Leveraging Multimodal Semantic Fusion for Gastric Cancer Screening via Hierarchical Attention Mechanism

Shuai Ding, *Member, IEEE*, Shikang Hu, Xiaojian Li, Youtao Zhang, *Member, IEEE*, and Desheng Dash Wu, *Senior Member, IEEE*

***Abstract*—Gastroscopy is a widely adopted method for locating gastric lesions and performing the early screening and diagnosis of gastric cancer (GC). However, the effectiveness of traditional GC screening methods depends on the medical skills of the gastroscopy specialist. A lack of knowledge and experience may lead to misdiagnosis and mistreatment, especially in small-scale hospitals. Recently, there has been a significant increase in studies on data-driven computer-aided diagnosis techniques. In this article, we propose a novel intelligent decision-making method for GC screening (ID-GCS), a multimodal semantic fusion-based data-driven decision-making system. ID-GCS exploits a hybrid attention mechanism to extract textual semantics from multimodal gastroscopy reports and performs semantic fusion to integrate the semantics of textual gastroscopy reports and images, resulting in improved interpretability of gastroscopy findings. We evaluated ID-GCS using a real gastroscopy report dataset, and experimental results show that compared with state-of-the-art methods, ID-GCS achieves better sensitivity and accuracy in GC screening.**

***Index Terms*—Gastric cancer (GC) screening, hybrid attention, interpretability, multimodal fusion.**

I. Introduction

GASTRIC CANCER (GC) is one of the most common causes of cancer deaths globally and therefore a major threat to public health. According to the statistics on cancer incidence and mortality from the International Agency for Research on Cancer, 1.03 million GC cases were reported in 2018, making GC the fifth-most common cancer worldwide. Additionally, GC accounted for 783 000 (third-highest) cancer-related global deaths in 2018; that is, approximately 89 people die from GC every hour. While the clinical practice has shown that the five-year survival rates for progressive and advanced GC are often less than 30% [1], appropriate surgical procedures and treatment regimens, such as endoscopic mucosal dissection and resection can prolong the life of 90% of patients when they are diagnosed at early stages [2]. However, the global early diagnostic rate of GC is unsatisfactory, reaching less than 20% in some developing counties [3]. The reasons for this include the following.

1) The symptoms of early-stage GC are not specific, making it difficult to distinguish GC from other chronic gastric diseases, such as gastritis and gastric ulcers.
2) It is challenging to improve the clinical diagnostic accuracy for GC due to a shortage and uneven distribution of experienced gastroscopy specialists.
3) A shortage of medical funds limits the worldwide adoption of novel gastroscopic diagnosis technologies [4].

Advanced information technologies [e.g., big data and artificial intelligence (AI)] and clinical assistant decision-making systems have been developing rapidly in recent years. Studies have shown that the exploitation of large-scale electronic medical records (EMRs) and various types of medical imaging data are beneficial for developing data mining, assistant analysis, intelligent diagnosis, and decision support systems to enhance nursing and treatment, lesion detection, disease risk prediction, emergency healthcare, and health management in clinical practice [5]–[9]. The adoption of novel information technologies has also greatly improved GC screening. For example, Ishihara *et al.* [10] constructed an intelligent gastroscopic image analysis system for clinical GC detection based on X-ray gastroscopic images, which exhibited higher GC detection performance than physicians. Oikawa *et al.* [11] proposed a computer-aided pathology system that performs a rapid, automated histological analysis based on histological section image data and can be used to detect GC and obtain automated pathological diagnoses. Garcia *et al.* [12] proposed a deep convolutional neural network (CNN)-based model for the automated detection of lymphocytes in GC immunohistochemistry images, achieving a high accuracy of over 96%.

Because gastroscopic images are objective depictions of gastric mucosal lesions that contain their most comprehensive

Manuscript received April 24, 2020; revised January 9, 2021 and April 10, 2021; accepted June 29, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 91846107, Grant 71571058, and Grant 71690235; in part by the Anhui Provincial Science and Technology Major Project under Grant 17030801001 and Grant 18030801137; and in part by the Fundamental Research Funds for the Central Universities under Grant PA2019GDQT0021. This article was recommended by Associate Editor Z. Liu. *(Corresponding author: Shuai Ding.)*

Shuai Ding and Xiaojian Li are with the School of Management and the Key Laboratory of Process Optimization and Intelligent Decision-Making (Ministry of Education), Hefei University of Technology, Hefei 230009, Anhui, China (e-mail: dingshuai@hfut.edu.cn; lixj90@hfut.edu.cn).

Shikang Hu is with Alibaba Search Algorithm Team, Hangzhou, Zhejiang, China (e-mail: huzhu_hfut@163.com).

Youtao Zhang is with the Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15213 USA (e-mail: zhangyt@cs.pitt.edu).

Desheng Dash Wu is with the School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: dwu@ucas.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2021.3096974.

Digital Object Identifier 10.1109/TSMC.2021.3096974

feature information, existing clinical decision support systems mainly rely on imaging data and rarely exploit data from other modalities, such as textual gastroscopy reports. Gastroscopy reports contain descriptions and subjective summaries of the manifestation, shape, and location of gastric lesions from gastroscopy specialists. Due to the extensive clinical experience and specialized knowledge of these specialists, gastroscopy report texts contain valuable information and thus should not be overlooked in GC screening. Moreover, as a clinical procedure, GC screening needs to be scientifically rigorous and interpretable; however, prevalent clinical decision support systems often treat the decision-making process as a "black box" due to their reliance on machine learning methods, especially deep learning. Therefore, they are often unable to provide an understanding of the rationale behind the prediction results obtained between data input and diagnosis output and hence are also unable to provide the interpretability required in evidence-based medicine.

In this article, we propose an intelligent decision-making method for GC screening (ID-GCS), a data-driven decision-making methodology based on multimodal semantic fusion. Our main contributions are summarized as follows.

1) We propose ID-GCS, a method based on multimodal semantic fusion. Utilizing a deep neural network with a hierarchical attention mechanism, ID-GCS integrates text and image semantics from multimodal gastroscopy reports. The method learns and fuses lesion feature representations in order to potentially improve the sensitivity and accuracy of clinical GC screening.
2) We propose a hierarchical attention mechanism within ID-GCS that merges semantic-level and feature-level attention patterns to fuse cross-scale semantic information. The mechanism employs a hybrid attention structure [i.e., textual semantic-based GC screening (TextGCS)] integrating CNN and gated recurrent unit (GRU) attention patterns. TextGCS captures "local patterns" resulting from specific writing formats in gastroscopy reports and global semantic representations, which enhances textual semantic extraction for cancer screening.
3) We verify the effectiveness of ID-GCS using a real-world dataset consisting of 8713 GC screening reports. The experimental results show that ID-GCS achieves significant improvements over state-of-the-art methods in terms of sensitivity and accuracy. We further demonstrate the interpretability of ID-GCS by visualizing the feature-level attention mechanisms and the classification activation maps on gastroscopy reports.

The remainder of this article is organized as follows: Section II surveys the relevant literature, including the application of computer-aided diagnosis (CADx) systems in cancer screening, research on the interpretability of neural networks, and the application of attention mechanisms in the natural language processing and medical image analysis. In Section III, we present the proposed ID-GCS method. In Section IV, we use a real-world medical dataset to verify the effectiveness of ID-GCS and its interpretability. Finally, we summarize and conclude the paper in Section V. In the Appendix, we add some additional comparative experiments and parameter settings.

II. Literature Review

A. Application of Computer-Aided Diagnosis Systems in Cancer Screening

In recent years, machine learning techniques have been widely adopted in CADx to assist medical experts in disease diagnosis [13]–[16]. Together with advances in AI technology, data-driven CADx systems have been extensively applied in the research of a variety of cancers (e.g., breast cancer, pancreatic cancer, esophageal cancer, lung cancer, and prostate cancer). Walczak and Velanovich [17] constructed a CADx system for predicting the survival rate of patients with pancreatic cancer that helps physicians and patients select the most satisfactory treatment regimens and guides clinical decisions. Relying on the analysis of histopathology images, Swager *et al.* [18] and Wang *et al.* [19] analyzed pathogenesis at the cell level and achieved accurate diagnosis of esophageal and prostate cancers, respectively. Jiao *et al.* [20], Setio *et al.* [21], and Quellec *et al.* [22] used AI technologies to improve the efficiency and accuracy of the clinical diagnosis of breast and lung cancer by detecting lesions from X-ray images, magnetic resonance images, and CT images, respectively. In 2017, Esteva *et al.* developed an AI-based CADx system that achieved a diagnostic accuracy of 95% in skin cancer screening, which is comparable to that of medical experts [9]. Zhang *et al.* [23] established multiple mapping relationships between pathological images and diagnostic reports using their proposed MDnet. This network can read the images, generate the corresponding diagnostic reports as well as attention areas and provide the corresponding visual interpretability for the diagnosis of bladder cancer. This ability has inspired further research in this field. In 2019, Zhu *et al.* [24] constructed a CNN computer-aided detection system based on endoscopic images that effectively predicts the invasion depth of the tumor and was shown to be conducive to screening patients for endoscopic resection. Specifically for GC screening, Li *et al.* [25] proposed a new multi-instance learning algorithm that extracts the features of two layers from the original monochrome CT images and constructed a CADx system for identifying the invasion depth of GC tumors. Kanesaka *et al.* [26] constructed a CADx system for identifying and distinguishing early GC from magnified-narrow band imaging images. Wang *et al.* [27] proposed an AdaBoost-based multicolumn CNN and a connected electronic gastroscopy system for intelligent and dynamic GC screening.

However, most existing CADx systems exploit unimodal medical data, with the majority focusing on imaging data. As a result, the large amounts of clinically available medical records and unstructured textual reports are not effectively utilized.

B. Interpretability of Neural Network

Model interpretability has emerged as a vital issue, both theoretically and practically, in the study of neural network-based CADx systems. Recent studies have focused on the representations and learning interpretability of deep neural networks. Interpretability studies in the computer vision field have included in the following.

1) Feature visualization of intermediate network layers: by successively visualizing the feature map of each intermediate convolutional layer, the features and abstract logic of a neural network can be better understood [28], [29].
2) Discovery and detection of image regions that directly contribute to the network output [30]–[33], such as inverse estimation of the planar regions of an image based on the gradient loss in the propagation feature map. Yang *et al.* [34] proposed two complementary frameworks that automatically determine the most contributive regions of the input scenes. Wang *et al.* [35] proposed a novel text-image embedding network (TieNet) that integrates a multilevel attention model into an end-to-end trainable CNN-recurrent neural network (RNN) architecture to highlight meaningful text and image regions.
3) Analysis of filters, for example, analytical graphs [36] and trees [37]. In natural language processing-related fields, interpretability primarily refers to the detection of important objects or key regions and the analysis of keywords or text segments that make important contributions to the decision-making process of the network. Recent studies have introduced attention mechanisms based on coder and decoder structures [38].

Research on the interpretability of neural networks is also important for analyzing the internal decision logic of neural network models, understanding the causal relationship between the input and output of the networks, and enhancing model reliability. Interpretable models mitigate the doubts of medical practitioners toward their performance so that these methods are more likely to be deployed in medical-related fields.

C. Application of Attention Mechanism in Natural Language Processing and Medical Image Analysis

Cognitive science and neuroscience recognize attention as a complex cognitive function. An individual can unconsciously choose to pay more attention to one part of information while ignoring other parts, focusing more on the key parts or the important content of interest and less on the other parts. Attention mechanisms have been widely used in computer vision [39], [40], natural language processing [41]–[43], and speech recognition [44], [45]. In medical-related fields, attention mechanisms have also been widely used in the following areas.

1) *Medical Image Segmentation:* A previous study [46] proposed a sparse annotation strategy based on attention-guided active learning for three-dimensional (3-D) medical image segmentation. The attention mechanism is used to improve segmentation accuracy and estimate the segmentation accuracy of each slice. Another study [47] proposed a novel semisupervised image segmentation method that simultaneously optimizes supervised segmentation and an unsupervised reconstruction objective. In [48], a volumetric attention (VA) model was proposed for 3-D medical image segmentation and detection.
2) *Medical Image Analysis:* A previous study [49] proposed a novel attention gate (AG) model for medical image analysis that automatically learns to focus on target structures of varying shapes and sizes. Wang *et al.* [50] proposed a novel, deep CNN, called Thorax-Net, for diagnosing 14 thorax diseases using chest radiography images and achieved good performance in experiments.
3) *Prediction and Diagnosis of Disease Risk based on EMR Data:* A previous study [51] proposed a novel approach, the attention-based multi-instance neural network (AMINet), to classify single diseases-based only on existing and valid information in real-world outpatient records. Another study [52] proposed deep attention models to predict the onset of high-risk vascular diseases based on the symbolic history of patients with hypertension. In [53] an attention-based cross-modal CNN (AXCNN) was introduced for predictive analytics exploiting EMRs. Utilizing both patient historical diagnostic records and patient demographics, another study [54] proposed a medical context attention (MCA)-based RNN.

III. ID-GCS Methodology

This section presents an overview of the proposed ID-GCS framework and then elaborates on its components in detail.

A. Framework of ID-GCS

Fig. 1 presents an overview of the ID-GCS framework, which integrates feature- and semantic-level attention mechanisms, including a text semantic extraction network (TextGCS) and a visual semantic extraction network (GCSNet). The framework consists of the following two steps.

1) *Semantic Extraction:* The hybrid attention-integrated TextGCS is designed to extract textual semantics from gastroscopy reports. Specifically, this process includes word embedding and extraction and fusion of attention using the local template matching attention of a CNN (CNN-attention), the sequence correlation attention of a GRU (GRU-attention), and hybrid attention weight-based textual semantic representations. GCSNet, consisting of four convolutional layers and three pooling layers, is used to extract visual semantics from gastroscopic images.
2) *Multimodal Fusion and Cancer Inference:* The ID-GCS framework adopts a semantic-level attention mechanism to integrate multimodal semantics (e.g., the semantics of gastroscopy report texts and gastroscopic images). Knowledge is generated based on the integrated multimodal semantic information to provide decision-making support for clinical GC screening.

To improve the interpretability of the ID-GCS framework, we adopted visualization methods to display the internal relationship between the GC screening output and the data input (the gastroscopy report texts and the gastroscopic images). The

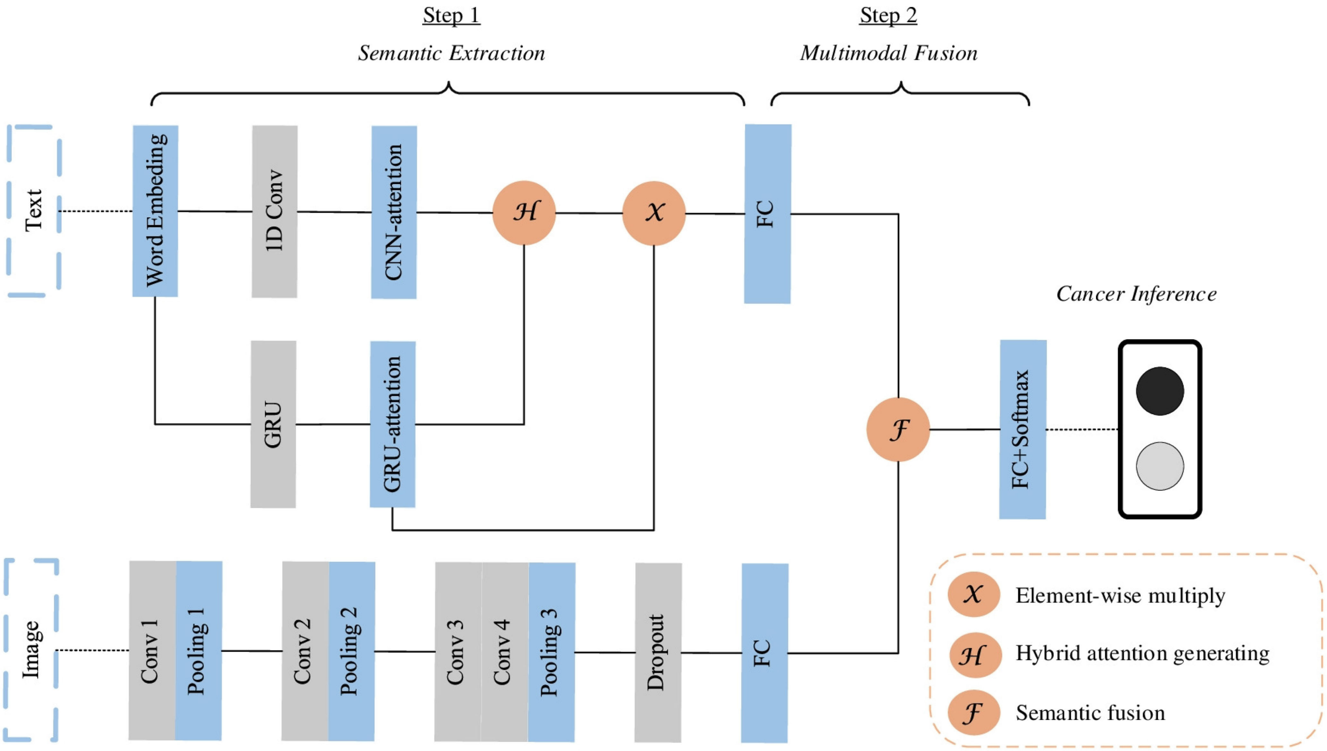


Fig. 1. Framework of ID-GCS.

attention-weighted visualization of the semantics of the gastroscopy report texts helps to highlight the words and text segments that are crucial to the diagnosis output. For the gastroscopic images, a classification activation map is used to visualize the key pixels, which helps detect key lesion regions and improve model interpretability.

B. Textual Semantic Extraction by TextGCS

A clinical gastroscopy report usually includes the described location, shape, and manifestation of lesions, all valuable information for GC screening. Physicians tend to follow certain wording styles when preparing gastroscopy reports, and the locations of textual expressions of important features are often close in context. As a result, internal "local patterns" are often observed in these reports. This manifests as a high co-occurrence frequency of certain Chinese characters, and we believe that this internal pattern can be captured by a CNN. As a summary of the examination, a gastroscopy report conveys all the semantics that should be considered. We use a GRU to extract the semantics of the gastroscopy report text from the entire gastroscopy report. Based on the uniqueness of the gastroscopy report data, we present a novel feature-level hybrid attention mechanism that integrates CNN-attention and GRU-attention for extracting the key semantic representations from gastroscopy reports. Fig. 2 shows the steps of TextGCS. Here, $W_t(t = 1.4)$ denotes the tth word in the original gastroscopy report, and ϕ denotes the zero padding, which ensures the same dimensions regardless of the input for further 1-D convolution.

First, the input text sequence is divided into words using the reverse maximum matching algorithm. Second, the input text sequence is converted to dense word vectors using the word embedding technique. Third, a hybrid attention model is exploited for the input sequence to combine CNN-attention and GRU-attention into feature-level hybrid attention. Finally, the final text semantic representation is generated using the feature-level hybrid attention to assign attention weights to the vector representations in the GRU hidden layers.

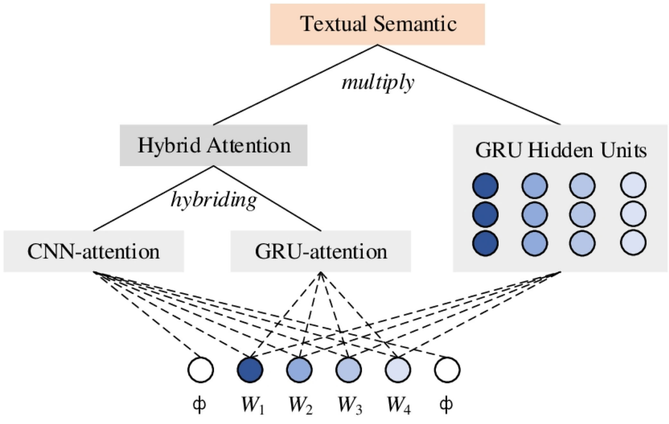


Fig. 2. Structure of TextGCS.

1) Word Embedding: Computers can perform only numerical calculations; therefore, for an input text sequence, a word embedding technique is used to embed the raw semantics of words or phrases, which enables them to be read into memory and calculated numerically. In this method, a dictionary is established based on the text corpus of the EMR, and the index location of each word or phrase in the dictionary is employed as its representation. Using a word embedding layer, the words or phrases in the report text are converted to low-dimensionality, dense vector representations as follows:

$$V = C \cdot W \tag{1}$$

where $V \in R^d$ is the low-dimensionality word vector representation, $W \in R^{|v|*d}$ is the word embedding matrix, and $C \in R^{|V|}$ is the high-dimensionality vector representation of length $|V|$.

2) CNN-Based Attention Extraction: Next, TextGCS exploits a CNN to extract the attention information from the gastroscopy report and capture the internal "local patterns" resulting from specific writing formats and rules in the report text. Generally, a gastroscopy report text sequence of length T is represented as

$$x_{1:T} = x_1 \oplus x_2 \oplus x_3 \cdots \oplus x_T \tag{2}$$

where $x_t \in R^d (t = 1, 2, ..., T)$ is the d-dimensional representation of the tth word in the text sequence. Let $w \in R^{d*l}$ denotes the filter of 1-D convolution operation and l denotes the length of the text sequence. The attention value of tth word can be computed by the convolution operation as follows:

$$c_t = f\left(w^T x_{t:t+l-1} + b\right) \tag{3}$$

where $x_{t:t+l-1}$ denotes a subsequence of length l in the text sequence of the original gastroscopy report and b denotes a bias. Considering the sparse activation and the possible gradient disappearance, the ReLU activation function is employed to conduct the convolution operation in TextGCS as follows:

$$f\left(w^T x_{t:t+l-1} + b\right) = \begin{cases} 0, & w^T x_{t:t+l-1} + b \leq 0 \\ w^T x_{t:t+l-1} + b, & w^T x_{t:t+l-1} + b > 0. \end{cases} \tag{4}$$

Each filter w can output the attention value c_t from the text sequence after convolution operation. Here, TextGCS employs multiple filters to ensure the stability of the attention caption to reduce noise interference in the computation of the CNN-attention value. Specifically, TextGCS uses multiple filters to capture the attention signals of the original report text and computes the average of the outputs of these filters as the stable attention $c_t \in R^T$ as follows:

$$c_t = \frac{1}{m} \sum_{i=1}^{m} c_t{}^i \tag{5}$$

where m is the number of filters and $c_t{}^i$ is the attention signal generated by the ith filter by a convolution operation. Each element c_t represents the attention weight at the corresponding location in the original text sequence.

3) GRU-Based Attention Extraction: The gastroscope report text is a global, holistic description of gastroscopy that includes the valuable semantics to be considered. In this study, to globally model the overall attention of the gastroscope report text sequence, we use the sequence correlation of the GRU. The hidden-state vector h_t encoded by the GRU is used as the semantic representation of the tth time step and can be obtained by

$$h_t = (1 - z_t) \cdot h_t^{'} + z_t \cdot h_{t-1} \tag{6}$$

where z_t is an update gate that determines whether input x_t at the current moment is disregarded. In other words, it can be used to determine whether the input x_t at the current moment is important to the semantic representation of the entire sequence input. We obtain z_t using a sigmoid activation function. Here, h_{t-1} is the hidden-state vector of the preceding moment input x_{t-1}, and $h_t^{'}$ is the candidate hidden-state vector jointly generated by the current moment input x_t and the hidden unit of the preceding moment h_{t-1}.

The update gate z_t and $h_t^{'}$ can be obtained

$$\begin{cases} z_t = \text{sigmoid}(W_{xz} x_t + W_{hz} h_{t-1}) \\ h_t^{'} = \tanh(W_{xh} x_t + r_t \cdot (W_{hh} h_{t-1})) \end{cases} \tag{7}$$

where $r_t = \text{sigmoid}(W_{xr} x_t + W_{hr} h_{t-1})$ is a reset gate that is used to control the impact of the hidden-state vector h_{t-1} at the preceding moment on the input x_t at current moment and can be obtained by a sigmoid activation function. The impact of the input x_t at the current moment on the semantic representation of the sequence is adjusted by adjusting its value. Finally, W_{xz}, W_{hz}, W_{xr}, W_{hr}, W_{xh}, and W_{hh} represent the parameter matrix.

In this study, the GRU model is used to encode the gastroscopy report text to obtain a vector representation of the report text sequence. Additionally, the attention distribution in the gastroscopy report sequence and the textual semantic representation are obtained by attention modeling.

GRU-attention is obtained by

$$a_t = \frac{\exp\left(b_t{}^T \cdot U\right)}{\sum_t \exp\left(b_t{}^T \cdot U\right)} \tag{8}$$

where a_t (i.e., the attention weight of the input at the current time step t) is the GRU-attention value of h_t obtained via normalization, and U is the correlation matrix, which can be used to calculate the importance (the attention weight) of h_t. The initial value of U is obtained by the network upon initialization and is updated as the network is trained, whereas b_t can be obtained by a layer from the multilayer perceptron as follows:

$$b_t = \tanh\left(w^T \cdot h_t + b_w\right) \tag{9}$$

where W is the parameter matrix and b_w is the offset vector.

4) Attention Fusion and Textual Semantic Representation: The preceding text describes the way in which CNN-attention and GRU-attention are obtained from the gastroscopy report sequence. The hidden-state vector of the GRU code is applied by TextGCS as the vector representation of the sequence input at each time step. The attention weights extracted by the CNN and those calculated based on the hidden-state vector code of the GRU are averaged to determine the feature-level hybrid attention, which is integrated with the hidden-state vector of the GRU at each time step based on the attention weight. Thus, the final textual semantic vector of the gastroscopy report input is computed as follows:

$$v = \sum_t \frac{(a_t + c_t)}{2} \cdot h_t \tag{10}$$

where c_t is the convolutional attention weight of the gastroscopy report text sequence at time step t, a_t is the attention weight of the hidden-state vector code of the GRU at time step t, h_t is the hidden-state vector representation of the GRU at each time step, and v is the final textual semantic vector representation of the gastroscopy report.

5) Interpretability of the Gastroscopy Report: In this study, the attention mechanism introduced in the extraction of the semantics of the gastroscopy report is designed not only to measure the influence of specific words on the overall semantic expression of the gastroscopy report but also to quantify the contribution of each word to the cancer inference, helping to identify words or texts with key contributions. If TextGCS receives a text sequence of length T, then for a gastroscopy report sequence of length $T^{'}$, a total of Δt "empty" items are added to ensure that the sequence inputs have the same length, as follows:

$$T = T^{'} + \Delta t. \tag{11}$$

Let $x_{1:\Delta t}$ denotes the "empty" input items with a dimensionality of $\Delta t * d$, let $x_{\Delta t+1:T}$ denotes the gastroscopy report sequence input, and let d denotes the dimensionality of the word embedding. The attention weight distribution of the sequence $x_{1:T}$ is calculated using the previously discussed hybrid attention mechanism. In this study, the attention weight of the gastroscopy report text sequence input $x_{\Delta t+1:T}$ is converted to multiples of β with the average attention weight β of $x_{1:\Delta t}$ as the reference. Subsequently, the gastroscopy report text sequence input $x_{\Delta t+1:T}$ is visualized based on the attention weight in the form of color levels. Specifically, β is obtained by

$$\beta = \frac{1}{\Delta t} \sum_{\Delta t} \frac{(a_t + c_t)}{2} \tag{12}$$

where a_t is the GRU-attention of the gastroscopy report text sequence at the time step t, and c_t is the CNN attention of the gastroscopy report text sequence at the time step t.

This method conspicuously highlights the words and text blocks that have a more significant role in the GC screening process and the prediction logic of the ID-GCS methodology and renders them interpretable.

C. Visual Semantic Extraction by GCSNet

1) Architecture of GCSNet: Gastroscopic images reveal the surface condition of the gastric mucosa. Gastric mucosal lesions often exhibit relatively simple characteristics, such as redness, whitening, eminence, and ulceration, and lack other, relatively complex characteristics. Existing, pretrained network architectures, such as AlexNet, Visual Geometry Group network (VGGNet), residential neural network (ResNet), and LeNet, are overly complex models. Although they are powerful in extracting features, due to the uniqueness of gastroscopic image data and the relative lack of reported data, these networks are extremely prone to overfitting in GC screening. By improving the adaptability of AlexNet, a network architecture with relatively low complexity, GCSNet is designed to extract semantics from gastroscopic images. The structure of GCSNet consists of four convolutional layers, three pooling layers, and a fully connected layer, as shown in Fig. 3.

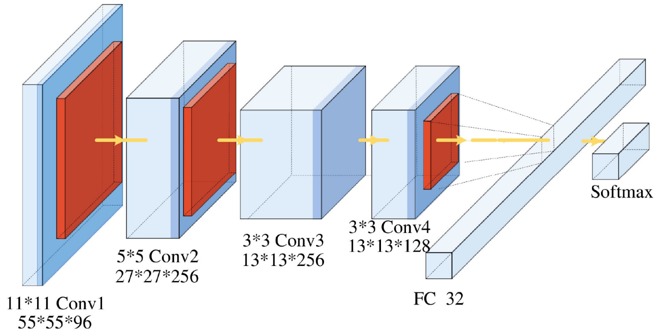


Fig. 3. Structure of GCSNet.

1) *Input Layer:* Gastroscopic images that differ in size are converted to images with dimensions of 224*224*3.
2) *Convolutional Layer C1:* A total of 96 11*11*3 convolution kernels are used to perform convolution operations on the gastroscopic image input. Each convolution kernel has a step size of four pixels.
3) *Pooling Layer P1:* A maximum pooling operation with a step size of 2 and a window size of 3 is performed on each feature map from the C1 layer.
4) *Convolutional Layer C2:* Similar to C1, convolution operations with 256 5*5*96 convolution kernels are performed, and 256 27*27 feature maps are obtained.
5) *Pooling Layer P2:* A maximum pooling operation with a step size of 2 and a window size of 3 is performed on each feature map from the C2 layer.
6) *Convolution Layer C3:* A total of 256 3*3*256 convolution kernels are used to perform convolution operations obtained from the P2 layer.
7) *Convolution Layer C4:* A total of 128 3*3*256 convolution kernels are used to perform convolution operations on the feature maps outputted from the C3 layers.
8) *Pooling Layer P3:* A maximum pooling operation with a step size of 2 and a window size of 3 is performed on each feature map from the C4 layer.
9) *Fully Connected Layer F1:* The F1 layer immediately precedes the output layer. All the feature maps from the P3 layer are expanded and connected to form a 1-D eigenvector, which serves as the input for the F1 layer. The F1 layer is a fully connected layer that consists of 32 neurons, the gastroscopic image semantic layer discussed in this study.

The preprocessed gastroscopic image data are trained on the designed GCSNet. The vector outputted from the last fully connected layer of GCSNet is applied as the image semantic vector $v \in R^d$ of the gastroscopic image data.

2) Interpretability of the Gastroscope Image: This section describes the use of Grad-CAM [30] to visually display the gastroscopic image inputs. For a given gastroscopic image input, the gradient of each channel in each feature map, which shows the GC outputted from the last convolutional layer in ID-GCS, is calculated, and a weight is assigned to each channel in the feature map based on the gradient. Thus, an activation heat map is obtained for the GC regions of the original gastroscopic image. The class activation heat map assists in

Algorithm 1 Multi-ATT

```
Input: each mode's semantic representation $v$
Output: final multimodal semantic vector representation $V$
Initialize the parameters:
parameter matrices $W$
offset vector $v_w$
correlation matrix $U_b$
for $v_t \in v$ do
    calculate $vb_t$ by (14)
end for
for $vb_t$ do
    calculate $va_t$ according by correlation matrix $U_b$ and formulation (13)
end for
calculate $V$ by formulation (15)
return $V$
```

comprehending which part of a gastroscopic image contributes to the model's final classification decision and highlights the image regions that have a particularly significant role in the GC screening process, contributing to ID-GCS interpretability.

D. Multimodal Semantic Fusion

In practice, physicians tend to attach different levels of importance to different modes of medical data when analyzing multisource, multimodal EMR data. The different modes of EMR data are assumed to contribute differently to cancer inference. Based on these assumptions, we propose a multimodal semantic fusion method that uses a semantic-level attention mechanism to perform a weighted fusion of the semantics of various modes by assigning them different attention weights. The multimodal semantic fusion method, named Multi-ATT, is illustrated in Algorithm 1.

Let v denote the sequence of the semantic vector representation, v_t denote the semantic vector of the tth modal medical data, va_t denote the semantic-level attention value of v_t (i.e., the gastroscopic image semantic vector and gastroscopy report textual semantic vector), and V denote the final multimodal semantic vector representation.

The data of each mode are treated as a sequence input. The hidden-layer vector at each time step is the semantic vector of the corresponding mode of data. The attention of each mode of medical data can be obtained by

$$va_t = \frac{\exp\left({vb_t}^T \cdot U_b\right)}{\sum_t \exp\left({vb_t}^T \cdot U_b\right)} \tag{13}$$

where U_b is the correlation matrix, which can be used to calculate the total importance of semantic vector v_t (i.e., the gastroscopic image semantic vector and gastroscopy report textual semantic vector) to the entire multimodal sequence input. The initial value of U_b is obtained during network initialization and updated as the network is trained, and vb_t can be obtained from a layer of the multilayer perceptron as follows:

$$vb_t = \tan\left(W^T \cdot v_t + v_w\right) \tag{14}$$

where W is the parameter matrix and v_w is the offset vector.

By weighted fusion of the semantic-level attention obtained from the calculation and the semantic vector of each mode of EMR data based on the attention, the final semantic vector is obtained, as shown in

$$v = \sum_t (va_t \cdot v_t) \tag{15}$$

where va_t is the attention weight of each mode of semantic representation and V is the final multimodal semantic vector representation obtained after fusion.

Finally, a softmax layer that contains two neurons is added to the end of the semantic extraction networks TextGCS and GCSNet, which outputs the GC screening result.

In summary, ID-GCS effectively extracts the modal semantic representations and performs multimodal semantic fusion, which significantly improves the sensitivity of GC screening, as shown in our experiments.

IV. Experiments

To evaluate the effectiveness of ID-GCS, we compared it with state-of-the-art methods used in other experimental studies. Next, we present the dataset and experimental settings and then describe the evaluation indicators to evaluate different algorithms. We studied textual semantic extraction, visual semantic extraction, and multimodal semantic fusion, and we exploited semantic visualization to verify the interpretability of ID-GCS.

A. Data Description and Experimental Environment

The experimental data used in this study were gastroscopy report texts and gastroscopy images acquired from December 2016 to September 2017 by the First Affiliated Hospital of Anhui Medical University, People's Republic of China. The gastroscopy images include eight parts of the body, including the esophagus, cardia, gastric body, fundus of the stomach, angle, antrum, pylorus, and duodenum. The gastroscopy reports include basic patient information, the gastroscopy findings, and the diagnosis. After the data preprocessing step, we obtained a dataset consisting of 8713 data samples, which included 2274-GC samples and 6439 noncancer samples. The sample distribution for different age groups and sexes is shown in Fig. 4. For the comparative analysis, the experimental data were divided into a training set, a verification set, and a test set at a ratio of 6:2:2 (5227, 1743, and 1743 samples, respectively). We used the training set to train the model and the validation set to validate the model and optimize the model parameters, improving its performance. The test set was used to evaluate the effect of the final model and to design experimental tables and experiments.

The experiments were carried out on a server with an Intel Xeon CPU (E5-2620, 2.10 GHz), 128 GB of memory, and 4 NVIDIA GeForce Titan X GPUs.

B. Metrics

To validate and evaluate the proposed method in this experiment effectively, the specificity, sensitivity, test accuracy,

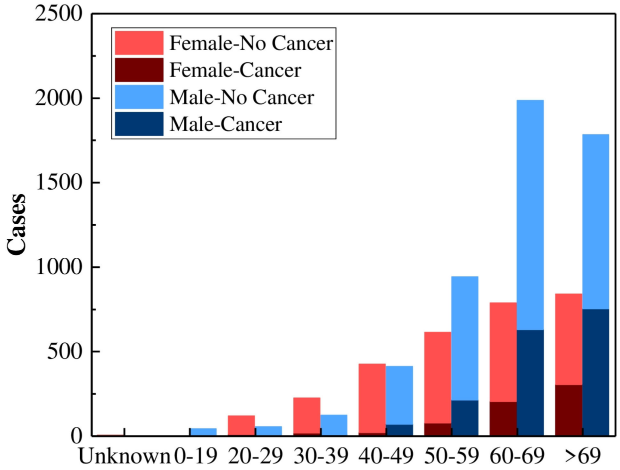


Fig. 4. Cases distribution for different age groups and genders.

F1-score, and area under the receiver operating characteristic (ROC) curve (AUC) were selected as performance metrics. The first four metrics are defined as follows:

$$\begin{cases} \text{Specificity} = \frac{|\text{TN}|}{|\text{TN}|+|\text{FP}|} \\ \text{Sensitivity} = \frac{|\text{TP}|}{|\text{TP}|+|\text{FN}|} \\ \text{Accuracy} = \frac{|\text{TP}|+|\text{TN}|}{|\text{TP}|+|\text{TN}|+|\text{FP}|+|\text{FN}|} \\ F1 - \text{Score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}} \end{cases} \quad (16)$$

where TP (true positive) refers to the set of patients who fall into the positive class and are correctly classified, FN (false negative) refers to the set of patients who fall into the positive class but are misclassified as negative, and TN (true negative) and FP (false positive) are similarly defined as Recall = |TP|/(|TP| + |FN|) and Precision = |TP|/(|TP| + |FP|). ROC curves are typically used in binary classification to study the output of a classifier, typically by plotting the true positive rate on the *Y*-axis and the false positive rate on the *X*-axis.

C. Baseline

To verify the validity of the text semantic extraction network TextGCS, visual semantic extraction network GCSNet, and multimodal semantic fusion method Multi-ATT based on a semantic-level attention mechanism, we chose the following methods and models in the fields of text classification, image classification, and multimodal data fusion for comparison.

For the comparative text classification experiment, we chose TextCNN, TextRNN, and RNN+attention as the baseline models. The TextCNN model is an algorithm proposed by Kim in 2014 for CNN-based text classification. The TextRNN model, which generally consists of word embeddings, consists of RNN encoding, concatenate, fully connected, and softmax layers; it has a flexible structure and can effectively solve the vanishing gradient and gradient explosion problems. The RNN+attention model introduces an attention mechanism that assigns higher weights to words that affect the classification results, thereby achieving excellent outcomes. In this study, long short-term memory (LSTM) and GRU were employed as the RNN units in the attention-based text classification experiments.

TABLE I
MULTIMODAL SEMANTIC FUSION METHOD COMPARISON

Model	ACC	Spec	Sen	AUC	F1-Score
GMU	0.9145	0.9195	0.9016	0.9760	0.8552
Concatenate	0.9122	0.9100	0.9180	0.9738	0.8541
AVG	0.9122	0.9171	0.8996	0.9742	0.8516
ID-GCS	0.9443	0.9490	0.9324	0.9804	0.9037

We chose the classical models AlexNet, ResNet, and VGG-16 as the baseline models for image classification. We trained 50, 101, and 152 layers of neural networks using ResNet units and pretrained the weights. ResNet won the first place in the 2015 ImageNet large-scale visual recognition challenge (ILSVRC), and it has fewer parameters than VGGNet, which simplifies the optimization required to achieve high accuracy.

We then chose the following three semantic fusion methods to compare with our multimodal semantic fusion technique.

1) Concatenation [35], [55], which converts the vector representations of multiple semantic modes into a new eigenvector representation by concatenation.
2) The gated multimodal unit (GMU) [56], which assigns weights to multimodal semantic information via a gated network and integrates multimodal semantics to obtain better feature representation.
3) Average (AVG), which is a decision-level fusion strategy that obtains a fused multimodal semantic representation by the equal-weighted fusion of the semantic vectors of all modes.

In the comparison of the multimodal fusion methods, TextCNN and VGG-16 were used to extract the sematic representations from the gastroscopy report texts and gastroscopy images, respectively.

D. Results and Discussion

In this study, classifiers were formed by adding an output layer containing two neurons to the end of existing semantic extraction network architectures, and they were compared with the baseline models selected in Section IV-C. In Section IV-D1, the selected semantic fusion methods are compared with the proposed multimodal semantic fusion model. In Section IV-D2, the selected textual semantic extraction methods are compared with the proposed multimodal fusion semantic model. In Section IV-D3, the selected visual semantic extraction methods are compared with the proposed multimodal fusion semantic model.

1) Comparisons of Multimodal Semantic Fusion Models: To evaluate the performance of the proposed ID-GCS in GC screening, we first compared ID-GCS with the TextGCS method and the visual semantic-based GC screening method (GCSNet); that is, we compared the results from the proposed unimodal and multimodal data fusion methods. We then compared ID-GCS with three semantic fusion models: 1) concatenation; 2) GMU; and 3) AVG.

Table I summarizes the comparison results from the different fusion methods. The accuracy of ID-GCS was 3.26%–3.52% higher than that of the GMU, concatenation, and AVG models. The sensitivity was 1.57%–3.65% higher than that of

TABLE II
COMPARISON OF TEXT-ORIENTED MODEL RESULTS

Model	ACC	Spec	Sen	AUC	F1-score
TextCNN	0.9094	0.9307	0.8545	0.9647	0.8407
TextRNN	0.9048	0.9307	0.8381	0.9596	0.8313
GRU+ATT	0.9099	0.9203	0.8831	0.9626	0.8459
LSTM+ATT	0.9082	0.9195	0.8790	0.9633	0.8428
TextGCS	0.9111	0.9203	0.8873	0.9654	0.8482
ID-GCS	0.9443	0.9490	0.9324	0.9804	0.9037

TABLE III
COMPARISON OF THE IMAGE-ORIENTED MODELS

Model	ACC	Spec	Sen	AUC	F1-score
AlexNet	0.8417	0.8733	0.7602	0.888	0.7288
ResNet152	0.8532	0.8827	0.7473	0.8919	0.7966
VGG-16	0.8579	0.8763	0.7832	0.8697	0.7971
GCSNet	0.8692	0.8916	0.8115	0.9173	0.7765
ID-GCS	0.9443	0.9490	0.9324	0.9804	0.9037

the other models, indicating that ID-GCS could significantly reduce the missed diagnosis rate of GC by more than 32.6% compared with the other models. The specificity of ID-GCS was 3.21%–4.29% higher than that of the other models, the AUC value was 0.45%–0.68% higher than that of the other models, and the F1-score was 5.67%–6.12% higher than that of the other models.

The superior results of ID-GCS over those of the other methods indicate its substantial advantage in GC screening. Compared with the concatenate and AVG multimode fusion methods, Multi-ATT, based on the attention mechanism, effectively recognizes the importance of the modal data for cancer screening. Given a certain weight, Multi-ATT can maximize the extracted modal semantics for GC screening and obtain good experimental results.

By using gating mechanisms and semantic integration, the GMU can distinguish the importance of multimode semantics. However, the semantic weight learning of the GMU relies only on the sigmoid activation function, and it is not sufficient for learning the importance of the semantics of each mode, potentially leading to information loss; therefore, it is slightly inferior to ID-GCS.

2) Comparisons of the Textual Semantic Extraction Models: Table II summarizes the comparison results for ID-GCS, TextCNN, TextRNN, LSTM-ATT, and GRU-ATT and shows that except for ID-GCS, all models exhibited relatively similar total accuracy and AUC values, with differences within 1%. The accuracy of ID-GCS was 3.64–4.37% higher than that of the other models, the F1-score was 6.54–8.71% higher than that of the other models, and the sensitivity was 5.08-11.25% higher than those of the other models. The ID-GCS method effectively reduced the possibility of missed GC diagnose by more than 55%.

The comparison shows that the proposed ID-GCS method is superior to TextCNN, TextRNN, LSTM-ATT, and GRU-ATT in all aspects. This is because TextGCS is designed to exploit gastroscopy reports by effectively capturing their inherent local and global patterns, achieving better expression of the semantics of the text of the gastroscopy reports. For cancer-related semantic capture, the important semantics within the short context of gastroscopy reports can be better identified by members responsible for the local semantic capture ability of the model. The ID-GCS method integrates multiple modes of data, that is, the gastroscopy report text and gastroscope images, and thus better captures the GC features with richer semantic information. Moreover, the analysis of our experimental results shows that the models with attention mechanisms performed better than those without, especially for sensitivity. This is because the models with an attention mechanisms assign larger weights to important features, and thus, they tend to give a better representation of the semantics of the report text.

3) Comparisons of the Visual Semantic Extraction Models: Table III summarizes the results from different models. The accuracy of ID-GCS was 8.64%–12.19% higher than that of the AlexNet, ResNet, and VGG models. The AUC value of ID-GCS was 6.88%–12.73% higher than that of the baseline models. The sensitivity of the ID-GCS method exhibited a notable advantage, with a value that was 14.90%–22.77% higher than that of the other models; that is, ID-GCS reduced the probability of a missed GC diagnosis by 72.9% with respect to the other models. In terms of the F1-score, the ID-GCS method was 13.27%–24.00% better than the baseline models.

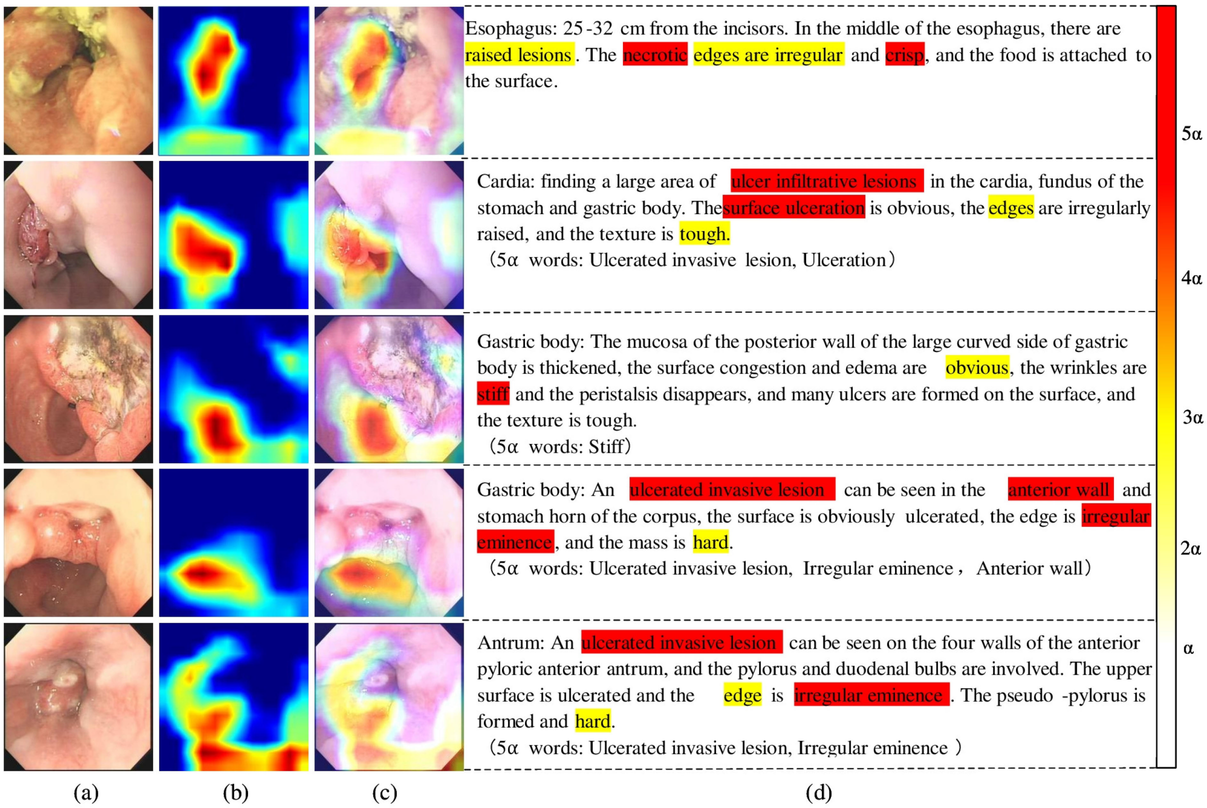
These comparative experimental results show that the proposed ID-GCS method exhibited better overall performance than the baseline models in terms of accuracy, specificity, sensitivity, AUC value, and F1-score. Compared with existing, pretrained network architectures, such as VGG, AlexNet, and ResNet, the ID-GCS method had slightly worse complexity and fitting. As a neural network designed mainly for analyzing gastroscopy images, GCSNet is effective in extracting imaging features and in identifying the key features associated with GC. Moreover, the low complexity of ID-GCS makes overfitting difficult. In summary, ID-GCS is a well-designed framework for GC screening.

E. Visualization of Semantics

To examine the presence of key regions in gastroscopy images and important contributing words in the gastroscopy report text that have a decisive impact on the diagnosis output, the visualization of the gastroscopy report text and images is discussed in this section. This visualization highlights the text blocks and image regions that contribute substantially to the model output to determine the prediction logic between the input data and the diagnostic conclusion. Images of five GC lesions and the report text for the corresponding regions were selected and visualized. The visualized features are shown in Fig. 5.

Importantly, the text input data of the gastroscopy report used in this article is written with Chinese characters; we have translated it to English to facilitate understanding of the results.

1) *Visualization of Important Regions in Gastroscopic Images:* An activation heat map is plotted based on the feature maps from the last convolutional layer. Grad-CAM [30] was used to visualize the key regions of



Esophagus: 25-32 cm from the incisors. In the middle of the esophagus, there are raised lesions . The necrotic edges are irregular and crisp, and the food is attached to the surface.

Cardia: finding a large area of ulcer infiltrative lesions in the cardia, fundus of the stomach and gastric body. Thesurface ulceration is obvious, the edges are irregularly raised, and the texture is tough.
（5α words: Ulcerated invasive lesion, Ulceration）

Gastric body: The mucosa of the posterior wall of the large curved side of gastric body is thickened, the surface congestion and edema are obvious, the wrinkles are stiff and the peristalsis disappears, and many ulcers are formed on the surface, and the texture is tough.
（5α words: Stiff）

Gastric body: An ulcerated invasive lesion can be seen in the anterior wall and stomach horn of the corpus, the surface is obviously ulcerated, the edge is irregular eminence, and the mass is hard.
（5α words: Ulcerated invasive lesion, Irregular eminence， Anterior wall）

Antrum: An ulcerated invasive lesion can be seen on the four walls of the anterior pyloric anterior antrum, and the pylorus and duodenal bulbs are involved. The upper surface is ulcerated and the edge is irregular eminence . The pseudo -pylorus is formed and hard.
（5α words: Ulcerated invasive lesion, Irregular eminence ）

Fig. 5. Focus of visualization of semantic prediction in ID-GCS: (a) raw images, (b) strongest feature map in last conv-layer, (c) overlay of original image and heat map, and (d) endoscopy findings with key contributing words.

the gastroscopic images. Fig. 5 shows the visualization of important pixels and regions in the gastroscopic images. Fig. 5(a) shows the original gastroscopic images. Fig. 5(b) shows the activation heat maps that correspond to the GC type, which reveals the key regions of the gastroscopic images that display important information. Fig. 5(c) shows the attention alert maps. This visualization allows the better detection of key lesions in gastroscopy images.

2) *Visualization of the Attention Weights of Key Text Segments:* In this study, the feature-level hybrid attention mechanism calculates the attention distribution for a gastroscopic report text sequence input. When constructing an assistive decision model for GC, an "empty" input is added to ensure that the sequence inputs have the same length. Therefore, the added "empty" input will be allocated a certain amount of attention during the calculation. When visualizing gastroscopy report text, each sequence input is visualized in the form of color levels with the average attention β of "empty" sequence inputs as the threshold. The average attention β can be obtained by (12). Specifically, words with an attention weight higher than 5β are highlighted in red (i.e., 5β words); words with an attention weight higher than 3β are highlighted in yellow; and relatively unimportant words (words with an attention weight lower than β) are not highlighted, as shown in Fig. 5(d). This visualization helps to highlight the presentation of lesions directly related to GC tumors in the report text so that it could be more easily detected by doctors, potentially improving the sensitivity of GC screening.

TABLE IV
COMPLETE EXPERIMENT RESULTS

Model	ACC	Spec	Sen	AUC	F1-score
AlexNet	0.8417	0.8733	0.7602	0.888	0.7288
ResNet50	0.9088	0.9530	0.7951	0.9503	0.8299
ResNet101	0.9094	0.9538	0.7951	0.9419	0.8308
ResNet152	0.8532	0.8827	0.7473	0.8919	0.7966
VGG-16	0.8579	0.8763	0.7832	0.8697	0.7971
GCSNet	0.8692	0.8916	0.8115	0.9173	0.7765
TextCNN	0.9094	0.9307	0.8545	0.9647	0.8407
TextRNN	0.9048	0.9307	0.8381	0.9596	0.8313
GRU-ATT	0.9099	0.9203	0.8831	0.9626	0.8459
LSTM-ATT	0.9082	0.9195	0.8790	0.9633	0.8428
FastText	0.9122	0.9410	0.8381	0.9705	0.8424
TextRCNN	0.9042	0.9394	0.8135	0.9642	0.8262
TextGCS	0.9111	0.9203	0.8873	0.9654	0.8482
GMU	0.9145	0.9195	0.9016	0.9760	0.8552
Concatenate	0.9122	0.9100	0.9180	0.9738	0.8541
AVG	0.9122	0.9171	0.8996	0.9742	0.8516
Text-Resnet50	0.8979	0.9076	0.8730	0.9518	0.8272
Text-Resnet101	0.9369	0.9673	0.8586	0.9756	0.8840
ID-GCS	0.9443	0.9490	0.9324	0.9804	0.9037

An analysis of the visualized gastroscopic images and report text shows that the GC lesion regions were satisfactorily detected [Fig. 5(b)], and the descriptions of the location and shape of the lesions in the gastroscopy report were assigned relatively high attention values. This finding is consistent with

TABLE V
DETAILED PARAMETER SETTINGS

Model	Parameters
ResNet50 ResNet101	Optimizer=SGD; learning rate=0.01; early stopping=yes; epochs=200
ResNet152	Optimizer=SGD; learning rate=0.01; decay=0.0001; early stopping=yes; epochs=500
VGG-16	Optimizer=SGD; learning rate=0.01; decay=0.0001; early stopping=yes; epochs=500; dropout=0.5
AlexNet	Optimizer=SGD; learning rate=0.01; decay=0.0001; early stopping=yes; epochs=500; dropout=0.7
GCSNet	Optimizer=SGD; learning rate=0.01; decay=0.0001; early stopping=yes; epochs=500; dropout=0.8
TextCNN	Filter sizes=[2, 3, 4, 5]; optimizer=Adam; learning rate=0.001; early stopping=yes; epochs=500; dropout=0.5
TextRNN	RNN cell: LSTM; optimizer=Adam; learning rate=0.001; early stopping=yes; epochs=200
GRU-ATT LSTM-ATT FastText TextRCNN	Optimizer=Adam; learning rate=0.001; early stopping=yes; epochs=200
TextGCS	Optimizer=Adam; learning rate=0.001; early stopping=yes; epochs=500
GMU Concatenate AVG Text-Resnet50 Text-Resnet101 ID-GCS	Optimizer=Adam; learning rate=0.0001; early stopping=yes; epochs=300

clinical experience and objective reality and demonstrates the interpretability of the proposed ID-GCS methodology. In clinical practice, the visualization of key areas within gastroscopic images and key report texts can remind doctors to focus on the corresponding areas and assist them in clinical decision-making. The doctor is the main decision-maker in this process, and the visualization of gastroscopy images and report text can be used as an aid. In this article, we make an interpretable effort through visualization, which is beneficial for analyzing the explanatory evidence in the two modalities, but our effort fails to use the semantic alignment method to mine and present the relationship between the two modalities. In the future, we plan to perform semantic alignment, which should help solve this problem.

V. CONCLUSION

In this article, we proposed ID-GCS, an intelligent decision-making method for GC screening based on multimodal semantic fusion. ID-GCS captures the clinical thinking and diagnostic approaches of physicians and integrates objective medical data and subjective experiential knowledge obtained from gastroscopy reports for GC screening. By adopting the hybrid attention visualization and Grad-CAM visualization technology, ID-GCS highlights image areas and text segments that are critical in the diagnosis of GC and illustrates the predictive rationale between the input data and the diagnostic results. Furthermore, the improved interpretability of ID-GCS helps to meet the requirements of evidence-based medicine. Our experimental results show that ID-GCS achieves significant improvements over state-of-the-art methods in GC screening.

As part of our future research, we intend to address the following limitations.

1) The overall performance of GCSNet for GC screening is lower than that of TextGCS, which indicates that GCSNet can be further improved.
2) The patients whose data were studied in this article were mainly from Anhui Province and several surrounding cities in China. We will collect more data to validate the effectiveness of our method.
3) This article focuses on implementing an algorithm that has yet to pass clinical trials. The algorithm needs to be further customized and enhanced before it can be potentially adopted in different hospital systems.

APPENDIX

A. Complete Record of Experimental Results

We recorded all the experimental results in Table IV.

In Table IV, we additionally recorded the experimental results of the ResNet50, ResNet101, TextRCNN, fastText, Text-Resnet50, and Text-Resnet101 models.

In terms of image models, the pretrained ResNet50 and ResNet101 models were better than the ResNet152 model; therefore, on the gastroscopy image dataset in this article, networks with deeper structures are not good choices. The superiority of the pretrained ResNet50 and ResNet101 models over the ResNet152 model shows that networks with lower depth and complexity can better simulate gastroscopy image datasets. Compared with GCSNet, pretrained ResNet50 and ResNet101 were more accurate and had a higher AUC value but were less sensitive. However, sensitivity is more important in GC screening. The comparison of VGG-16, AlexNet, ResNet152, and GCSNet was discussed in the experimental part of this article and will not be repeated here.

In terms of text models, TextRCNN and fastText had no advantages over models, such as TextCNN, TextRNN, RNNAttention, and TextGCS, and indeed were slightly inferior in terms of sensitivity.

In terms of multimodal semantic fusion, we supplemented the comparative experiments of Text-Resnet50 and Text-Resnet101 by integrating TextGCS as the text semantic extraction module and Multi-ATT as the semantic fusion method Multi-ATT, but their overall performance was not as good as that of ID-GCS.

TABLE VI
GCSNET NETWORK DESIGN AND PERFORMANCE

Convolutional Layers	Fully Connected Layers	Accuracy
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.3 \rightarrow 512 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9374 \quad val \quad acc: 0.8721$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.3 \rightarrow 512 \rightarrow 0.3 \rightarrow 32 \rightarrow 2$	$acc: 0.9989 \quad val \quad acc: 0.8703$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 512 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9991 \quad val \quad acc: 0.8474$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 512 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9044 \quad val \quad acc: 0.8744$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 512 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9143 \quad val \quad acc: 0.8657$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 512 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9001 \quad val \quad acc: 0.8646$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 256 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9878 \quad val \quad acc: 0.8749$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.3 \rightarrow 256 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9280 \quad val \quad acc: 0.8738$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.3 \rightarrow 256 \rightarrow 0.3 \rightarrow 32 \rightarrow 2$	$acc: 0.9400 \quad val \quad acc: 0.8623$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 256 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.8897 \quad val \quad acc: 0.8669$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 256 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9199 \quad val \quad acc: 0.8744$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 256 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.8677 \quad val \quad acc: 0.8646$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 1024 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.8763 \quad val \quad acc: 0.8709$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.3 \rightarrow 1024 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9337 \quad val \quad acc: 0.8692$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.3 \rightarrow 1024 \rightarrow 0.3 \rightarrow 32 \rightarrow 2$	$acc: 0.9238 \quad val \quad acc: 0.8388$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 1024 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9287 \quad val \quad acc: 0.8698$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 1024 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9224 \quad val \quad acc: 0.8772$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 1024 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.8937 \quad val \quad acc: 0.8652$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.2 \rightarrow 32 \rightarrow 2$	$acc: 0.9079 \quad val \quad acc: 0.8405$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.3 \rightarrow 32 \rightarrow 2$	$acc: 0.9172 \quad val \quad acc: 0.8623$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9879 \quad val \quad acc: 0.8543$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9105 \quad val \quad acc: 0.8732$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.8 \rightarrow 32 \rightarrow 2$	$acc: 0.9377 \quad val \quad acc: 0.8698$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$	$0.2 \rightarrow 32 \rightarrow 2$	$acc: 0.9079 \quad val \quad acc: 0.8405$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.3 \rightarrow 32 \rightarrow 2$	$acc: 0.9320 \quad val \quad acc: 0.8508$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9273 \quad val \quad acc: 0.8703$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9076 \quad val \quad acc: 0.8732$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.8 \rightarrow 32 \rightarrow 2$	$acc: 0.9310 \quad val \quad acc: 0.8876$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 512 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9014 \quad val \quad acc: 0.8749$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 512 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.8941 \quad val \quad acc: 0.8646$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 512 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9268 \quad val \quad acc: 0.8503$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 512 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.8956 \quad val \quad acc: 0.8744$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 256 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9174 \quad val \quad acc: 0.8462$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 256 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9040 \quad val \quad acc: 0.8698$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 256 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.8816 \quad val \quad acc: 0.8761$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 256 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9182 \quad val \quad acc: 0.8709$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 1024 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9016 \quad val \quad acc: 0.8669$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 1024 \rightarrow 0.5 \rightarrow 32 \rightarrow 2$	$acc: 0.9106 \quad val \quad acc: 0.8789$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.5 \rightarrow 1024 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9036 \quad val \quad acc: 0.8657$
$96 \rightarrow 256 \rightarrow 384 \rightarrow 256$	$0.7 \rightarrow 1024 \rightarrow 0.7 \rightarrow 32 \rightarrow 2$	$acc: 0.9360 \quad val \quad acc: 0.8497$

B. Detailed Parameter Settings of Each Method

In this section, we show the key parameter settings for the models in Table V. As shown above and in the table, many models are considered in this article. The parameters are mainly related to the optimizer, learning rate, early termination, and iteration steps. SGD refers to the gradient descent method, Adam refers to the adaptive moment estimation optimizer, the learning rate indicates the learning rate used, early stopping indicates whether the early termination strategy was implemented, the number of epochs indicates the maximum number of iterations, filter sizes refer to the convolution kernel sizes used in the TextCNN model, dropout indicates the proportion of neurons randomly deleted during the training process, and decay refers to the decrease in the learning rate.

C. Discussion on the Depth and Width of GCSNet

Because more complex neural network models did not necessarily perform satisfactorily for cancer screening on the gastroscopy image dataset, this study used AlexNet and adaptively improved the network structure of the model, resulting in the proposed GCSNet. During network adjustment, we used the accuracy of the model with the validation dataset as a metric. The experimental results for different arrangements of convolutional and fully connected layers are shown in Table VI.

TABLE VII
CANDIDATE ARCHITECTURES

Candidate model	Convolutional Layers	Fully Connected Layers
Architecture 1	$96 \rightarrow 256 \rightarrow 256 \rightarrow 128$	$0.8 \rightarrow 32 \rightarrow 2$
Architecture 2	$48 \rightarrow 192 \rightarrow 256 \rightarrow 128$	$0.8 \rightarrow 32 \rightarrow 2$
Architecture 3	$48 \rightarrow 192 \rightarrow 384 \rightarrow 128$	$0.8 \rightarrow 32 \rightarrow 2$

The convolutional layers column represents the structural design of the convolution layers used for feature extraction in GCSNet, whereas the fully connected layers column represents the structural design after the first fully connected layer in GCSNet. Decimals between layers indicate that dropout was used. For example, the native AlexNet convolutional layer is expressed as $96 \rightarrow 256 \rightarrow 384 \rightarrow 384 \rightarrow 256$, with a unit size of 4096 and a fully connected layer of $4096 \rightarrow 4096 \rightarrow 1000$.

The experiments showed that a more streamlined network effectively suppressed the occurrence of overfitting, thereby achieving better results on the gastroscopy dataset used in this article and indicating that it is better to connect a larger dropout to the first fully connected layer. After experimental analysis of various network depths and widths, we finally identified the following three candidate structures, as show in Table VII. Architecture 1 (GCSNet) was chosen through further experiments, and its performance details are given in Section IV of this article.

References

[1] C. Allemani *et al.*, "Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): Analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries," *Lancet*, vol. 391, no. 10125, pp. 1023–1075, 2018.

[2] M. Pellise *et al.*, "Endoscopic mucosal resection for large serrated lesions in comparison with adenomas: A prospective multicentre study of 2000 lesions," *Gut*, vol. 66, no. 4, pp. 644–653, 2017.

[3] G. H. Kim, P. S. Liang, S. J. Bang, and J. H. Hwang, "Screening and surveillance for gastric cancer in the united states: Is it needed?" *Gastrointest. Endosc.*, vol. 84, no. 1, pp. 18–28, 2016.

[4] C. Hamashima and R. Goto, "Potential capacity of endoscopic screening for gastric cancer in Japan," *Cancer Sci.*, vol. 108, no. 1, pp. 101–107, 2017.

[5] B. Pourbabaee, M. J. Roshtkhari, and K. Khorasani, "Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2095–2104, Dec. 2018.

[6] Y. Chen, S. Ding, Z. Xu, H. Zheng, and S. Yang, "Blockchain-based medical records secure storage and medical service framework," *J. Med. Syst.*, vol. 43, no. 1, p. 5, 2018.

[7] J. De Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, 2018.

[8] S. Wang, X. Chang, X. Li, G. Long, L. Yao, and Q. Z. Sheng, "Diagnosis code assignment using sparsity-based disease correlation embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3191–3202, Dec. 2016.

[9] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[10] K. Ishihara, T. Ogawa, and M. Haseyama, "Detection of gastric cancer risk from X-ray images via patch-based convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, 2017, pp. 2055–2059.

[11] K. Oikawa, A. Saito, T. Kiyuna, H. P. Graf, E. Cosatto, and M. Kuroda, "Pathological diagnosis of gastric cancers with a novel computerized analysis system," *J. Pathol. Informat.*, vol. 8, p. 5, Feb. 2017.

[12] E. Garcia, R. Hermoza, C. B. Castanon, L. Cano, M. Castillo, and C. Castanneda, "Automatic lymphocyte detection on gastric cancer IHC images using deep learning," in *Proc. IEEE 30th Int. Symp. Comput. Based Med. Syst. (CBMS)*, Thessaloniki, Greece, 2017, pp. 200–204.

[13] S. Ding, L. Li, Z. Li, H. Wang, and Y. Zhang, "Smart electronic gastroscope system using a cloud–edge collaborative framework," *Future Gener. Comput. Syst.*, vol. 100, pp. 395–407, Nov. 2019.

[14] S. Ding, Z. Li, X. Liu, H. Huang, and S. Yang, "Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model," *Inf. Sci.*, vol. 499, pp. 12–24, Oct. 2019.

[15] L. Khelifi and M. Mignotte, "A novel fusion approach based on the global consistency criterion to fusing multiple segmentations," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 9, pp. 2489–2502, Sep. 2017.

[16] M. Li and Z.-H. Zhou, "Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 37, no. 6, pp. 1088–1098, Nov. 2007.

[17] S. Walczak and V. Velanovich, "Improving prognosis and reducing decision regret for pancreatic cancer treatment using artificial neural networks," *Decis. Support Syst.*, vol. 106, pp. 110–118, Feb. 2018.

[18] A.-F. Swager *et al.*, "Computer-aided detection of early barrett's neoplasia using volumetric laser endomicroscopy," *Gastrointest. Endosc.*, vol. 86, no. 5, pp. 839–846, 2017.

[19] Z. Wang, C. Liu, D. Cheng, L. Wang, X. Yang, and K.-T. Cheng, "Automated detection of clinically significant prostate cancer in MP-MRI images based on an end-to-end deep neural network," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1127–1139, May 2018.

[20] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A parasitic metric learning net for breast mass classification based on mammography," *Pattern Recognit.*, vol. 75, pp. 292–301, Mar. 2018.

[21] A. A. A. Setio *et al.*, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.

[22] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel, "Multiple-instance learning for anomaly detection in digital mammography," *IEEE Trans. Med. Imag.*, vol. 35, no. 7, pp. 1604–1614, Jul. 2016.

[23] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 3549–3557.

[24] Y. Zhu *et al.*, "Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy," *Gastrointest. Endosc.*, vol. 89, no. 4, pp. 806–815, 2019.

[25] C. Li, C. Shi, H. Zhang, Y. Chen, and S. Zhang, "Multiple instance learning for computer aided detection and diagnosis of gastric cancer with dual-energy CT imaging," *J. Biomed. Informat.*, vol. 57, pp. 358–368, Oct. 2015.

[26] T. Kanesaka *et al.*, "Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging," *Gastrointest. Endosc.*, vol. 87, no. 5, pp. 1339–1344, 2018.

[27] H. Wang, S. Ding, D. Wu, Y. Zhang, and S. Yang, "Smart connected electronic gastroscope system for gastric cancer screening using multi-column convolutional neural networks," *Int. J. Prod. Res.*, vol. 57, no. 21, pp. 6795–6806, 2019.

[28] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8689, 2014, pp. 818–833.

[29] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 4829–4837.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 618–626.

[31] M. Sultana, P. P. Paul, and M. L. Gavrilova, "Social behavioral information fusion in multimodal biometrics," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2176–2187, Dec. 2018, 00016.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

[33] D. Kumar, A. Wong, and G. W. Taylor, "Explaining the unexplained: A class-enhanced attentive response (CLEAR) approach to understanding deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Honolulu, HI, USA, 2017, pp. 36–44.

[34] S. Yang, W. Wang, C. Liu, and W. Deng, "Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 53–63, Jan. 2019.

[35] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9049–9058.

[36] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S.-C. Zhu, "Interpreting CNN knowledge via an explanatory graph," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4454–4463.

[37] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting CNNs via decision trees," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 6261–6270.

[38] J. Du, L. Gui, R. Xu, and Y. He, "A convolutional attention model for text classification," in *Proc. Nat. CCF Conf. Nat. Lang. Process. Chin. Comput.*, 2017, pp. 183–195.

[39] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2014, pp. 2204–2212.

[40] Y. Fang, C. Zhang, J. Li, M. P. Da Silva, and P. Le Callet, "Visual attention modeling for stereoscopic video," in *Proc. IEEE Int. Conf.*

Multimedia Expo Workshops (ICMEW), Seattle, WA, USA, 2016, pp. 1–6.

[41] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2015, pp. 379–389.

[42] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. Human Lang. Technol. (NAACL-HLT)*, 2016, pp. 1480–1489.

[43] L. Wang, Z. Cao, G. De Melo, and Z. Liu, "Relation classification via multi-level attention CNNs," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguist. (Vol. 1 Long Papers)*, 2016, pp. 1298–1307.

[44] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Shanghai, China, 2016, pp. 4945–4949.

[45] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Stockholm, Sweden, 2017, pp. 949–953.

[46] Z. Zhang, J. Li, Z. Zhong, Z. Jiao, and X. Gao, "A sparse annotation strategy based on attention-guided active learning for 3D medical image segmentation," 2019. [Online]. Available: arXiv:1906.07367.

[47] S. Chen, G. Bortsova, A. G.-U. Juárez, G. van Tulder, and M. de Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, 2019, pp. 457–465.

[48] X. Wang, S. Han, Y. Chen, D. Gao, and N. Vasconcelos, "Volumetric attention for 3D medical image segmentation and detection," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, 2019, pp. 175–184.

[49] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[50] H. Wang, H. Jia, L. Lu, and Y. Xia, "Thorax-net: An attention regularized deep neural network for classification of thoracic diseases on chest radiography," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 475–485, Feb. 2020.

[51] Z. Wang, J. Poon, S. Sun, and S. Poon, "Attention-based multi-instance neural network for medical diagnosis from incomplete and low quality data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Budapest, Hungary, 2019, pp. 1–8.

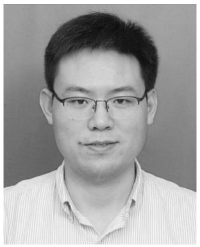
[52] Y. J. Kim, Y.-G. Lee, J. W. Kim, J. J. Park, B. Ryu, and J.-W. Ha, "Highrisk prediction from electronic medical records via deep attention networks," 2017. [Online]. Available: arXiv:1712.00010.

[53] B. L. P. Cheung and D. Dahl, "Deep learning from electronic medical records using attention-based cross-modal convolutional neural networks," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Las Vegas, NV, USA, 2018, pp. 222–225.

[54] W. Lee, S. Park, W. Joo, and I.-C. Moon, "Diagnosis prediction via medical context attention networks using deep generative modeling," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, Singapore, 2018, pp. 1104–1109.

[55] I. Gallo, A. Calefati, S. Nawaz, and M. K. Janjua, "Image and encoded text fusion for multi-modal classification," in *Proc. Digit. Image Comput. Techn. Appl. (DICTA)*, Canberra, ACT, Australia, 2018, pp. 1–7.

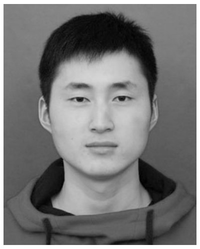
[56] J. E. A. Ovalle, T. Solorio, M. M.-Y. Gómez, and F. A. González, "Gated multimodal units for information fusion," in *Proc. 5th Int. Conf. Learn. Represent.*, Toulon, France, Apr. 2017, pp. 1–17.



Shuai Ding (Member, IEEE) received the Ph.D. degree in management information systems from the Hefei University of Technology, Hefei, China, in 2011.

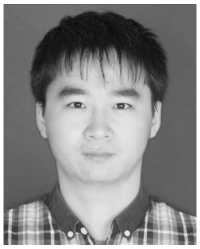
He is currently an Associate Professor with the School of Management, Hefei University of Technology. He has been a Visiting Scholar with the University of Pittsburgh, Pittsburgh, PA, USA. He has authored/coauthored more than 40 scientific articles at various top venues, including IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, and IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS. His primary research interests are in medical data mining, artificial intelligence, and automatic diagnosis system.

Dr. Ding has served as a reviewer for various journals and conferences.



Shikang Hu received the B.S. and M.S. degrees in management information systems from the Hefei University of Technology, Hefei, China, in 2016 and 2019, respectively.

He is currently working as a Senior Algorithm Engineer with the Search Algorithm Team, Alibaba Group. His research interests include multimodal deep learning, medical data mining, and decision support systems.



Xiaojian Li received the bachelor's degree in automation from Southeast University, Nanjing, China, in 2011, and the joint Ph.D. degree from the Department of Automation, the University of Science and Technology of China, Hefei, China, and the Department of Mechanical and Biomedical Engineering, the City University of Hong Kong, Hong Kong, in 2017.

During this period, the research areas are control science and engineering and mechanical and biomedical engineering. His current research field is medical micro-robot.

Youtao Zhang (Member, IEEE) received the Ph.D. degree in computer science from the University of Arizona, Tucson, AZ, USA, in 2002.

He is currently an Associate Professor of Computer Science with the University of Pittsburgh, Pittsburgh, PA, USA. His current research interests include computer architecture and memory systems, and hardware-assisted AI/ML.

Dr. Zhang was a recipient of the U.S. National Science Foundation Career Award in 2005. He is also the coauthor of several papers that received paper awards. He is a member of ACM.

Desheng Dash Wu (Senior Member, IEEE) is with the School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 100 papers in refereed journals, such as *Production and Operations Management*, *Decision Support Systems*, *Decision Sciences*, *Risk Analysis*, and IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS. His research interests include enterprize risk management in operations, performance evaluation in financial industry, and decision sciences.

Dr. Wu was an Associate Editor/Guest Editor for the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS and *Omega*. He is an Elected Member of the European Academy of Sciences and Arts.