.869, and the F1-score also has a fall from 0.854 to 0.798, hence this is the biggest reduction among the three branches, with decreases of 0.054 and 0.056, respectively. After removing the endoscopy branch, the AUC and F1-score decreased to 0.881 and 0.812, respectively; and After we have taken away the clinical-serological branch, the AUC and F1-score have decreased to 0.903 and 0.832. These results give us the idea that the pathology branch possesses a stronger influence upon patient-level explanation, while the endoscopy branch brings about a notable part of morphological improvements. Although the clinical-serology branch does not fix the upper bound of the model, it thus can greatly decrease undulations in the fusion course. Figure 6(b) gives the AUC together with the inference time for each patient, hence it shows that the multimodal model does not enter the extreme area of "high accuracy but high time consumption," and its inference time is 0.392 s per patient. By comparison, the time difference between the dual-modal and multimodal models is merely 0.084 s/patient, however it brings about synchronous promotions in AUC, F1-score, and calibration performance. The subgroup heatmap that is shown in Figure 6(c) further proves that the multimodal model obtains an AUC higher than 0.90 in the patients who have H. pylori positive result, negative result, Baseline OLGIM I–II, Baseline OLGIM III–IV, gastric antral predominant lesions, gastric body predominant lesions, multifocal distribution, males, females, patients aged <55 years, and patients aged ≥55 years.

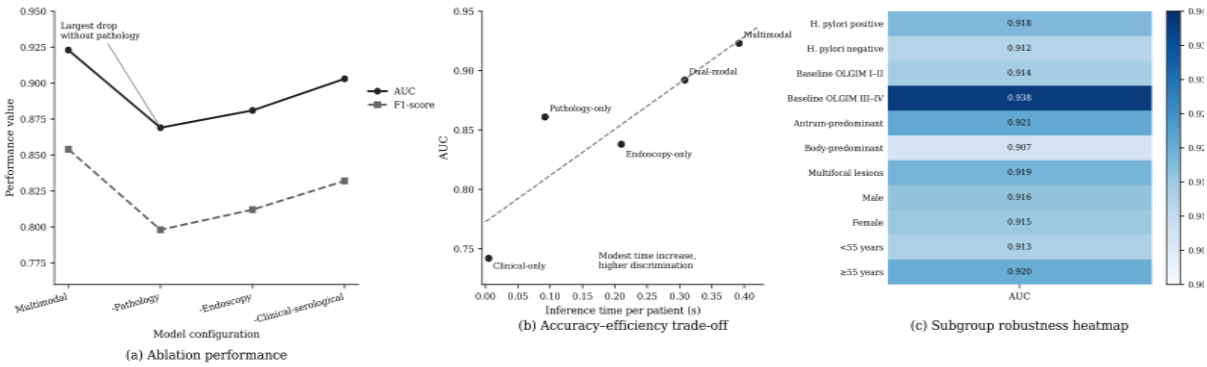

Figure 6 Ablation, efficiency, and subgroup robustness of the proposed system.

From the results of this section, the advantages brought by the joint modeling of the third modality are not only reflected in the improvement of single-point AUC, but also in the ranking ability, probability calibration, threshold decision-making, and subgroup stability. Recent similar studies have demonstrated that real-world gastric lesion AI systems possess the potential to identify pathological gastric abnormalities and lesions requiring treatment. Automated endoscopic grading of gastric intestinal metaplasia can further promote risk stratification, while the SHAP model based on clinical indicators indicates that phenotypic variables have interpretable value in assessing the risk of intestinal metaplasia[19-21].

3.3 Error Sources, Representative Cases, and Clinical Implications of System Deployment

This section primarily addresses three questions: where the errors of the model are mainly concentrated; what exactly the model relies on to make judgments in representative cases; and which steps need to be controlled beforehand if the system is integrated into the actual review process. To explain the basis for the model's judgments in real cases and identify where the errors primarily come from, we further present representative responsive and non-responsive cases, as shown in Figure 7. Figure 7(a) illustrates a typical responder. At baseline, this patient exhibited multifocal mucosal texture disorder, focal whitening, and vascular exposure in the gastric antrum-gastric angle. After the Month 6 review, the abnormal area significantly contracted,in the pathological burden score has the decrease from 5.8 to 3.2, the OLGIM grade has the drop from III to I, PGI has the increase from 56.4 ng/mL to 72.8 ng/mL, PGR has the increase from 4.12 to 5.36, and IL-6 has the decrease from 8.3 pg/mL to 5.1 pg/mL. This model has carried out the assignment of a response probability which is 0.91. The Grad-CAM heatmap in Figure 7(c) mainly covers the gland boundary disorder region and local abnormal blood vessel region in the baseline image. By Month 6, the heatmap significantly shrunk, indicating that the heatmap mainly focuses on the core structure of the lesion rather than staying at the level of overall tone changes.

Figure 7(b) gives one example of one non-responder case. This patient already had patchy intestinal metaplasia, and had focal irregular microsurface predominance in the gastric body at the time of Baseline. After six calendar months passed, the endoscopic look presented only restricted improvement, with the pathological load score having a reduction from 5.4 to 5.1, OLGIM grade II having an escalation to grade III, G-17 alone having a decrease from 13.1 pmol/L to 12.4 pmol/L, and TNF-α having a decrease from 18.9 pg/mL to 17.8 pg/mL, which shows that improvement that is comparatively small exists. The model has given out a probability of non-response which is 0.87. The SHAP ordering which is shown in Figure 7(d) makes it known that the alterations on pathological burden, OLGIM stage change, PGR variation, IL-6 variation, lesion position, and H. The pylori condition are the chief variants that cause the divergence in forecast outcomes between the two patients. When we look from the background of one representative case, this dual-channel explanation method helps to combine "at which place the model is looking" and "which kinds of variables are making prediction probability become higher or lower" inside this same framework. Recent overviews on explainable artificial intelligence in human gastrointestinal canal also emphasize that hot region tracking and parameter assignment must be connected to particular clinical objects, instead of staying on the separate visualization level[22].

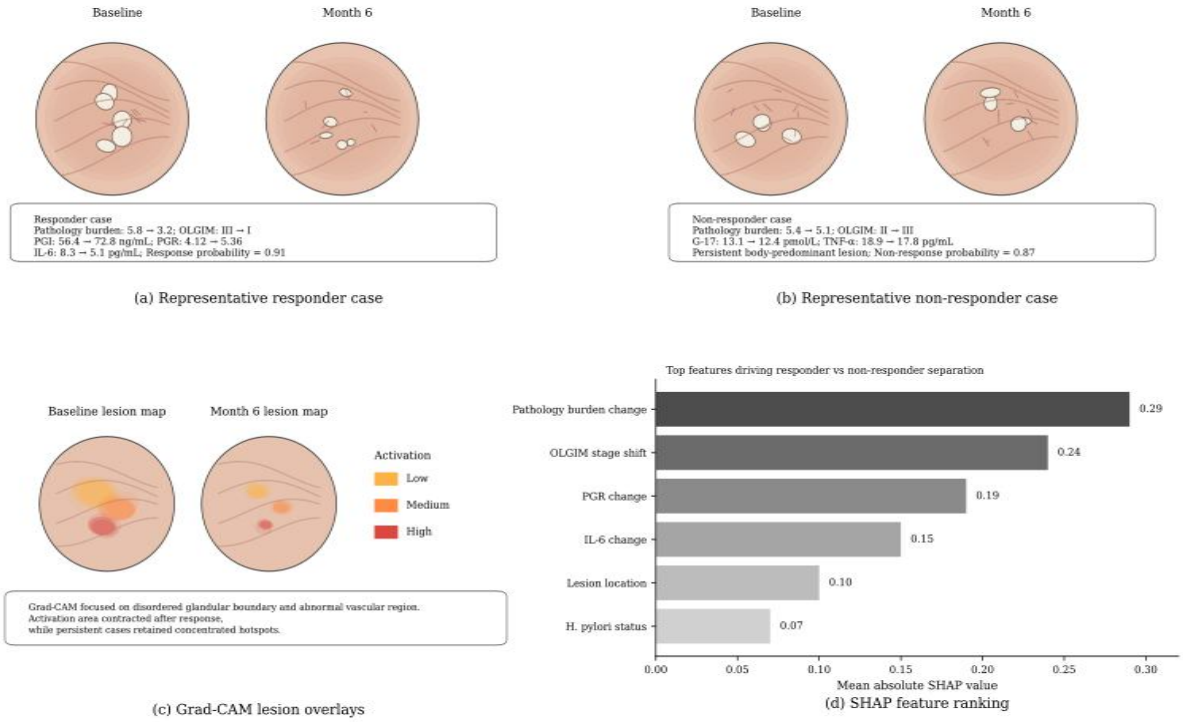

Figure 7: Representative response and non-response cases with model interpretation.

Beside representative cases, we on our side continue to make summary of the distribution of wrong judgments, subgroup deviation, and main origins of mistakes, which is displayed in Table 5. According to the combined prediction results at patient level from cross-validation and independent testing, altogether 33 main wrong judgments were recorded, among them there are 18 false positive cases and 15 false negative cases. This shows us that mistakes are not uniformly spread among every case, but are gathered together in circumstances like boundary period change, multiple focus damage, low grade endoscopic pictures, and not synchronous alterations between pathology examination and endoscopic examination. When we carry out further decomposition, we can discover that borderline stage change and multifocal lesion cases together take up 54.5% (18/33) of all wrong judgments; Wrong estimations that connect with low-quality endoscopic image frames take up 21.2% (7/33). False-positive instances are more frequently encountered in patients that have active inflammation of H. pylori which has not wholly gone down yet but has got short-term betterment on surface appearance, hence false-negative results are more frequently seen in patients who have a dominant gastric body, smaller lesion region, or patchy spreading of intestinal metaplasia.

Table 5: Error distribution, subgroup deviations, and major misclassification patterns

Error category	False-positive, n	False-negative, n	Dominant deviation pattern	Suggested handling
Borderline stage shift	7	4	Endoscopic improvement present, but pathological downgrade did not fully meet response threshold	Add pathology-priority review for near-threshold cases
Multifocal lesions	4	3	Cross-site inconsistency between local visual regression and residual histological burden	Use site-aware re-check and multi-site aggregation
Low-quality endoscopic frames	3	4	Blur, mucus, specular reflection, or unstable focus weakened lesion morphology signals	Trigger image-quality gate before inference
Patchy IM / focal LGIN mismatch	2	2	Localized histological abnormalities were underrepresented in visible surface changes	Add targeted biopsy confirmation
Active H. pylori-related inflammation	1	1	Inflammatory background mimicked partial response or masked true persistence	Reassess together with inflammatory markers
Body-predominant subtle lesions	1	1	Mild structural distortion produced weak visual contrast	Increase weight of pathology and serological branch

The distribution of these errors indicates that model errors do not primarily stem from "complete misinterpretation", but rather from the tension between multimodal evidence: endoscopy has shown some improvement, but pathology has not yet experienced a comparable decrease; or pathological sites have migrated, but the visible phenotype remains in a local residual stage. Correspondingly, recent reviews on the quality control of endoscopic AI have pointed out that image quality monitoring, in-process quality control, and standardization of image interpretation should not be regarded as external issues of the model, but rather as part of system deployment[23].

From the angle of deployment, the system in this research is more fit for being put into the "second reader" and "risk re-assessment interface" positions in the real checking process, hence it does not take the place of independent decision-making of pathologists or endoscopists. To patients who have high probability of response but still are in the borderline stage of transition, the system's value lies in prompting that the re-checking of pathological positions and image quality is necessary, therefore it does not directly lower the intensity of follow-up visits. To the patients that have continuously high non-response probability and have not obtained improvement on both pathological and serum indicators, therefore this system can be used as an auxiliary interface for review arrangement and risk re-assessment. Recent summaries on the clinic practice putting-in of gastrointestinal AI have found that workflow connection, rule conforming, data drift watching, and human-machine cooperation are pre-conditions for putting out use, hence it shows that the coming of high-effect models into clinic work does not only rely on test set marks, but also relies on their capability to connect with photograph norms, report systems, and check decision points[24].

Through deeper investigation, the arrangement risks in actual situation environments also

include too much dependence on automatic results, ignoring the cross-center distribution differences, and not enough alert for mistake probability. The recent arguments about gastrointestinal artificial intelligence have already clearly pointed out that, diagnostic deviation, not enough transparency, and unclear responsibility limits, once the model walks into clinical application, therefore will amplify the results of a single wrong judgment. Therefore, this research has combined error summing-up, typical case explanation, picture quality control, and manual check interfaces into the one same section. The purpose is not to add an extra explanation to the model, but instead to place the interpretation foundation on checkable image areas and variables, thus clearly marking the system's usable scopes, therefore making it more in line with safe use norms in real-world checking procedures.[25].

4 Conclusion

This article puts emphasis on the effect assessment of Curcuma wenyujin intervention for precancerous lesions of stomach cancer, therefore it establishes a multi-mode intelligent analysis frame which is made specially for patient-level follow-up situations. This research does not put emphasis on recognizing lesion in one single check-up, but instead it puts pathological change, endoscopic appearance, and clinical serum index into same evaluation coordinate system, thus to together evaluate degree of improvement, stage change, and probability of response after intervention.

(1) At the object organization and data level, this paper constructs a structured sample consisting of 228 patients, 5,472 endoscopic images, 1,824 pathological site labels, and 26 clinical serum variables, and achieves longitudinal alignment according to Baseline, Month 3, and Month 6. This data organization method enables efficacy evaluation to no longer rely on a single pathological conclusion or a single endoscopic observation, but can track the continuous changes in lesion burden over time, serving as a foundation for subsequent expansion work.

(2) At the methodological and result levels, this paper constructs a multimodal system consisting of an endoscopic branch, a pathological branch, and a clinical-serological branch, simultaneously outputting efficacy response and stage migration results. The results of this paper show that the Wenyu Jin group is superior to the control group in terms of response rate, pathological response rate, OLGIM/OLGA downstage, mucosal load reduction, and improvement in inflammatory indicators; the corresponding multimodal model achieves an AUC of 0.923, an F1-score of 0.854, and a Brier score of 0.108 in patient-level response recognition, with fluctuations in different subgroups controlled within a relatively small range, indicating that the system not only possesses discriminative ability but also has good calibration performance, and remains stable in key reading indicators, closer to the usage requirements in real review processes. In other words, this paper completes patient-level joint interpretation and can be integrated into the review interpretation chain to support review arrangements and risk reassessment.

(3) In this article, there are still quite a few boundaries which have a necessity for being retained. First of all, the main analysis group comes from only one center, hence there can still be distribution differences between the sample structure and the actual clinical crowd. Secondly, the existing follow-up time is mostly 6 months, hence the evaluation for migration and recurrence risks in longer time periods is still not enough. Thirdly, the boundary stage movement, multiple-focus damages, and low-grade endoscopic pictures are still the main origins of mistakes. In the time that comes, we can still go forward through adding multi-center forward-looking samples, stretching the follow-up time window, and carrying out a combined flow which includes picture quality controlling, pathology-first checking, and cross-center standard adjusting, therefore to let the system get more in line with the safe use demands in the

actual checking route.

References

[1] Dinis-Ribeiro, M., Libânio, D., Uchima, H., et al. (2025). Management of epithelial precancerous conditions and early neoplasia of the stomach (MAPS III): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG) and European Society of Pathology (ESP) guideline update 2025. Endoscopy, 57(5), 504–554. DOI: 10.1055/a-2529-5025.

[2] Dinis-Ribeiro, M., Shah, S., El-Serag, H., et al. (2024). The road to a world-unified approach to the management of patients with gastric intestinal metaplasia: a review of current guidelines. Gut, 73, 1607–1617. DOI: 10.1136/gutjnl-2024-333029.

[3] Tong, Q. Y., Pang, M. J., Hu, X. H., et al. (2024). Gastric intestinal metaplasia: progress and remaining challenges. Journal of Gastroenterology, 59(4), 285–301. DOI: 10.1007/s00535-023-02073-9.

[4] Fang, S., Liu, Z., Qiu, Q., et al. (2024). Diagnosing and grading gastric atrophy and intestinal metaplasia using semi-supervised deep learning on pathological images: development and validation study. Gastric Cancer, 27(2), 343–354. DOI: 10.1007/s10120-023-01451-9.

[5] Tao, X., Zhu, Y., Dong, Z., et al. (2024). An artificial intelligence system for chronic atrophic gastritis diagnosis and risk stratification under white light endoscopy. Digestive and Liver Disease, 56(8), 1319–1326. DOI: 10.1016/j.dld.2024.01.177.

[6] Li, N., Yang, J., Li, X., et al. (2024). Accuracy of artificial intelligence-assisted endoscopy in the diagnosis of gastric intestinal metaplasia: a systematic review and meta-analysis. PLOS ONE, 19(5), e0303421. DOI: 10.1371/journal.pone.0303421.

[7] Yu, S., Jiang, H., Xia, J., et al. (2025). Construction of machine learning-based models for screening the high-risk patients with gastric precancerous lesions. Chinese Medicine, 20(1), 7. DOI: 10.1186/s13020-025-01059-4.

[8] Tao, J., Zhang, Z., Meng, L., et al. (2025). Risk prediction model for precancerous gastric lesions based on magnifying endoscopy combined with narrow-band imaging features. Frontiers in Oncology, 15, 1554523. DOI: 10.3389/fonc.2025.1554523.

[9] Zou, T. H., Gao, Q. Y., Liu, S., et al. (2024). Effectiveness and safety of Moluodan in the treatment of precancerous lesions of gastric cancer: a randomized clinical trial. Journal of Digestive Diseases, 25(1), 27–35. DOI: 10.1111/1751-2980.13251.

[10] Luo, Y., Zhu, L., Ren, Z., et al. (2025). Curcumae Rhizoma: An anti-cancer traditional Chinese medicine. Chinese Herbal Medicines, 17(3), 428–447. DOI: 10.1016/j.chmed.2025.04.006.

[11] Wang, L., Lian, Y. J., Dong, J. S., et al. (2025). Traditional Chinese medicine for chronic atrophic gastritis: efficacy, mechanisms and targets. World Journal of Gastroenterology, 31(9), 102053. DOI: 10.3748/wjg.v31.i9.102053.

[12] Borgli, H., Thambawita, V., Smedsrud, P. H., et al. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data, 7, 283. DOI: 10.1038/s41597-020-00622-y.

[13] Bravo, D., Frías-Ordoñez, J. S., Vera Polanía, F., et al. (2025). GastroHUN an endoscopy dataset of complete systematic screening protocol for the stomach. Scientific Data, 12(1), 102. DOI: 10.1038/s41597-025-04401-5.

[14] Wang, C., Ge, J., Niu, Y., et al. (2025). A fully annotated pathology slide dataset for early gastric cancer and precancerous lesions. Scientific Data, 12(1), 1326. DOI: 10.1038/s41597-025-05679-1.

[15] Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. DOI: 10.1136/bmj-2023-078378.

[16] Riley, R. D., Archer, L., Snell, K. I. E., et al. (2024). Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ, 384, e074820. DOI: 10.1136/bmj-2023-074820.

[17] Van Calster, B., Collins, G. S., Vickers, A. J., et al. (2025). Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. The Lancet Digital Health, 7(12), 100916. DOI: 10.1016/j.landig.2025.100916.

[18] Saelmans, A., Seinen, T., Pera, V., et al. (2025). Implementation and Updating of Clinical Prediction Models: a systematic review. Mayo Clinic Proceedings: Digital Health, 3(3), 100228. DOI: 10.1016/j.mcpdig.2025.100228.

[19] Chang, Y. H., Shin, C. M., Lee, H. D., et al. (2024). Real-World Application of Artificial Intelligence for Detecting Pathologic Gastric Atypia and Neoplastic Lesions. Journal of Gastric Cancer, 24(3), 327–340. DOI: 10.5230/jgc.2024.24.e28.

[20] Almeida, E., Martins, M. L., Marques, D., et al. (2025). Artificial intelligence for endoscopic grading of gastric intestinal metaplasia: advancing risk stratification for gastric cancer. Endoscopy, 57(11), 1254–1260. DOI: 10.1055/a-2657-9906.

[21] Wang, Y., Bi, J., Song, S., et al. (2025). Identifying gastric intestinal metaplasia risk based on clinical indicators: a machine learning predictive model based on the SHAP methodology. Frontiers in Pharmacology, 16, 1602191. DOI: 10.3389/fphar.2025.1602191.

[22] Mascarenhas, M., Mendes, F., Martins, M., et al. (2025). Explainable AI in Digestive Healthcare and Gastrointestinal Endoscopy. Journal of Clinical Medicine, 14(2), 549. DOI: 10.3390/jcm14020549.

[23] Mori, Y., Misawa, M. (2025). Quality assessment in endoscopy “artificial intelligence in endoscopy”. Best Practice & Research Clinical Gastroenterology, 76, 102006. DOI: 10.1016/j.bpg.2025.102006.

[24] El-Sayed, A., Lovat, L. B., Ahmad, O. F. (2025). Clinical Implementation of Artificial

Intelligence in Gastroenterology: Current Landscape, Regulatory Challenges, and Ethical Issues. Gastroenterology, 169(3), 518–530. DOI: 10.1053/j.gastro.2025.01.254.

[25] El-Sayed, A., Lovat, L. B., Ahmad, O. F. (2025). Clinical Implementation of Artificial Intelligence in Gastroenterology: Current Landscape, Regulatory Challenges, and Ethical Issues. Gastroenterology, 169(3), 518–530. DOI: 10.1053/j.gastro.2025.01.254.