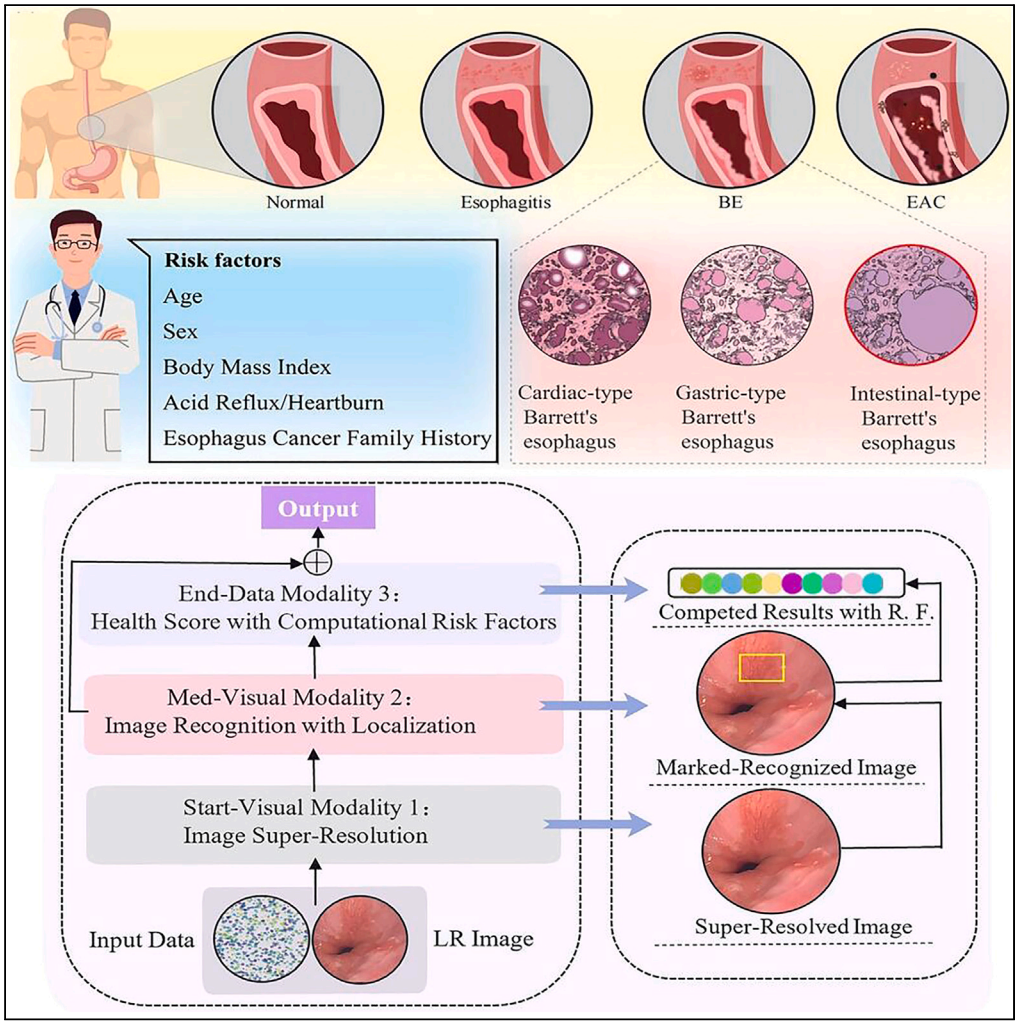
iScience

CellPress
OPEN ACCESS

Article

Multimodal integration for Barrett's esophagus



Shubin Liu, Shiyu Peng, Mengxuan Zhang, Ziyuan Wang, Lei Li

medicalp123@hotmail.com (S.P.)
leili@scu.edu.cn (L.L.)

Highlights

Multimodal framework to tackle tasks from various modalities

High-level modality combines low-level modality to form final risk grading

SR pipeline enable disease identification with high-accuracy positioning

Liu et al., iScience *27*, 108437
February 16, 2024
https://doi.org/10.1016/j.isci.2023.108437

iScience

CellPress
OPEN ACCESS

Article

Multimodal integration for Barrett's esophagus

Shubin Liu,[1] Shiyu Peng,[2,*] Mengxuan Zhang,[3] Ziyuan Wang,[1] and Lei Li[1,4,*]

SUMMARY

The esophageal adenocarcinoma is facing a worldwide challenge: early prediction and risk assessment in clinical Barrett's esophagus (BE). In recent years, the growing interests have been witnessed in prediction and risk assessment in clinical BE. However, the resolution is limited, and the system is huge and expensive for the existing devices. Inspired by the principle of collaboration between human eye vision and brain cortex in data processing, here we propose multimodal learning framework to tackle tasks from various modalities, which can benefit from each other. To our findings, the experimental result indicates that low-level modality can directly affect high-level modality and form the final risk grading based on contribution, which maximizes the clinical performance of medical professionals based on our findings.

INTRODUCTION

The discipline of esophageal cancer is facing a major global challenge: preventing its early onset and reducing its severity.[1] According to the statistics for 2020,[2] the global incidence rate of the disease reached an astonishing 604,000 cases, ranking 7th in the world's most common malignant tumors. Unfortunately, the mortality rate of esophageal cancer has claimed the lives of approximately 54,000 patients, ranking 6th in terms of mortality. Esophageal cancer can be roughly divided into two main types: esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC). In recent years, the incidence rate of EAC has risen sharply, with an annual incidence rate of 0.7 cases per 100,000 people. EAC is with the characteristics of occult onset, difficulty in early diagnosis, high malignancy, and poor prognosis, which makes more complex.[3] The development of EAC is closely related to the presence of Barrett's esophagus (BE), characterized by a clear boundary between esophageal squamous epithelium and gastric columnar epithelium under endoscopy. In patients with upward displacement of $\geq$1 cm at the junction of BE and gastroesophageal mucosa, metaplastic columnar epithelium replaced the normal stratified squamous epithelium in the lower esophageal segment.[4] Metaplasia can be gastric fundus gland metaplasia, cardiac epithelial metaplasia, or special intestinal metaplasia (SIM).[5] Among them, BE with intestinal metaplasia has a higher risk of cancer transformation. However, the exact mechanism of BE carcinogenesis is not fully understood, highlighting the importance of identifying BE risk factors and developing relevant computer models for early detection. Traditional medicine categorizes the presence of columnar epithelium in the BE based on histology: gastric fundus type, cardiac type, and intestinal type.[6] Among them, intestinal-type BE is considered a precancerous lesion closely related to the occurrence of cancer. Mature intestinal epithelial cells can acquire additional mutations and then develop into dysplasia and cancer. The detection and risk grading (RG) of intestinal-type BE are facing key challenges worldwide. While it's true that many people with BE do not experience any symptoms, this does not necessarily mean they are out of harm's way. In fact, the danger of a disease going unnoticed can be even more dangerous, since by the time it's discovered, it may be in a late stage. For clinical BE, low resolution (LR) images still dominate due to limitations in optical and computational principles. The LR-endoscopic images can be attributed to several factors. (1) Many current endoscopic imaging devices, such as the Olympus CF-140L and EC-3890Li, are limited to resolutions below 1920p. (2) Endoscopic images are often downsampled to lower resolutions before being transmitted to medical systems, which is due to limited storage space and the high cost of data transfer. In this case, the images are often compressed and downsampled to reduce data size and transmission time, thereby reducing image resolution. However, while high-definition endoscopes offer good image quality, they do come with some disadvantages, such as higher cost and larger size. These factors can make it difficult for some medical facilities to justify the purchase of high-definition endoscopes, particularly in cases where lower resolution imaging is considered to have the potential to complete the clinical tasks at hand. Additionally, the larger size of high-definition endoscopes may also pose challenges in terms of maneuverability and patient comfort during procedures. In recent years, it is exciting that artificial intelligence technology has made significant progress in cancer identification and prognosis,[7–10] especially the use of Conventional Neural Networks (CNN),[11] which has shown extraordinary originality in these fields. For example, a CNN network was successfully used in the task of automatically detecting gastric cancer and the segmentation of rectal cancer. Given the user-friendliness of the device, the focus of the issue appears to be shifting from simply purchasing more expensive equipment to finding ways to achieve higher accuracy with low-cost endoscopes. Here we discussion the inherent logical

[1]School of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China
[2]Department of Gastroenterology, First Affiliated Hospital of Shihezi University, Xinjiang 832061, China
[3]Faculty of Science, The University of Melbourne, Parkville, VIC 3010, Australia
[4]Lead contact
*Correspondence: medicalp123@hotmail.com (S.P.), leili@scu.edu.cn (L.L.)
https://doi.org/10.1016/j.isci.2023.108437

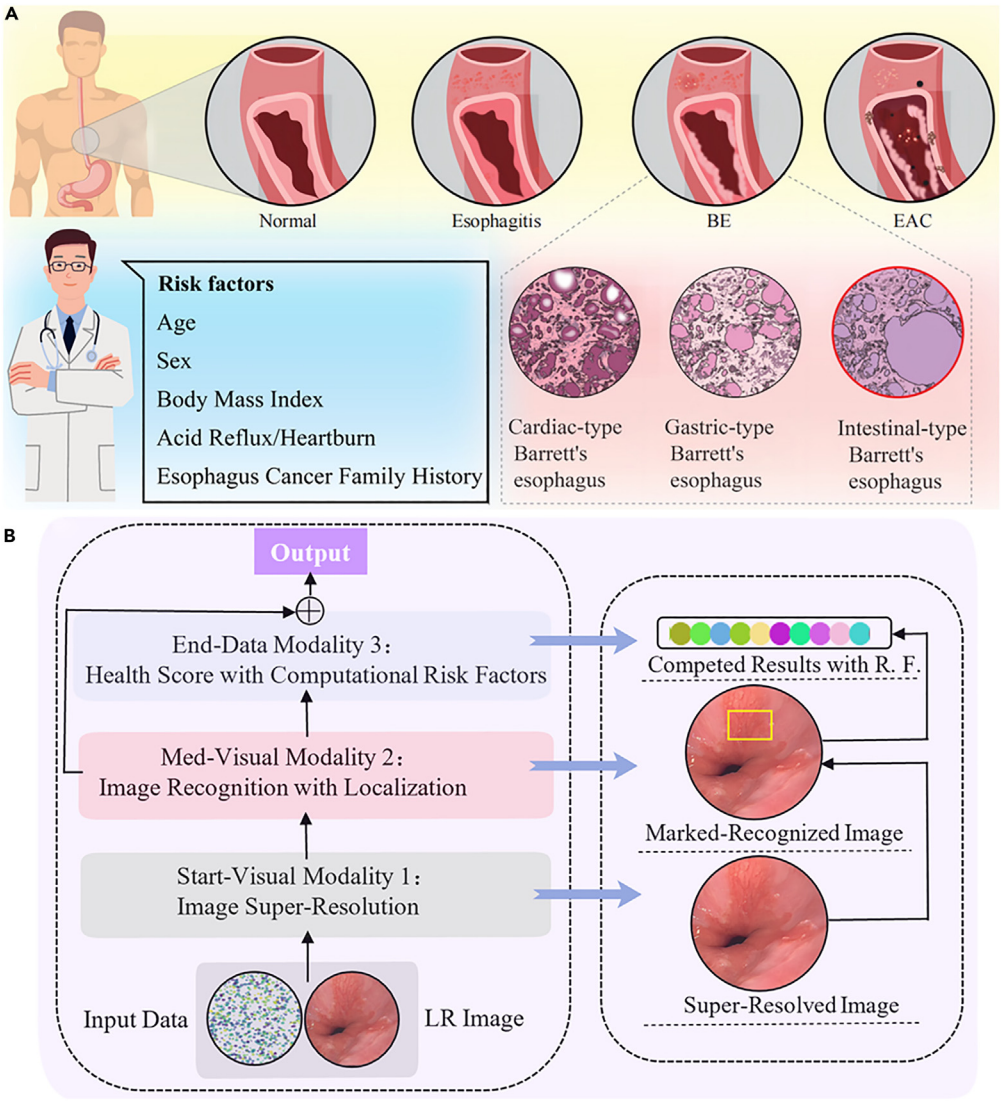


Figure 1. Principle and concept for BE with multimodal learning framework

(A) Development process of various pathological states of EAC and risk factor of BE.

(B) Multimodal computational process.

relationship between super resolution (SR), disease identification (DI), and RG for BE, to the best of our knowledge, our work is the first attempt to perform multimodal tasks in clinical BE.

Inspired by the cooperation between human vision and cerebral cortex in data processing and the principle of attention, here we describe a multimodal learning framework to tackle tasks from various modalities, which can benefit from each other. In the visual modality, we developed 2 distinct yet collaborative real-time pipelines (SR and DI) that effectively filter out irrelevant smooth regions and focus on specific texture details. Additionally, we investigated the significant impact of different resolutions on real-time DI. In the data modality, we designed a real-time data pipeline (RA) to handle high-level tasks. We predicted the score of future health assessments and employed an ablation design to explore the impact of multiple risk factors on health, thereby maximizing the clinical performance of medical professionals based on our findings.

Methods

Principle and concept

EAC patients experience various pathological states based on the severity of their condition, including normal, esophagitis, BE, and EAC, as described in Figure 1A. Perhaps esophagitis may not have a fundamental impact on EAC, but BE does, which evolves directly into EAC. BE can

Table 1. Basic information of BE patients with different pathological subtypes

Groups	SIM%	Non SIM%	total	χ	P
Age (years)				24.322	0.000
<50	5(4.5)	188(25.6)	193		
≫50	106(95.5)	547(74.4)	653		
Gender				10.593	0.001
Male	85(76.6)	445(60.5)	530		
Female	26(23.4)	290(39.5)	316		
BMI				29.282	0.001
<24	16(14.4)	239(32.5)	255		
24 ≪BMI<28	48(43.2)	342(46.5)	390		
BMI≫28	47(42.3)	154(21.0)	201		
Marriage				2.22	0.528
Married	102(91.9)	688(93.6)	790		
Divorced	3(2.7)	26(3.5)	29		
Widowed	5(4.5)	18(2.4)	23		
Unmarried	1(0.9)	3(0.4)	4		
degree of education				2.528	0.64
Bachelor's degree or above	29(26.1)	204(27.8)	233		
Junior college	36(32.4)	211(28.7)	247		
High school	12(10.8)	89(12.1)	101		
Junior high school	14(12.6)	124(16.9)	138		
Primary school and below	20(18.0)	107(14.6)	127		
Smoking habits				0.000	0.991
Yes	46(41.4)	305(41.5)	351		
No	65(58.6)	430(58.5)	495		
Drinking habits				1.923	0.165
Yes	35(31.5)	282(38.4)	317		
No	76(68.5)	453(61.6)	529		

be divided into gastric type, cardiac type, and intestinal type based on the morphological characteristics of mucosal epithelium, with intestinal type being the focus of this study. With a two-year follow-up record for clinical BE, the health information of 846 patients was closely monitored. Natural-visual information will be further processed and interpreted by the cerebral cortex, although the light signal is collected by the visual organ, it will be converted into neural signal for cerebral cortex to explain. Both are indispensable, and attention plays important roles in both visual information and neural information. Inspired by this, here we describe a multimodal learning framework illustrated in Figure 1B to tackle tasks from various modalities, which can benefit from each other. In the visual modality, we develop 2 distinct yet collaborative real-time pipelines with attention (SR and DI) that effectively filter out irrelevant smooth regions and focus on specific texture details. Additionally, we investigate the significant impact of different resolutions on real-time DI. In the data modality, we design a real-time data pipeline with attention (RA) to handle high-level tasks. We predict the score of future health assessments and employ an ablation design to explore the impact of multiple risk factors on health, such as how are these risk factors ranked. It can be expected that a high-precision prognosis system will reduce the investment in manpower, material resources, and efficiency.

Clinical BE medical research

In this study, we included a total of 846 patients with BE. We collected data on their age, gender, BMI, margin, degree of education, smoking and drinking habits, belching, absolute tension, acid reflux/heartburn, foreign body sensation, high-fat diet, anxiety, anorexia, palpitations, chest pain, cough, sleep status, hypertension, coronary heart disease, diabetes, esophagitis, *Helicobacter pylori* infection, family history of esophageal cancer, and divided them into two groups based on pathological types: SIM and non-SIM. We then performed a single-factor logistic regression analysis and included significant single factors in a multivariate analysis to determine the independent risk factors for SIM. We assigned scores based on these independent risk factors and conducted computer simulation analysis. For Table 1, the basic data from patients with BE is collected, including age, gender, BMI, margin, degree of education, smoking and drinking habits. We then performed a single-factor logistic regression analysis and found that age, gender, and BMI were statistically significant ($p < 0.05$).

Table 2. Clinical symptoms of BE patients with different pathological subtypes

Groups	SIM%	Non SIM%	total	χ	P
Belching				0.748	0.387
Yes	18(16.2)	97(13.2)	115		
No	93(83.8)	638(86.8)	731		
Abdominal distension				0.469	0.493
Yes	25(22.5)	145(19.7)	170		
No	86(77.5)	590(80.3)	676		
Acid reflux/heartburn				8.248	0.004
Yes	101(91.0)	711(96.7)	812		
No	10(9.0)	24(3.3)	34		
Foreign body sensation				0.179	0.673
Yes	55(49.5)	380(51.7)	435		
No	56(50.5)	355(48.3)	411		
High-fat diet				0.121	0.727
Yes	3(2.7)	16(2.2)	19		
No	108(97.3)	719(97.8)	827		
Anxiety				0.893	0.345
Yes	11(9.9)	54(7.3)	65		
No	100(90.1)	681(92.7)	781		
Anorexia				1.432	0.231
Yes	4(3.6)	48(6.5)	52		
No	107(96.4)	687(93.5)	794		
Palpitate				0.532	0.466
Yes	13(11.7)	105(14.3)	118		
No	98(88.3)	630(85.7)	728		
Chest pain				0.045	0.831
Yes	26(12.7)	179(87.3)	205		
No	85(76.6)	556(75.6)	641		
Cough				0.924	0.336
Yes	5(4.5)	51(6.9)	56		
No	106(95.5)	684(93.1)	790		
Sleep status				0.427	0.514
Yes	22(19.8)	166(22.6)	188		
No	89(80.2)	569(77.4)	658		

For Table 2, the clinical symptoms from patients with BE were also collected, including belching, absolute tension, acid reflux/hardburn, foreign body sensation, high-fat diet, anxiety, anorexia, palpitations, chest pain, cough, and sleep status. Results from a single-factor logistic regression analysis showed that acid reflux/heartburn was statistically significant ($p < 0.05$).

In Table 3, the comorbidities in BE patients were collected, which included hypertension, coronary heart disease, diabetes, esophagitis, *Helicobacter pylori* infection, family history of esophageal cancer, and single factor logistic regression analysis. The results showed that the family history of esophageal cancer had statistical significance ($p < 0.05$).

And in Table 4, the inclusion of single factor meaningful factors in multivariate analysis suggests that age, gender, BMI, acid reflux/hazard, family history of esophageal cancer are all significant.

RESULTS

The interplay between human visual perception and the cortical brain is achieved through the transmission and processing of neural pathways, whose complex structures and precise regulation enable us to perceive a wide range of visual information and process it rapidly and accurately. In the visual processing of the cortical brain, attention mechanisms regulate the activity of neurons to more effectively process

Table 3. Complication of diseases in BE patients with different pathological subtypes

Groups	SIM%	Non SIM%	total	χ	P
Hypertension				11.122	0.001
Yes	47(42.3)	198(26.9)	245		
No	64(57.5)	537(73.1)	601		
Coronary heart disease				1.973	0.16
Yes	21(18.9)	102(13.9)	123		
No	90(81.1)	633(86.1)			
Diabetes				3.333	0.068
Yes	22(19.8)	98(13.3)	120		
No	89(90.2)	637(86.7)	726		
Esophagitis				5.062	0.167
LA-A	65(58.6)	506(68.8)	571		
LA-B	18(16.2)	87(11.8)	105		
LA-C	25(22.5)	121(16.5)	146		
LA-D	3(2.7)	21(2.9)	24		
Helicobacter pylori infection				0.173	0.678
Yes	43(38.7)	300(40.8)	343		
No	68(61.3)	435(59.2)	503		
Family history of esophageal cancer				4.592	0.000
Yes	19(17.1)	7(1.0)	26		
No	92(82.9)	728(99.0)	820		

visual information of objects we focus on and filter out irrelevant or secondary information. Inspired by this principle, here we describe a multi-modal learning framework to tackle tasks from various modalities, which can benefit from each other. In this section, the visual modality is illustrated in Figure 2, here we describe 2 distinct yet collaborative real-time pipelines that effectively filter out irrelevant smooth regions, and focus on specific texture details.

For the real-time SR pipeline in Figure 2A, it is mainly composed of feature extraction, shrinking, non-linear mapping, expanding, coordinate attention and deconvolution operation. Parametric Rectified Linear Unit (PRELU) is selected as the activation function, and each layer is activated using it. The feature extraction includes a 2D conv layer with 5 × 5 kernel and a PRELU function. The shrinking includes a 2D conv layer with 1 × 1 kernel and a PRELU function. The non-linear mapping includes 4 2D conv layers with 3 × 3 kernel and 4 PRELU functions. The expanding includes a 2D conv layer with 1 × 1 kernel and a PRELU function. The coordinate attention includes a coordinate attention layer and a PRELU function, where ⊗ represents dot product operation. Here the coordinate attention will automatically focus on the area of interest, which greatly alleviates the limitations. The deconvolution includes a deconv layer with 9 × 9 kernel to generate a 4× SR image. The output SR image is further into the DI pipeline as input. In this section, an ingenious idea is to apply specific optimization function to the texture area and the smooth area. For example, while only MSE is used in the smooth area, L1 and MSE are used for joint training in the texture area, to help retain as much texture details as possible. The loss function in the texture area can be described as:

$$Loss_1 = W_1*MSE + W_2*L1 \quad \text{(Equation 1)}$$

Table 4. SIM multivariate logistic regression analysis of risk factors

variable	β	SE	Wald	P	OR	95%CI
Age	2.207	0.513	18.530	0.000	9.089	3.327-24.829
gender	0.714	0.266	7.200	0.007	0.490	0.291-0.825
BMI				0.000		
BMI	1.294	0.339	14.601	0.000	3.648	1.878–7084
BMI	0.888	0.252	12.415	0.000	2.429	1.483-3.980
Acid reflux/heartburn	1.277	0.458	7.757	0.005	3.585	1.460-8.804
Family history of esophageal cancer	3.354	0.535	39.3.9	0.000	29.626	10.032–81.688

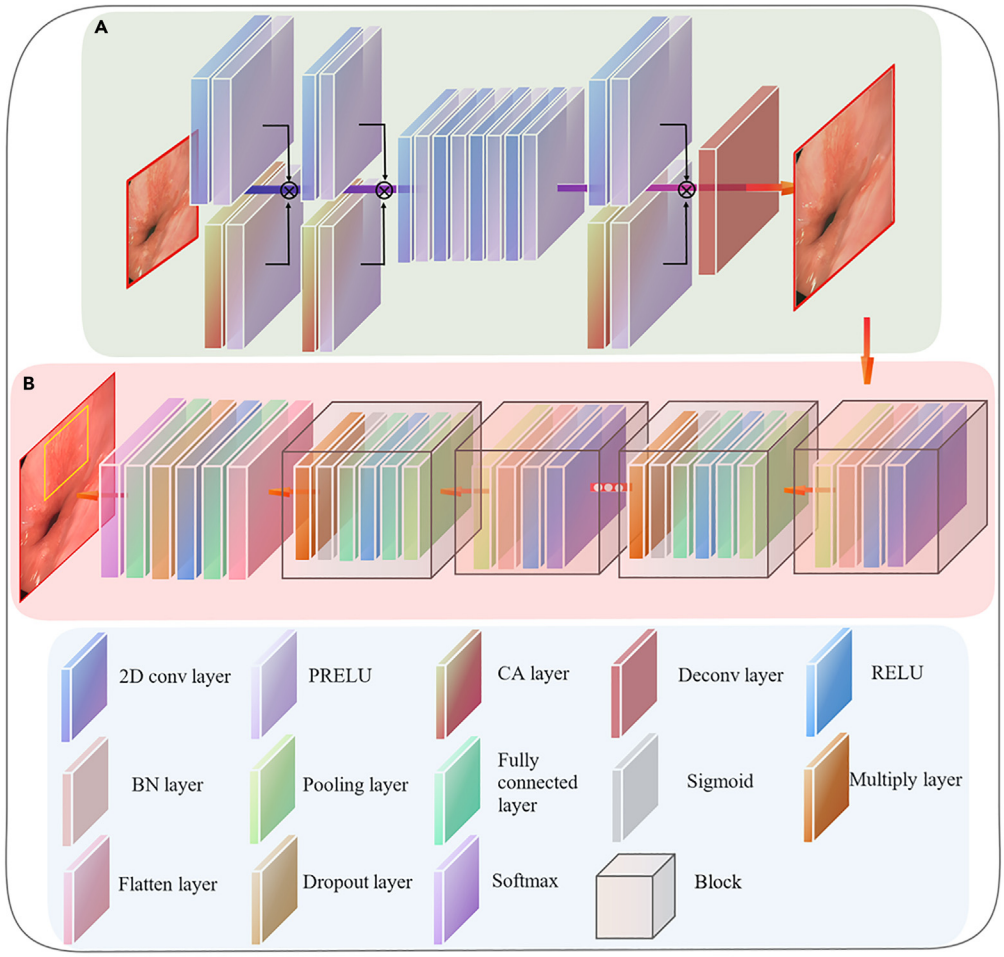


Figure 2. Overall framework of visual modality

(A) Overall architecture for SR pipeline.

(B) Overall architecture for DI pipeline.

the loss function in the smooth area can be described as:

$$Loss_2 = MSE \quad \text{(Equation 2)}$$

the loss function of the final joint optimization can be described as:

$$Total\_Loss = Loss_1 + Loss_2 \quad \text{(Equation 3)}$$

then the deduction result is described as:

$$Total_{Loss} = (1 + W_1)\left[\underset{\theta}{Min}\ \frac{1}{n}\Sigma_{i=1}^{n}\left\|F_1\left(I_{real}{}^{L};\theta\right) - I_{real}{}^{L}\right\|_2^2\right] + W_2 * \left[\underset{\theta}{Min}\ \frac{1}{n}\Sigma_{i=1}^{n}\left|F_1\left(I_{real}{}^{L};\theta\right) - I_{real}{}^{L}\right|\right] \quad \text{(Equation 4)}$$

where W_1 represents the weight of MSE in the texture area and W_2 represents the weight of L1 in the texture area, $I_{real}{}^{L}$ and $I_{real}{}^{T}$ are the i-th LR and Ground Truth (GT) image pair, and F_1 ($I_{real}{}^{L}$; θ) is the network output for $I_{real}{}^{L}$ with parameters θ.

For the real-time DI pipeline in Figure 2B, it is mainly composed of 4 conv blocks, 4 attention blocks and 1 compound blocks. Here each conv block includes RELU function following 1 conv layer with 3 × 3 kernel, a BatchNormalization layer and a Maxpooling layer. The attention block includes an Average Pooling layer, 2 Fully connected (FC) layers, a Multiply layer following a Sigmoid function. We created a dataset

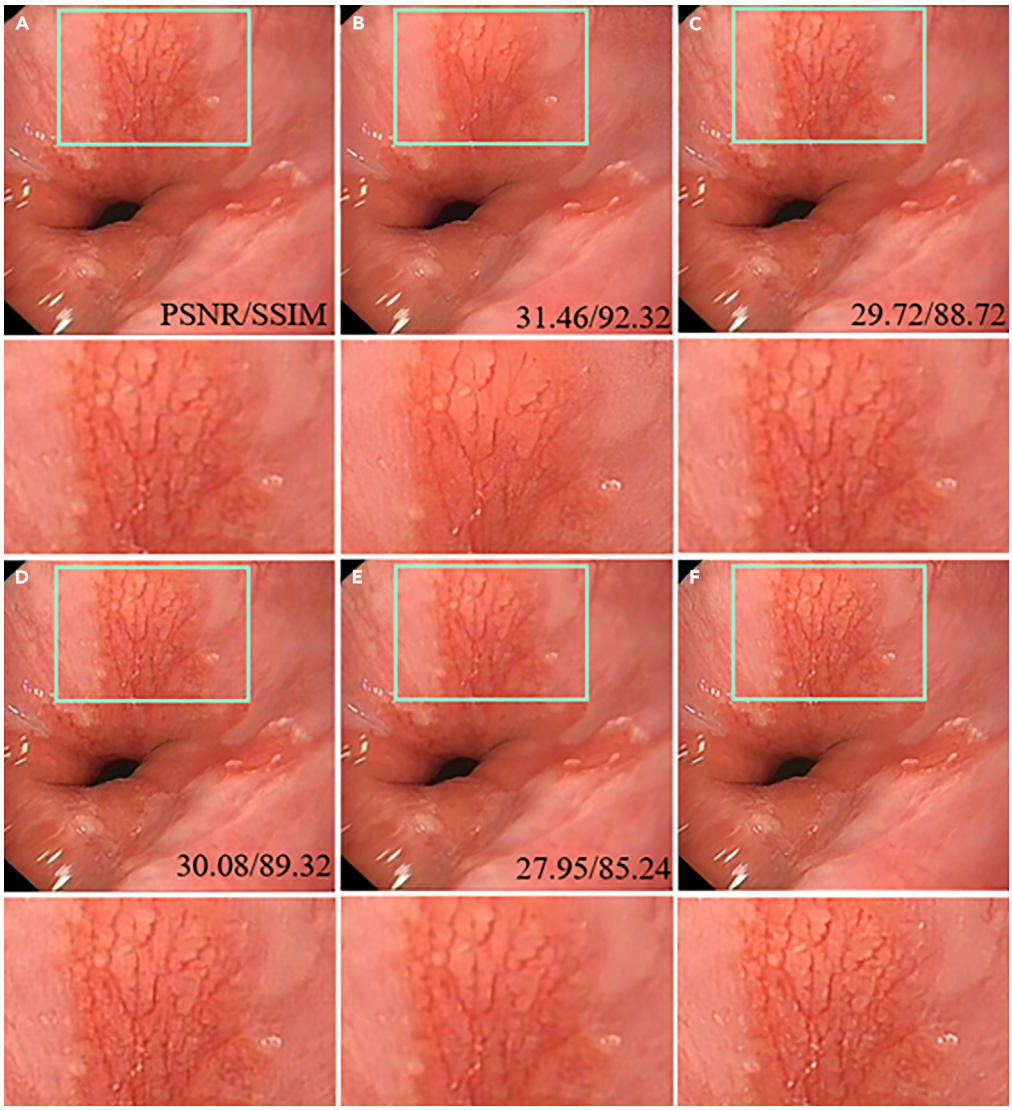


Figure 3. Comparison-SR results with DI

(A) Bicubic.
(B) Ours.
(C) EDSR.[14]
(D) FALSR.[15]
(E) ESPCN.[16]
(F) Ground Truth.

including 1,500 images with gastric-fundus-type BE, cardiac-type BE and intestinal-type BE. Our DI pipeline aims to identify and label the lesion area of intestinal-type BE. In this pipeline, the Cross-Entropy H_t is utilized as the loss function:

$$H_t = \Sigma_{i=1}^{n} I_{real}{}^{T} \log F_2\left(I_{real}{}^{S}; \theta\right) \quad \text{(Equation 5)}$$

where $I_{real}{}^{T}$ represents GT value, and $F_2\left(I_{real}{}^{S}; \theta\right)$ is the network output for SR input $I_{real}{}^{S}$ with parameters θ. Our work is performed on a PC platform (Intel Core i5-8600K CPU @3.6GHz + GTX1070) equipped with Windows10 operating system. For the model implementation and training, the baseline architecture is based on Fast Super-Resolution Convolutional Neural Networks (FSRCNN)[12] and Coordinate attention.[13] The model uses Adam optimizer for parameter optimization, with a learning rate of 10^{-3}, epochs of 200 and an upscaling factor of 4×.

Figure 3 illustrates the comparison-SR results with DI, the extensive experimental indicates that our method significantly outperforms other state-of-the-art methods. Benefiting from the attention mechanism and the joint training of multiple optimization functions in

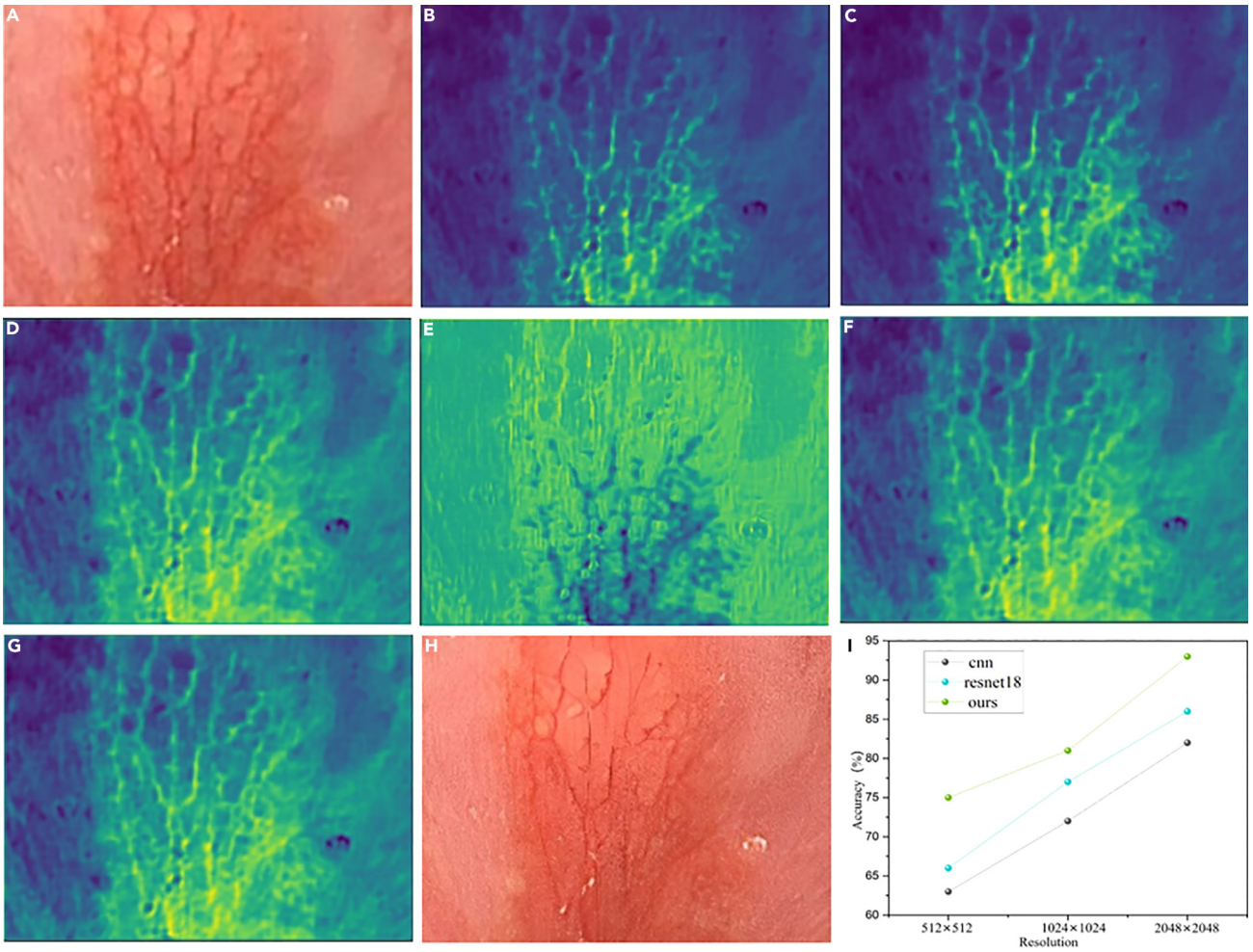


Figure 4. Visual maps of the calculation process and the logical relationship between SR and DI

(A) Input.

(B–G) Visual maps of the calculation process.

(H) SR result.

(I) Impact of various resolution on accuracy.

different regions, our experimental result takes the lead. In clinical BE, the blessed performance of SR and DI technology is crucial, for instance, the area reconstructed by SR is exactly the labeled DI area. As shown in Figure 3, the texture details are discovered using our pipeline. However, the details of insignificant flat areas can be ignored. Inspired by biomimetic principle, the proposed pipeline places more attention on the label area and ignores interference from flat areas. In addition, quantitative analysis of the Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM) was also recorded in Figure 3, and our method shows competitive advantages over classical SR methods. In this study, a dataset of 1,500 images including intestinal-type BE and other-type patients was collected from 846 patient populations included in the study according to different pathological subtypes (Table 1). The data were split in a ratio of 7:2:1 between training/validation/testing. In the visual modality, SR pipeline is responsible for converting LR images into high-resolution (HR) images that preserve texture details; the result of the SR pipeline is processed as input by DI pipeline to obtain the final result. The outcome accuracy for disease identification only achieves 83.6% without SR assistance; however, our proposed vision modality achieves accuracy at 94.1%.

Figure 4 shows the visual maps of the model calculation process and the logical relationship between SR and DI. Figures 4A–4H illustrate the visualization map of each layer from the network. Different colors indicate different levels of attention on the network, here yellow represents higher attention and blue represents lower attention. According to our findings, the pipeline intentionally avoids interference from flat areas to maximize the utilization of texture information; this hypothesis is lent credence by visual maps in Figure 4. As illustrated in Figure 4I, the result implies the DI task can benefit from the SR task. The accuracy increases almost linearly with the increase of resolution, which is also

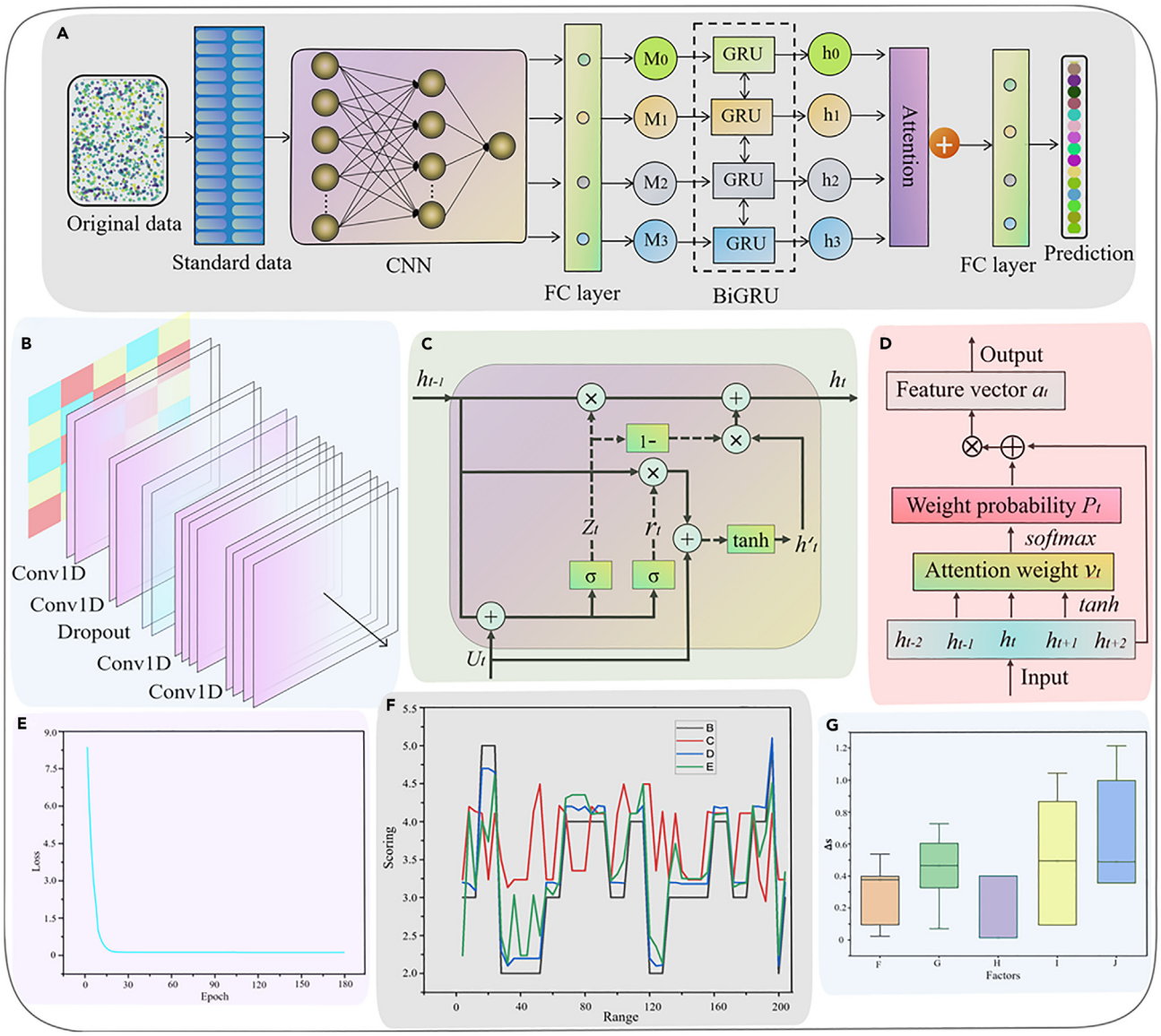


Figure 5. Overall structure and result of the RG pipeline

(A) Procedure of the RG pipeline.

(B) CNN component.

(C) BiGRU component.

(D) Attention component.

(E) Test loss.

(F) Comparison-prediction result with various methods.

(G) Boxplot of the risk grading calculated using RG pipeline.

applicable to different classification networks such as CNN and Resnet18.[17] This result proves our previous conjecture, which benefits our clinical BE.

Conclusions

Inspired by the cooperation between human vision and cerebral cortex in data processing and the principle of attention, here we describe a multimodal learning framework to tackle tasks from various modalities, which can benefit from each other. In the visual modality, we developed 2 distinct yet collaborative real-time pipelines (SR and DI) that effectively filter out irrelevant smooth regions and focus on specific texture details. Additionally, we investigated the significant impact of different resolutions on real-time DI. In the data modality, we designed

a real-time data pipeline (RA) to handle high-level tasks (Figure 5). We predicted the score of future health assessments and employed an ablation design to explore the impact of multiple risk factors on health, thereby maximizing the clinical performance of medical professionals based on our findings.

Limitations of the study

In this study, we found that resolution cannot be infinitely improved. When a certain critical condition is reached, the best advantage of mutual benefit between the two pipelines can achieve dynamic equilibrium.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Computational details
- QUANTIFICATION AND STATISTICAL ANALYSIS

ACKNOWLEDGMENTS

This work was sponsored by National Natural Science Foundation of China under grant No. 61927809 and 61975139.

AUTHOR CONTRIBUTIONS

S.B.L., L.L., and S.Y.P. conceived the project. M.X.Z. proposed the principle; S.B.L. and Z.Y.W. collected the data and conducted the experiments; S.B.L. designed the algorithms and analyzed the data; the authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: June 21, 2023
Revised: July 10, 2023
Accepted: November 9, 2023
Published: November 14, 2023

REFERENCES

1. Gong, K.D., Lu, R., Bergamaschi, T.S., Sanyal, A., Guo, J., Kim, H.B., Nguyen, H.T., Greenstein, J.L., Winslow, R.L., and Stevens, R.D. (2023). Predicting intensive care delirium with machine learning: model development and external validation. Anesthesiology *138*, 299–311.
2. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA. Cancer J. Clin. *71*, 209–249.
3. Wong, M.C.S. (2021). et al. Global burden, risk factors, and trends of esophageal cancer: an analysis of cancer registries from 48 countries. Cancers *13*.
4. Spechler, S.J., and Souza, R.F. (2014). Barrett's esophagus. N. Engl. J. Med. *371*, 836–845.
5. Jankowski, J.A., Harrison, R.F., Perry, I., Balkwill, F., and Tselepis, C. (2000). Barrett's metaplasia. Lancet *355*, 203–208.
6. Sharma, P., McQuaid, K., Dent, J., Fennerty, M.B., Sampliner, R., Spechler, S., Cameron, A., Corley, D., Falk, G., Goldblum, J., et al. (2004). A critical review of the diagnosis and management of Barrett's esophagus: the AGA Chicago Workshop. Gastroenterology *127*, 310–330.
7. Trebeschi, S., Van Griethuysen, J.J.M., Lambregts, D.M.J., Lahaye, M.J., Parmar, C., Bakers, F.C.H., Peters, N.H.G.M., Beets-Tan, R.G.H., and Aerts, H.J.W.L. (2017). Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. Sci. Rep. *7*, 5301.
8. Hirasawa, T., Aoyama, K., Tanimoto, T., Ishihara, S., Shichijo, S., Ozawa, T., Ohnishi, T., Fujishiro, M., Matsuo, K., Fujisaki, J., and Tada, T. (2018). Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer *21*, 653–660.
9. Horie, Y., Yoshio, T., Aoyama, K., Yoshimizu, S., Horiuchi, Y., Ishiyama, A., Hirasawa, T., Tsuchida, T., Ozawa, T., Ishihara, S., et al. (2019). Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointest. Endosc. *89*, 25–32.
10. Forrest, I.S., Petrazzini, B.O., Duffy, Á., Park, J.K., Marquez-Luna, C., Jordan, D.M., Rocheleau, G., Cho, J.H., Rosenson, R.S., Narula, J., et al. (2023). Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. Lancet *401*, 215–225.

11. Xin, R., Zhang, J., and Shao, Y. (2020). Complex network classification with convolutional neural network. Tsinghua Sci. Technol. *25*, 447–457.
12. Dong, C., Loy, C., and Tang, X. (2016). Accelerating the super-resolution convolutional neural network. Proc. Eur. Conf. Comput. Vis. 391–407.
13. Xie, C., Zhu, H., and Fei, Y. (2021). Deep coordinate attention network for single image super-resolution. IET Image Process. *16*, 273–284.
14. Lim, B., Son, S., Kim, H., Nah, S., and Lee, K.M. (2017). Enhanced deep residual networks for single image super-resolution. IEEE Conf. Comput. Vis. Pattern Recognit. 136–144.
15. Chu, X., Zhang, B., Ma, H., Xu, R., and Li, Q. (2020). Fast, accurate and lightweight super-resolution with neural architecture search. IEEE, 59–64.
16. Shi, W., Caballero, J., Huszár, F., Totz, J., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proc. ICCV, 1874–1883.
17. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition (IEEE).
18. Chen, T.Q., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Organisms/strains		
Age (years)	<50	193
	≫50	653
Gender	Male	530
	Female	316
Marriage	Married	790
	Divorced	29
	Widowed	23
	Unmarried	4
Degree of education	Bachelor's degree or above	233
	Junior college	247
	High school	101
	Junior high school	138
	Primary school and below	127
Others	This paper	N/A
Critical commercial assays		
Olympus-290	Olympus	N/A
CV-290	Olympus	N/A
CLV-290SL	Olympus	N/A
GIF-HQ290	Olympus	N/A
OEV262H	Olympus	N/A
WM-NP2	Olympus	N/A
Software and algorithms		
Pycharm	JetBrains	https://www.jetbrains.com/pycharm/download/?section=window/
Python	Open-source software	https://www.python.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lea contact, Leili (leili@scu.edu.cn).

Materials availability

This study did not generate new unique materials.

Data and code availability

- All data reported in this paper is available within the paper.
- The original code in this paper is available from the lead contact upon reasonable request.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon reasonable request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

846 patients from our hospital were randomly assigned to the experimental group.All experimental procedures were approved by College of Sichuan University.

METHOD DETAILS

Computational details

With the screening of medical risk factors in the previous section, the final risk factors were confirmed. Here the RG pipeline is created to predict the risk values and risk grading according to various risk factors, as illustrated in Figure 5. For non-numeric data, such as gender, professional medical personnel use numerical coding to assign values to risk factors. The overall structure of RG pipeline is illustrated in Figure 5A, which mainly includes a CNN module, a bidirectional gated recurrent unit (BiGRU) module, an attention module and multiple Fully connected (FC) modules. As illustrated in Figure 5B, CNN module includes 4 1D convolution layers and 2 maxpooling layer, which is to extract local features with small dimension. Here the size of 1D convolution kernel is set to 2 and its number is 32, stride = 1, padding = 0. The FC modules are introduced to improve the nonlinearity of the pipeline, and its output is computed by iterating from 1 to T using:

$$U_{(x)} = \sigma\left(W_f^T x + b\right) \quad \text{(Equation 6)}$$

where W_f and b are the weight matrix and bias vectors respectively for the hidden layer, and σ is activation function. BiGRU is a recurrent network unit that excellent at capturing key information of time series for either long or short term. The main structure of GRU includes an update gate and a reset gate. The architecture of the GRU cell is illustrated in Figure 2C. The logic of GRU cell is as follows:

$$r_t = \sigma(W_r U_t + U_r h_{t-1} + b_r) \quad \text{(Equation 7)}$$

$$z_t = \sigma(W_z U_t + U_z h_{t-1} + b_z) \quad \text{(Equation 8)}$$

$$h'_t = tanh(W_h U_t + U_h(r_t h_{t-1}) + b_h) \quad \text{(Equation 9)}$$

$$h_t = z_t h_{t-1} + (1 - z_t) h'_t \quad \text{(Equation 10)}$$

where σ is the sigmoid function and tanh is the hyperbolic tangent function. *Ur, z* are the weight matrixes for the previous vector h_{t-1}, and h'_t is a candidate activation. Vector r_t, z_t denote the reset gate and the renew gate vector.

The BiGRU includes the forward GRU and the reverse GRU. The forward GRU generates a forward feature vector sequence $\{\overrightarrow{h_1} \cdots \overrightarrow{h_t}\}$, while the reverse GRU generates a reverse feature vector sequence $\{\overleftarrow{h_1} \cdots \overleftarrow{h_t}\}$. Hence, the final feature vector sequence *ht* can be computed using:

$$\text{ht} = \beta t \overleftarrow{h_t} + \alpha t \overrightarrow{h_t} + bt \quad \text{(Equation 11)}$$

where β_t is the output weight of information for backward propagation GRU unit at time *t*, α_t is the output weight of information for forward propagation GRU unit at time *t*, and b_t is the corresponding offset.

Attention architecture imitates how human brain calculates information, which is of great significance for the improvement of prediction performance. Multi head attention is introduced to solve the problem of the proportion of input vector. The data that contribute more is given a greater proportion, which is of great significance to the improvement of performance accuracy. Multi head attention in Figure 5D illustrates the detailed update process as follows:

$$v_t = \sigma(h_t) \quad \text{(Equation 12)}$$

$$P_t = \frac{\exp(v_t)}{\sum_{k=1}^{m} \exp(v_k)} \quad \text{(Equation 13)}$$

$$a_t = \sum_{t=1}^{m} P_t h_t \quad \text{(Equation 14)}$$

where σ is activation function tanh, h_t is the feature vector from BiGRU module, softmax function generates probability vector P_t, and a_t is the generated attention vector.

The test loss in Figure 5E is continuously iterated and updated until it remains stable, proving that this method has good robustness and high accuracy. In Figure 5F, the RG pipeline outperforms competitive advantage over the comparison methods, here B-GT; C-CNN; D-RG; E- XGBoost.[18] As illustrated in Figure 5G, the final risk factors are confirmed which are Age (I), Gender (F), BMI (G), Acid reflux/heartburn (J) and Family history of esophageal cancer (H), respectively. Continuous ablation experiments are specially designed to verify the specific impact of different risk factor on the evaluation value. To our findings, J accounts for the largest proportion, followed by I, followed by G, F, and H, as illustrated in Figure 5G. This has to some extent inspired medical personnel to pay more attention to J's clinical manifestations, paving the way for modern medicine. In the whole multimodal framework, the end-to-end optimization benefits the two modalities, and here the Prompt idea is used to achieve high-precision prediction of the output of BE, and the output results of visual modals and data modalities need to be unified into the same dimension. For example, features are extracted from images and extracted from parameters, and then they

are fused into a unified feature vector, and then the two feature vectors are weighted and stitched into a unified feature vector to achieve the final assessment according to the final feature vector parameters. The combined formula can be described as:

$$Out = \gamma f(\cdot) + \delta g(\cdot) \quad \text{(Equation 15)}$$

where $f(\cdot)$ is the function that converts the image to a vector representation, and $g(\cdot)$ is the function that converts the argument to a vector representation. γ and δ are weight matrices that are used to weight the representation of images and parameters.

Relatively speaking, our method has been validated based on images and data information from 846 patients, but the acquisition of datasets remains a challenge. In the future, multi-center studies will be considered to address the problem of insufficient samples and achieve wider clinical validation.

QUANTIFICATION AND STATISTICAL ANALYSIS

Use SPSS 26.0 statistical software for data organization and statistical analysis. Age, BMI, and other measurement data are represented by means and statistically described. Clinical symptoms, disease complications, and other counting data are compared and analyzed using the Perason chi square test and Fisher's exact probability method. The significance of differences was tested using one-way ANOVA. Then, single factor meaningful inclusion in multivariate logistic regression analysis further clarifies independent risk factors. $p < 0.05$ is considered statistically significant.