LKAN: LLM-based Knowledge-aware Attention Network for Clinical Staging of Liver Cancer

Ya Li, *Member, IEEE*, Xuecong Zheng, Jiaping Li, Qingyun Dai, Chang-Dong Wang, *Senior Member, IEEE*, and Min Chen, *Fellow, IEEE*

***Abstract*— Clinical staging of liver cancer (CSoLC), an important indicator for evaluating the degree of deterioration of primary liver cancer cells (PLCCs), is key in the diagnosis, treatment, and rehabilitation of liver cancer. In China, the current CSoLC adopts the China liver cancer (CNLC) staging, which is usually evaluated by clinicians based on the patient's radiology reports. Therefore, inferring clinical information from unstructured radiology reports can provide auxiliary decision support for clinicians. The key to solving the challenging task is to guide the model to pay attention to the staging-related words or sentences, and the following issues may occur: 1) Imbalanced categories: The symptoms of liver cancer in the early- or mid-stage are not obvious, resulting in more data in the end-stage. 2) Domain sensitivity of liver cancer data: The liver cancer dataset contains a large amount of domain knowledge, and the conventional methods can exacerbate out-of-vocabulary, which greatly affects the accuracy of classification. 3) Free-text and lengthy report: The radiology report of liver cancer sparsely describes various lesions with domain-specific terms, which poses difficulties in mining key information related to staging. To tackle these challenges, this article proposes a large language model (LLM)-based Knowledge-aware Attention Network (LKAN) for CSoLC. First, for maintaining semantic consistency, LLM and a rule-based algorithm are integrated to generate more diverse and reasonable data. Second, unlabeled radiology corpus of liver cancer are pre-trained to introduce domain knowledge for subsequent representation learning. Third, attention is improved by incorporating both global and local features, which can provide professional guidance for the classifier to focus on the important information. Compared with the baseline models, the classification accuracy of LKAN has achieved the best results with 90.3% Accuracy, 90.0% Macro_F1 score, and 90.0% Macro_Recall. The code is available at https://github.com/xczhh/Supplemental-Material.**

***Index Terms*— Clinical staging of liver cancer (CSoLC), Chinese radiology reports, natural language processing, domain knowledge.**

This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFF1201200, in part by NSFC under Grant 62276277, and in part by the Program of Guangdong Provincial Department of Education under Grant 2021ZDZX1079. (Corresponding authors: Jiaping Li and Chang-Dong Wang.)

Ya Li, Xuecong Zheng and Qingyun Dai are with the School of Electronics and Information, Guangdong Polytechnic Normal University, Guangzhou 510660, Guangdong, China, and with the Guangdong Provincial Key Laboratory of Intellectual Property & Big Data, Guangzhou, 510665, China (e-mail: liya2829@gpnu.edu.cn; xuecongzh@gmail.com; dqy@gpnu.edu.cn).

Jiaping Li is with the Department of Interventional Oncology, the First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510080, China (e-mail:lijiap@mail.sysu.edu.cn).

Chang-Dong Wang is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China, and with the Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou 510000, China (e-mail: changdongwang@hotmail.com).

Min Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: minchen@ieee.org).

I. Introduction

PRIMARY liver cancer (PLC) is the third leading cause of cancer death in the world, causing over 830,000 deaths each year, of which nearly half are from China [1]–[3]. As a key step in cancer diagnosis and treatment, liver cancer staging is an important factor for doctors in choosing treatment plans and prognoses for their patients [4]–[6]. In China, clinical staging of liver cancer (CSoLC) requires clinicians to evaluate patients' various indicators following the China liver cancer (CNLC) [7] staging. However, radiology reports typically consist of lengthy and unstructured medical sentences. It is a complex and time-consuming process to mine staging-related information from radiology reports to make judgments that challenges a clinician's experience and energy. As a result, the fatigue and experience of doctors hinder the accuracy of staging. In addition, China faces the problems of uneven distribution of medical resources [8], [9] and the concentration of high-level specialized hospitals in large cities, leading to overcrowding in better hospitals and further aggravating the overloaded workload of clinicians. Therefore, obtaining accurate CSoLC is a challenging task.

The current research on clinical staging of cancer mainly focuses on methods based on medical images and radiology reports. In the image-based methods, Pan et al. [10] leveraged convolutional neural network (CNN) to learn the features of brain tumors from 195 multiphase magnetic resonance imaging (MRI) images and classified them into high- or low-stage. Patil et al. [11] proposed a melanoma cancer staging method combining CNN with similarity measure for text processing (SMTP) as loss function, and divided melanoma into three stages. De et al. [12] used soft-label sequential regression to classify prostate cancer into five levels using the Gleason grade group (GGG) as a grading index based on bi-parametric magnetic resonance imaging (bp-MRI). Huang et al. [13] graded laryngeal cancer tumor (LCT) by identifying lesion regions of interest (LROIs) based on the degree of staining, number, morphology, and interrelationship of atypical squamous cells in histopathological images. Fan et al. [14]

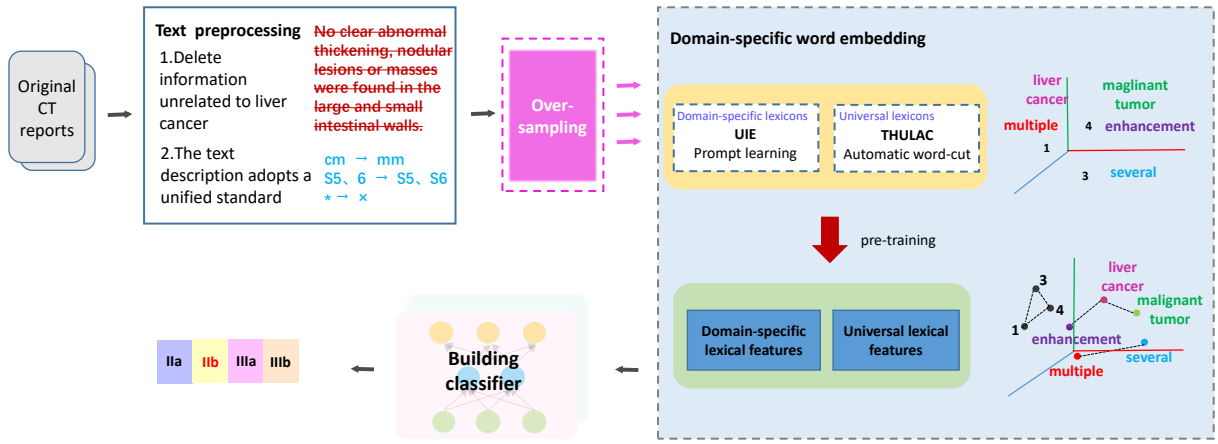


Fig. 1. The flowchart of the proposed LKAN method.

jointly predicted Ki-67 and tumor grade with a multitask learning framework by separately utilizing radiomics from tumor MRI series. The above studies have performed very well in cancer staging. However, the image-based methods rely on radiologists' annotations, which are time-consuming and difficult to be applied in clinical practice [15]–[17].

As an alternative to the image-based labeling, radiology reports have the potential to avoid human annotation efforts [18]–[20]. In recent years, radiology reports have been written by radiologists. As they integrate the expert's cognition and judgment in text form, the workload of data annotation is greatly reduced. Consequently, methods based on radiology reports have been received attention. Fink et al. [21] developed a pulmonary embolism scoring system on the basis of structured radiology reports and compared its diagnostic performance with the conventional clot burden index. Fink et al. [22] exploited structured radiology reports to train a deep natural language processing (NLP) model for classifying tumor response categories (TRCs) based on free-text oncology reports. D'Anniballe et al. [16] developed rule-based algorithms (RBA) for the extraction of disease labels and trained attention-guided recurrent neural networks (RNN) to classify radiology reports into multiple diseases. Eskreis et al. [23] developed a deep learning model for automated background parenchymal enhancement (BPE) categorization based on breast MRI reports. Although several studies have demonstrated the value of using radiology reports in various clinical decision-making processes, their application in assisting CSoLC remains an open question.

The radiology report serves as a communication bridge between radiologists and clinicians. However, acquiring a sufficient number of authentic radiology reports poses a challenge, which impedes the progress of the CSoLC task. Besides, the following issues may occur when applying radiology reports to CSoLC: 1) Imbalanced categories: The symptoms of liver cancer in the early- or mid-stage are not obvious, resulting in more data in the end-stage. 2) Domain sensitivity of liver cancer data: The liver cancer dataset contains a large amount of domain knowledge. Conventional methods may exacerbate out-of-vocabulary, and impair the classification performance. 3) Free-text and lengthy report: The radiology report of liver cancer sparsely describes various lesions with domain-specific terms, which hinders the mining and utilization of critical information related to staging.

Motivated by the above analysis, this article proposes a LLM-based Knowledge-aware Attention Network (LKAN) to assist clinicians in CSoLC. Fig. 1 illustrates the flowchart of LKAN. Firstly, in order to address the issue of imbalanced categories, we propose a novel data augmentation method that maintains semantic consistency by combining a large language model (LLM) with RBA to generate more diverse and reasonable data. Secondly, out-of-vocabulary increases the difficulty for the conventional language models to adapt to the liver cancer dataset. Therefore, we use a universal and domain-specific lexicon automated generation method and pre-train a domain-specific word embedding to alleviate the domain sensitivity issue. Finally, unstructured and lengthy textual features are difficult to utilize. Descriptions of various lesions are scattered throughout sentences or paragraphs, rendering it difficult for the model to recognize and infer staging-related information. To tackle this issue and improve the ability to identify the precise CSoLC, the attention block in the hierarchical attention network (HAN) [24] is improved for capturing the important and scattered information by incorporating both global and local features. Experimental results show that the proposed LKAN method has achieved the best results with 90.3% Accuracy, 90.0% Macro_F1 score, and 90.0% Macro_Recall. Our proposed method is important for providing rapid diagnosis and screening for clinicians in the early stage of treatment and reducing their workload of reading reports, this method can also offer decision-making assistance for interns and in areas with a shortage of healthcare resources.

Specifically, our contributions are threefold:

1) Category imbalance is a long-standing in medical and other domains. A novel over sampling method is proposed by combining LLM and RBA to generate more diverse samples. This approach may offer new insights and perspectives for addressing the challenge of category imbalance.
2) It is necessary to integrate domain knowledge into the accurate classification of the domain sensitivity data. We

TABLE I
SUMMARY OF NOTATIONS

Symbols	Definitions and descriptions
m	The number of sentences in a report
n	The number of Chinese characters in a sentence
L_n	The label set
$\mathbf{S}$	The matrix containing multiple sentences
$\mathbf{s}_u$	The u-th sentence in report
$s_{u,t}$	The t-th word in sentence u
v	The number of Chinese words in the corpus
d	The dimension of Chinese words
$\mathbf{W}_e \in \mathbb{R}^{v\times d}$	The word embedding matrix with length v
$\mathbf{X}_u \in \mathbb{R}^{1\times n\times d}$	The word vectors obtained from $\mathbf{s}_u$ by word embedding
$\mathbf{x}_{u,t} \in \mathbb{R}^{1\times 1\times d}$	The t-th word vector with d-dimensions
p	The settings of units in gate recurrent unit
$\overrightarrow{\mathbf{H}}_{u,t} \in \mathbb{R}^{1\times p}$	The forward hidden state obtained from $\mathbf{x}_{u,t}$ by gate recurrent unit
$\overleftarrow{\mathbf{H}}_{u,t} \in \mathbb{R}^{1\times p}$	The backward hidden state obtained from $\mathbf{x}_{u,t}$ by gate recurrent unit
$\mathbf{H}_{u,t} \in \mathbb{R}^{1\times 2p}$	The overall hidden state of the whole sentence centered around $\mathbf{x}_{u,t}$ obtained by fusion
$\mathbf{H}_u \in \mathbb{R}^{1\times n\times 2p}$	The sentence representation of $\mathbf{X}_u$
$\mathbf{f}_u^{avg} \in \mathbb{R}^{1\times 2p}$	The feature map obtained by global average pooling
$\mathbf{f}_u^{max} \in \mathbb{R}^{1\times 2p}$	The feature map obtained by global max pooling
$\mathbf{z}_{u,t}$	The hidden representation of $\mathbf{H}_{u,t}$
$\mathbf{u}_w$	The context vector for measuring the importance of Chinese words
$\mathbf{a}_{u,t}$	The importance weight
$\mathbf{f}_u^{att} \in \mathbb{R}^{1\times d}$	The feature map obtained by attention
$\mathbf{f}_u \in \mathbb{R}^{1\times(d+4p)}$	The overall feature map obtained by fusion
$\boldsymbol{\tau} \in \mathbb{R}^{1\times d}$	The output tensor of $\mathbf{s}_u$
N	The total number of reports
M	The number of stages
$\boldsymbol{\rho} \in \mathbb{R}^{N\times M}$	The predicted probability set of each stage
$\mathbf{W}_c$	The trainable parameters of fully connected layers
$\mathbf{b}_c$	The bias of fully connected layers
$\mathbf{Y} \in \mathbb{R}^{N\times M}$	The ground truth set of each stage
$\mathcal{L}$	The output of loss function

explore the effectiveness of constructing domain-specific word embeddings and provide references for subsequent work.

3) Chinese radiology reports are utilized for CSoLC and achieved high accuracy, which explores a new auxiliary method for clinical staging of other cancers.

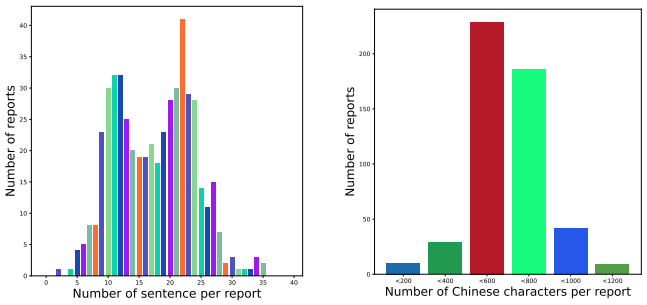
The notations that are used throughout the paper are summarized in Table I. The remainder of this article is organized as follows. In Section II, we briefly introduce the dataset used for CSoLC and provide the details of LKAN implementation. In Section III, we introduce our extensive and comparative experiments and report the results of our proposed LKAN and related methods. In Section IV, we discuss and visually analyze our method. We summarize our work in Section V.

II. MATERIALS AND METHODS

A. Dataset

1200 liver cancer digital radiology reports from January 2018 to January 2022 have been collected from the First Affiliated Hospital of Sun Yat-sen University. This study was approved by the Clinical Research and Experimental Animal Ethics Committee of the First Affiliated Hospital of Sun Yat-sen University (No. 2023780) on November 17, 2023. Due to the retrospective nature of this study, the requirement for informed consent was waived. 402 radiology reports with scan sites not in the abdomen are excluded, and 292 radiology reports are excluded as they are unrelated to PLC. Thus, the final dataset includes 506 radiology reports with clear staging information on liver cancer, and their staging results are annotated by three senior clinicians with 10+ years of experience in liver cancer treatment. There are 50 stage IIa, 70 stage IIb, 150 stage IIIa, and 236 stage IIIb. Each stage is labeled with 0, 1, 2, and 3, respectively. There are no stage I tumors on our dataset because the number is quite small at the time of the initial data collection (i.e., only 5 samples) and are excluded due to noncompliance with the selection criteria (i.e., unclear staging information). The distributions of sentences per report and report length on this dataset are shown in Fig. 2. It is worth noting that the reports typically contain 10 to 24 sentences (Fig. 2(a)), most of which exceed 600 Chinese characters in length (Fig. 2(b)). According to CNLC, Fig. 3 shows that the staging results are gradually differentiated by four indicators, namely extrahepatic metastasis, vascular invasion, number of tumors, and size of tumors, respectively. Meanwhile, if the current indicator can fully determine the staging result, subsequent indicators won't have a decisive role in clinical staging. Therefore, four staging results are determined by different indicators. Table III shows that only stage IIa needs to be jointly determined by four indicators.

In the radiology report, the locations, sizes, and metastasis of the lesions are detailed by the radiologists. However,



(a) Number of Sentences per report (b) Sentence length

Fig. 2. The distributions of the number of sentences per report and the sentence length on the dataset.

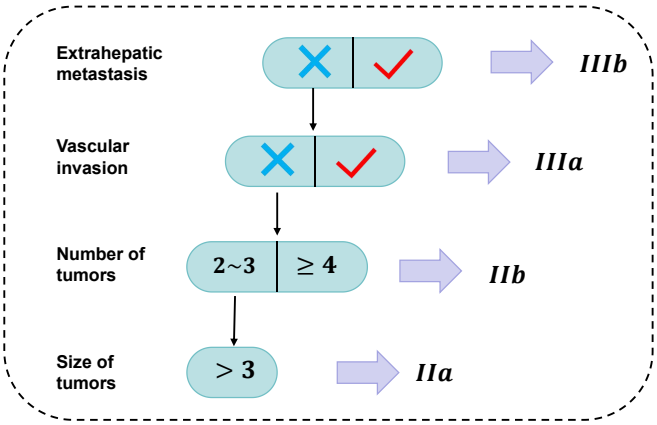


Fig. 3. The flowchart of CNLC.

TABLE II
EXAMPLE SENTENCES. BOLD WORDS REPRESENT THE KEYWORDS RELATED TO CLINICAL STAGING PREDICTION, ITALICIZED WORDS REPRESENT THAT THIS INFORMATION IS IMPORTANT BUT CANNOT DETERMINE THE FINAL STAGING, AND UNDERLINED WORDS REPRESENT THAT THE TUMOR IS NOT HCC

Sentences	Staging	Indicators
No filling defect is found in the main portal vein and its branches. There is no dilation of the intrahepatic and extrahepatic bile ducts. **An enlarged lymph node** with a size of *approximately 26.6mm* × *21.4mm* can be seen **in the hilar region of the liver**, which is **significantly enhanced after the enhanced scanning**.	IIIb	Extrahepatic metastasis
In *liver S6*, a slightly low-density and circular-like *shadow* with a *diameter* of *approximately 18mm* is observed. After *the enhanced scan*, the lesion shows *equal density in the arterial phase* and *low density in the portal phase*. There is an oval **filling defect** shadow on the left branch of **the portal vein**, with **mild enhancement** after **the enhanced scanning**.	IIIa	Vascular invasion
There are **multiple** circular low-density **lesions** in **liver S3** and **S5-S8**; the **largest** is in **liver S3**, with a diameter of approximately 18mm. During **the arterial phase**, **nodular enhancement** is observed at the edge of the low-density lesions in **liver S6** and **S8**, and **low-density** shadows are observed during **the portal** and **delayed phases**. **A lesion** in **S5** shows **circular enhancement**. **The lesion** in **liver S3** shows **uneven enhancement of the entire tumor** during **the arterial phase**, while it shows **low-density shadows** during **the portal phase**. No enhancement was observed in the remaining nodules. Another low-density lesion with a diameter of approximately 15mm is observed in liver S7, with clear boundaries. No enhancement was observed after enhanced scanning.	IIb	Number of tumors >3 (lesions in S4, S5, S6, S8 are HCC, but lesion in S7 is a cyst.)
The liver margin is not smooth, the proportion of liver lobes is imbalanced, and liver fissures are widened. **Two** circular low-density **shadows** can be seen in **liver S4** and **S5/6**, with sizes of approximately 31mm × 28mm × 28mm, **33**mm × **33**mm × 32mm. **Uneven enhancement in the arterial phase**, with a degree of enhancement close to or slightly lower than the surrounding liver parenchyma. However, **enhancement in the portal phase** is **decreased**, with slight compression of adjacent liver veins.	IIa	Size of tumors >3 Number of tumors = 2 (S4, S5/6)

*English translations are presented in the table for illustration, But all reports are written in Chinese (Table I in Supplemental Material). The data format provided by the clinicians is as follows: the reports and their corresponding staging are shown in the first two columns of the table. We list the indicators in the last column to explain how clinicians summarize information related to CSoLC from the radiology reports based on CNLC.

TABLE III
THE DECISIVE INDICATORS FOR DIFFERENT STAGING RESULTS

Indicators	Extrahepatic metastasis	Vascular invasion	Number of tumors	Size of tumors
IIIb	✓	—	—	—
IIIa	×	✓	—	—
IIb	×	×	>3	—
IIa	×	×	2∼3	>3

✓ represents *Yes*, × represents *No*, — represents *Not important*.

determining the specific staging result requires distinguishing whether the lesion is hepatocellular carcinoma (HCC) or whether metastasis occurs. These are not directly provided in the radiology report and need to be further inferred based on the subjective experience of clinical doctors. Therefore, paying attention to important words and sentences from free-text is crucial. The four indicators and their significant imaging characteristics are summarized below:

- Extrahepatic metastasis: HCC has metastasized outside of the liver, and the common sites of metastasis are the hepatic hilum, peritoneum, lungs, and bone [25]. Extrahepatic metastasis is often associated with end-stage liver cancer.
- Vascular invasion: The portal vein shows filling defects or cancerous thrombi.
- Number of tumors: The number of HCC lesions is four or more. Under the CT enhancement scan, HCC typically presents a "fast-in and fast-out" pattern [26]. In radiology reports, the prominent characteristic of HCC is typically described as enhancing in the arterial phase and attenuating in the portal venous or delayed phases (with its density lower than that of the surrounding liver parenchyma).
- Size of tumors: As an important indicator of liver cancer staging, the size of a tumor is usually described by its longest and shortest diameters.

For further illustration of our work, Table II lists the example sentences with the corresponding staging, which are annotated by professional clinicians. Firstly, we can see that when advanced symptoms occur (such as lymph node metastasis in the hepatic portal area in Case 1 and filling defect in the portal vein in Case 2), the number, location, and enhancement of the tumor will no longer affect the staging results. Secondly, despite multiple lesions appearing in case 3, the lesion in liver S7 did not exhibit the typical imaging characteristics of liver cancer, and could only be considered as a cyst rather than HCC. Finally, the two circular and low-density shadows in Case 4 match the imaging characteristics of liver cancer, therefore they are classified as stage IIa.

B. Text preprocessing

As depicted in Fig. 2, our dataset comprises lengthy radiology reports. However, only a few sentences are extremely relevant to CSoLC (Fig. 2 in Supplemental Material). Unrelated sentences are considered noisy data, which lengthens the report and increases computational power consumption. Hence, it is

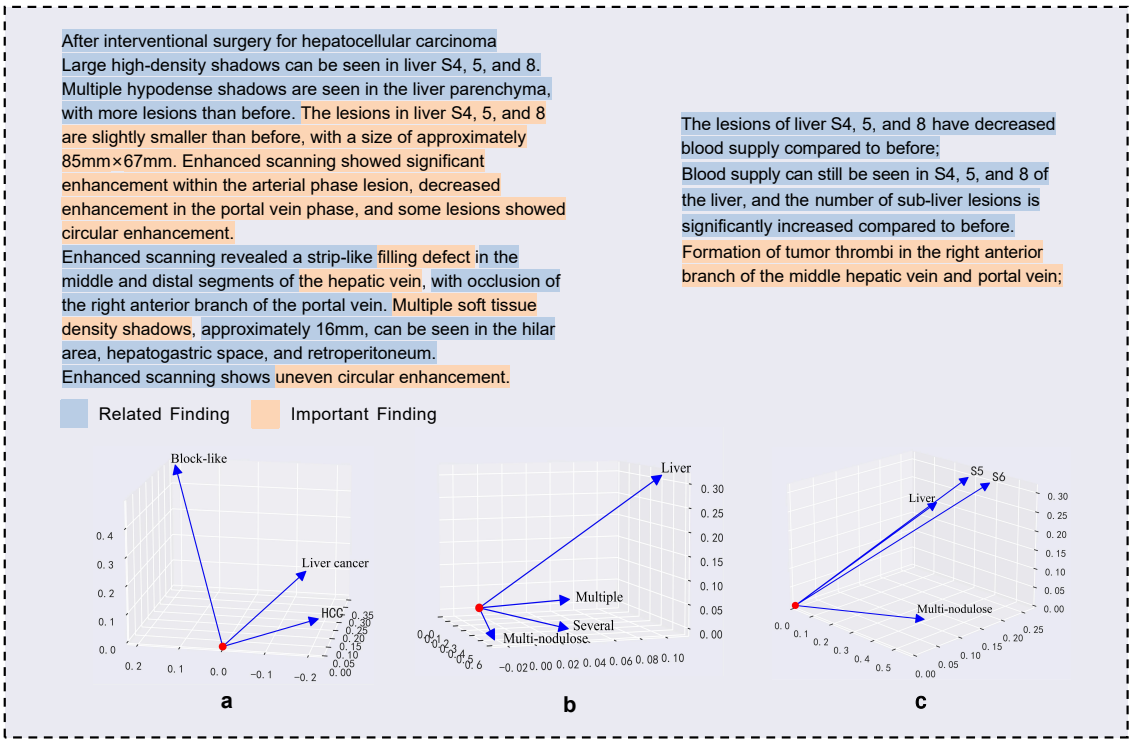


Fig. 4. An example of Chinese radiology reports. Blue blocks represent normal findings; Orange blocks represent relevant findings; Others indicate that they are not related to CSoLC. The distribution of pre-training word vectors is shown in (a). The spatial relationships of similar words are shown in (b). *All reports are written in Chinese (Fig. 2 in Supplemental Material); English translations are presented in the figure for illustration.

necessary to preprocess all reports before further modeling (as illustrated in Fig. 4).

In detail, we first remove all descriptions that are unrelated to the CSoLC from the report by regular-expression-matching. In radiology reports, tumor sizes are often expressed with various length units. We standardize the unit to millimeters and then convert all measurements accordingly. Next, a Chinese stopword corpus is utilized to filter meaningless stopwords from the text, and those words may contain Chinese auxiliary words that are not essential components of the radiology reports. After all the above operations, the length of reports is reduced to 150.

The difference between Chinese and English is that words in English can be easily recognized since the space token is a good approximation of a word divider [27]. In Chinese, Chinese characters are used as the basic unit, but words have more semantics. At this step, we first segment the text into individual words and then convert them into corresponding indices based on the word-index vocabulary. Words outside the vocabulary are marked with [UNK]. We use a maximum sequence length of 150 (the maximum length in the dataset) and pad shorter sequences with 0.

C. Over-sampling

Faced with the imbalanced datasets, many tasks leverage pre-training language models [28] as the basis and then fine-tune them using limited samples [29]–[31]. These methods are effective, but they are also limited by the differences between the source domain and the target domain. In fact, the majority of general domain text is substantively different from biomedical text, raising the prospect of negative transfer that actually hinders the target performance [32]. Besides, text data augmentation can also overcome the sample size limit and work together with other few-shot learning methods in NLP [33]–[35]. LLMs store a large amount of knowledge, and using them to generate reliable samples can reduce the burden on human annotation [35].

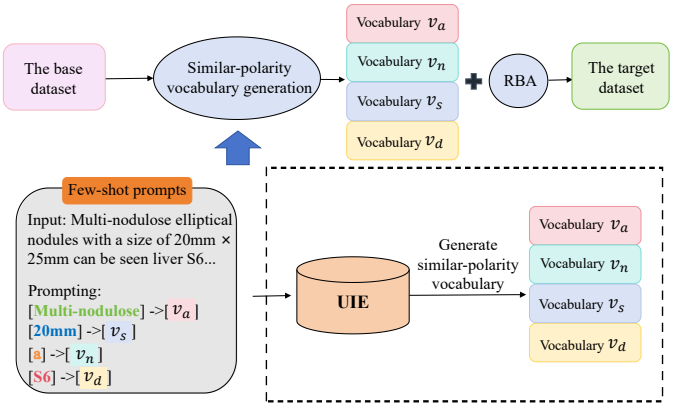


Fig. 5. The flowchart of over sampling. UIE is prompted by few-shot prompts construction to generate similar-polarity vocabulary. The target dataset is generated by the similar-polarity vocabulary with RBA. Four vocabularies are different and are generated based on the morphology, number, size, and growth site of the tumors, resulting in $\boldsymbol{v_a}$, $\boldsymbol{v_n}$, $\boldsymbol{v_s}$, and $\boldsymbol{v_d}$ respectively.

The goal of data augmentation is to produce sensible and diverse new samples that maintain semantic consistency [35].

Existing studies indicate that pre-training language models can help augment a dataset with new samples with similar semantic meaning [36], [37], but they still face a lack of diversity and faithfulness. In this study, we propose a method called AugUIE, which uses UIE [38] (a popular LLM based on ERNIE 3.0 [39]) combined with RBA to generate auxiliary samples under CNLC rules.

Two strategies, namely inversion and imitation [40] are used to over-sample the minority categories. There are differences in the number of tumors and the size of tumors between stage IIa and stage IIb. In the inversion strategy, artificial texts are generated by reversing words into their counterparts that carry the opposite polarity. Conversely, we create new samples from the original categories in the imitation strategy. The examples of the two strategies are illustrated in Fig. 6.

The overall framework of AugUIE is shown in Algorithm 1. Given a base dataset $\{(s_n, L_n)\}_{n=1}^{N_{data}}$ with a label space $y_j \in Y_n$, and a novel dataset $\{(target\_s_n, target\_L_n)\}_{n=1}^{N_{data}}$ with the same label space $y_j \in Y_n$. Fig. 5 shows that UIE generates different vocabularies $v = (v_a, v_n, v_d, v_s)$ by few-shot prompts. Each vocabulary is different, but the words within each vocabulary have similar polarities. The application of these four vocabularies aims to enrich tumor features through similar word transformations, thereby enhancing the diversity of generated sentences. Specifically, v_a represents the morphology of the tumors, v_n represents the number of tumors, v_s represents the size of the tumors, and v_d represents the growth site of the tumors. Furthermore, to prevent non-standard sentence generation, the inversion and imitation strategies are used under the constraints of CNLC rules to replace words from v and generate trusted sequences. The base dataset has the same number of pairs as the target dataset.

Algorithm 1 The framework of AugUIE for data augmentation

Input: $\{(s_n, L_n)\}_{n=1}^{N_{data}}$, the dataset of sequence pairs.
Input: θ, initial UIE after prompt learning.
Input: $\{target\_L\}_{n=1}^{N_{data}}$, label for target conversion.
Output: $\{(target\_s_n, target\_L_n)\}_{n=1}^{N_{data}}$, the dataset of target sequence pairs.

for i $= 1, 2, ..., N_{data}$ **do**
  $v_a, v_n, v_d, v_s \leftarrow \mathrm{UIE}(s_i, L_i|\theta)$
  if ($target\_L_i$ == IIa) **then**
    w_s is randomized but not less than 30mm.
    w_n is required to be no greater than 3.
  else
    if ($target\_L_i$ == IIb) **then**
      w_n is assigned a value from v_n.
    end if
  end if
  w_a,w_d is assigned values from v_a,v_d respectively.
  $target\_s_i \leftarrow [s_{i1}, \ldots, w_a, w_d, s_{iv}, w_s, w_n, \ldots, s_{il}], l \in [1, maxlen], v \in [1, l)$.
end for
return $\{(target\_s_n, target\_L_n)\}_{n=1}^{N_{data}}$.

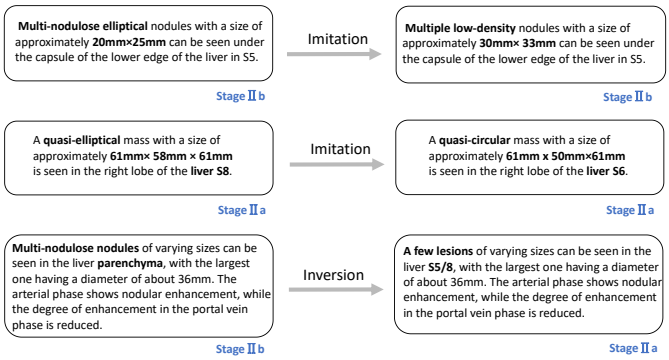


Fig. 6. Examples of "inversion" and "imitation" strategies. English translation in the figure is only for illustration, and all experiments are based on Chinese words (Fig. 3 in Supplemental Material).

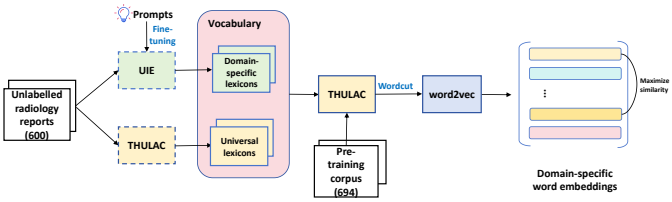


Fig. 7. The flowchart of domain-specific word embedding. To alleviate the out-of-vocabulary problem, two components are used to construct a vocabulary: a domain-specific vocabulary and a universal vocabulary. Firstly, Domain-specific vocabulary is generated by UIE with few-shot prompts, while universal vocabulary is generated by using THULAC. Secondly, vocabulary is used by THULAC to split the pre-training corpus into target words. Finally, Target sentences are pre-trained with word2vec to capture contextual information among medical words, thereby facilitating better adaptation of the subsequent classifier to the staging task.

D. Domain-specific word embedding

Using Chinese radiology reports as experimental data for the clinical staging task is limited by the domain sensitivity issue. Due to the vast semantic differences from general tasks [41], applying existing pre-trained language models may results in the appearance of out-of-vocabulary words or unavailable sub-words [32], greatly diminishing the model's learning capacity. To mitigate the out-of-vocabulary problem, domain-specific vocabulary is necessary for subsequent representation learning. However, this requires equipping the vocabulary with domain-specific knowledge and entails a significant amount of labor [42]. Therefore, we propose an adaptive vocabulary generation method based on prompt learning to preserve the integrity of biomedical terms, minimizing the number of subword units required to encode them [43].

In detail, The vocabulary includes both domain-specific and universal words. Domain-specific vocabulary reflects professional knowledge, which also requires maintaining the integrity of biomedical terms [43]. As shown in Fig. 7, domain words are generated from a dataset of unlabeled radiology reports by UIE with few-shot prompts. This step not only replaces the process of manual annotation by doctors but also accurately extracts medical terms. Universal words such as the number, size, and spread of tumors determine the final diagnostic result. In the experiment, THU lexical analyzer for Chinese (THULAC) [44] is used for generating universal

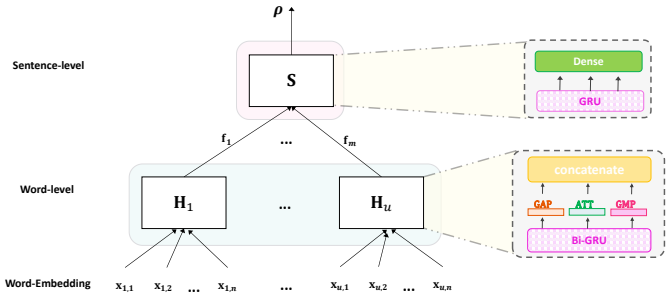


Fig. 8. The architecture of the proposed word encoder model.

words from the general corpus.

The embedding block mainly tackles the problem of feature acquisition. In order to capture domain-specific knowledge and universal semantic meaning, in this work, unlabeled radiology reports from the dataset (506 of which are used for CSoLC) combined with domain-specific vocabulary are pre-trained on word2vec. Fig. 4 (a), (b), and (c) reflect that Similar medical words will come closer together, which brings convenience to subsequent sequence learning.

E. Building classifier

1) Structure: As depicted in [24], documents have a hierarchical structure: words form sentences, and sentences form a document. The hierarchical structure constructs document representations by constructing sentence representations and aggregating them into document representations. In radiology reports, the sentences often contain a complex mixture of information describing not only the lesion of interest but also other related lesions [45]. Taking CSoLC as an example, the final staging result is only related to the comprehensive information composed of some words and sentences. Therefore, we use the hierarchical structure in [24] as the foundation (Fig. 8) and apply it to CSoLC by modifying the internal network structure. Experiments have shown that the hierarchical structure is helpful for processing and classifying the free-text and lengthy radiology reports.

2) Encoded layer:

a) Word encoder: The architecture of the word encoder model is shown in Fig. 8. Assume that a radiology report $\mathbf{S}$ has m sentences and each sentence contains n words. For a sentence, each word is represented as a vector with d-dimensions through embedding matrix $\mathbf{W}_e$,

$$\mathbf{x}_{u,t} = \mathbf{W}_e s_{u,t}, t \in [1, n], \tag{1}$$

$$\mathbf{X}_u = [\mathbf{x}_{u,1}, \mathbf{x}_{u,2}, \ldots, \mathbf{x}_{u,n}], \tag{2}$$

where $s_{u,t}$ with $t \in [1, n]$ represents the t-th word in the u-th sentence, $\mathbf{W}_e$ represents the word embedding, $\mathbf{x}_{u,t}$ represents the t-th word vector in the u-th sentence, and $\mathbf{X}_u$ represents the word vectors of $\mathbf{s}_u$.

Compared to RNN and long short-term memory (LSTM), gate recurrent unit (GRU) with a less complex structure involves fewer parameters, which reduces the amount of computation while ensuring accuracy [46], [47]. At the word-level, we use Bi-GRU to extract the features of semantic relationships. In detail, words are fed into a Bi-GRU network composed of forward and backward GRU cells in parallel. Then, groups of semantic-correlated features that enfold the information of forward and backward are generated and merged.

$$\overrightarrow{\mathbf{H}}_{u,t} = GRU(\overrightarrow{\mathbf{x}}_{u,t}), t \in [1, n], \tag{3}$$

$$\overleftarrow{\mathbf{H}}_{u,t} = GRU(\overleftarrow{\mathbf{x}}_{u,t}), t \in [n, 1], \tag{4}$$

$$\mathbf{H}_{u,t} = f_{fusion}([\overrightarrow{\mathbf{H}}_{u,t}, \overleftarrow{\mathbf{H}}_{u,t}]), \tag{5}$$

$$\mathbf{H}_u = [\mathbf{H}_{u,1}, \mathbf{H}_{u,2}, \ldots, \mathbf{H}_{u,n}], \tag{6}$$

where $\overrightarrow{\mathbf{H}}_{u,t}$ represents the forward hidden state, $\overleftarrow{\mathbf{H}}_{u,t}$ represents the backward hidden state, $\mathbf{H}_{u,t}$ represents the information of the whole sentence centered around $\mathbf{x}_{u,t}$, and $\mathbf{H}_u$ represents the overall sentence representation of $\mathbf{X}_u$.

For capturing the important and scattered information [48], the attention block is used to provide professional guidance for the subsequent classifier to focus on the keywords. Both global and local features are incorporated by concatenating global average pooling (GAP), global max pooling (GMP), and attention (ATT) to improve the robustness to noise. GAP obtains the global information [46], [47], [49], while GMP and Attention generate the most important information [47], [49].

To alleviate the problem of over-fitting, dropout [50] is applied before the word encoder.

The output of GAP $\mathbf{f}_u^{avg}$, Attention $\mathbf{f}_u^{att}$, and GMP $\mathbf{f}_u^{max}$ for the i-th sentence feature map can be defined as follows:

$$\mathbf{f}_u^{avg} = avg\{\mathbf{H}_u\}, \tag{7}$$

$$\mathbf{f}_u^{max} = max\{\mathbf{H}_u\}, \tag{8}$$

$$\begin{aligned} \mathbf{z}_{u,t} &= tanh(\mathbf{W}_w \mathbf{H}_{u,t} + \mathbf{b}_w) \\ \mathbf{a}_{u,t} &= \frac{exp(\mathbf{z}_{u,t}^\top \mathbf{u}_w)}{\sum_t exp(\mathbf{z}_{u,t}^\top \mathbf{u}_w)} \\ \mathbf{f}_u^{att} &= \sum_t \mathbf{a}_{u,t} \mathbf{H}_{u,t} \end{aligned} \tag{9}$$

where $\mathbf{W}_w$ and $\mathbf{b}_w$ represent the weight matrix and bias vector.

To comprehensively consider the features from different operations, the feature fusion can be formulated as:

$$\mathbf{f}_u = f_{fusion}([\mathbf{f}_u^{avg}, \mathbf{f}_u^{max}, \mathbf{f}_u^{att}]). \tag{10}$$

$$\mathbf{S} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m], \tag{11}$$

where $f_{fusion}(\cdot)$ represent the concatenation operation.

b) Sentence encoder: Finally, GRU is used to encode sentences and obtain an abstract representation of the document, which is an important feature of CSoLC.

$$\boldsymbol{\tau} = GRU(\overrightarrow{\mathbf{S}}). \tag{12}$$

where $\boldsymbol{\tau}$ represents the high-level representation of a report.

TABLE IV
HYPERPARAMETER SETTING

Hyperparameter	Value
Max sequence	150
Vocab size	20000
Batch size	16
Dropout	0.4
Embedding vector size	300
Optimizer	Adam
Learning rate	0.01
Epochs	30

F. Classification and loss function

Our model is trained using categorical cross-entropy as a loss function to reduce the amount of error that could happen.

$$\boldsymbol{\rho} = softmax(\mathbf{W}_c\boldsymbol{\tau} + \mathbf{b}_c), \tag{13}$$

$$\mathcal{L} = -\frac{1}{N}\sum_{i}\sum_{c}^{M}\mathbf{Y}_{i,c}log(\boldsymbol{\rho}_{i,c}), \tag{14}$$

where N represents the total number of reports, M represents the number of stages, $\mathbf{Y}_{i,c}$ represents the true value of report i belonging to stage c (set 0 or 1), and $\boldsymbol{\rho}_{i,c}$ represents the predicted probability of report i belonging to stage c.

III. EXPERIMENTS AND RESULTS

A. Experimental settings

A series of experiments are conducted on the dataset in a supervised learning setting. The dataset is divided into a training set, validation set, and test set in the ratio of 3:1:1. The model is trained on the training set, the best model is selected on the validation set, and performance is evaluated on the test set.

The model is implemented through the Python library Keras 2.3.1 and TensorFlow 1.4.0. The trained model is used to predict the data in the test set for CSoLC. The hyperparameters of the model used in the experiment are shown in Table IV. The experimental results show that the model has achieved the optimal performance with the above parameter configurations.

B. Evaluation metrics

In our experiments, six evaluation metrics are Precision (Pr), Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), Macro_F1 (F), and Macro_Recall (R), which are used to compare the staging performance of different models.

To better reflect the performance of the baselines and LKAN under the imbalanced dataset, Pr represents the probability of correct predictive results to the total predictions, which can be calculated as

$$Pr = \frac{TP}{TP + FP}, \tag{15}$$

where TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives. Sens and Spec are used to measure mis- and missed diagnosis rates of the model in staging prediction,

$$Sens = \frac{TP}{TP + FN}, \tag{16}$$

$$Spec = \frac{TN}{FP + TN}. \tag{17}$$

To evaluate the overall performance of staging performance, Acc reveals the accuracy of the model in predicting CSoLC, which can be calculated as

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}. \tag{18}$$

F and R can be expressed as follows:

$$F = \frac{1}{N}\sum\nolimits_{i=1}^{N} F1_i, \tag{19}$$

$$R = \frac{1}{N}\sum\nolimits_{i=1}^{N} Recall_i, \tag{20}$$

where $F1$ is regarded as a harmonic mean, and $Recall$ represents the probability of correct prediction in all positive samples. They can be obtained as follows:

$$F1 = \frac{2TP}{2TP + FP + FN}, \tag{21}$$

$$Recall = \frac{TP}{TP + FN}. \tag{22}$$

As the score of evaluation metrics increases, the performance of the model improves.

C. Baselines and experimental analysis

To demonstrate the effectiveness of the proposed LKAN, eleven representative baselines are chosen for comparison, which mainly include the conventional approaches (i.e., random forest (RF), decision tree (DT), and support vector machine (SVM)), and the deep learning based methods (i.e., LSTM, HAN).

1) Traditional methods: For CSoLC, our model is compared to the following traditional machine learning methods built with the hand-crafted features by bag of words (Bow) [51]:

- RF is based on the idea of ensemble learning, by randomly selecting the number of features and training data, and taking out the test label with the highest number of times for the same prediction data as the final prediction label.
- DT stands for the C4.5 tree-induction algorithm, which constructs decision trees based on the local best choices, providing superior performance and interpretability for machine learning.
- SVM [52] maps the data into a high-dimensional space and makes it possible to classify the data by finding a hyperplane in that space.

TABLE V
COMPARISON OF DIFFERENT METHODS ON THE CANCER REPORT CLASSIFICATION. BOLD SCORES REPRESENT THE BEST RESULTS AND UNDERLINED SCORES REPRESENT THE SECOND-BEST RESULTS

Method		IIa			IIb			IIIa			IIIb			Acc	F1	Recall
		Pr	Sens	Spec	Pr	Sens	Spec	Pr	Sens	Spec	Pr	Sens	Spec			
RF		0.36	0.36	0.99	0.00	0.00	**1.00**	0.23	0.23	0.93	0.88	0.88	0.19	0.41	0.34	0.43
DT		0.67	0.36	0.98	0.42	0.67	0.68	0.48	0.40	0.82	0.71	0.59	0.87	0.53	0.52	0.50
SVM		0.36	0.36	0.99	0.43	0.43	0.85	0.57	0.57	0.80	0.73	0.73	0.74	0.57	0.55	0.61
LSTM		0.50	0.71	0.95	0.73	0.37	0.95	0.59	0.81	0.75	0.74	0.73	0.85	0.66	0.62	0.65
HAN		0.42	0.71	0.93	0.77	0.67	0.93	0.80	0.77	0.91	0.82	0.81	0.89	0.76	0.71	0.74
BERT		0.55	0.85	0.95	0.59	0.53	0.87	0.72	0.64	0.89	0.68	0.73	0.79	0.66	0.65	0.69
BioBERT		0.42	0.71	0.94	0.48	0.33	0.87	0.50	0.11	0.95	0.45	0.77	0.42	0.45	0.42	0.48
DualCL		0.50	**1.00**	0.93	**0.96**	0.80	0.98	**0.91**	**0.83**	**0.96**	0.84	0.86	0.90	0.85	0.82	0.87
KPT		0.11	**1.00**	0.50	0.00	0.00	**1.00**	0.44	0.67	0.62	0.00	0.00	**1.00**	0.26	0.18	0.42
Chat-GPT	1-shot	0.67	0.29	0.99	0.00	0.00	**1.00**	0.17	0.11	0.75	0.46	**0.95**	0.34	0.41	0.29	0.34
	5-shot	**1.00**	0.14	**1.00**	0.50	0.03	0.99	0.21	0.17	0.73	0.46	0.91	0.37	0.41	0.28	0.31
Huatuo-GPT-II	1-shot	0.09	0.57	0.63	0.32	0.27	0.80	0.18	0.11	0.78	0.52	0.30	0.84	0.25	0.24	0.31
	5-shot	0.10	0.57	0.68	0.29	0.17	0.86	0.33	0.19	0.83	0.56	0.51	0.77	0.33	0.29	0.36
LKAN		0.86	0.85	0.99	0.90	**0.93**	0.97	0.88	0.80	0.95	**0.91**	**0.95**	0.94	**0.90**	**0.90**	**0.90**

2) Neural network methods: For deep learning based approaches, a series of experiments with the following models are carried out.

- LSTM is a memorable recurrent neural network, which helps to maintain the long-term dependency when processing longer sentences. LSTM's hidden units are set to 100, with 0.2 recurrent dropout probability.
- HAN [24] is a deep learning method that calculates attention scores at word- and sentence-level to discover context-aware content. After several experiments, the model achieves the highest accuracy when the parameters are set as follows: The dimension of word embedding is set to 200, and the sentence length is specified to 150. The dropout in word encoder is 0.2, and the dropout in sentence encoder is 0.3.
- Bidirectional encoder representation from transformers (BERT) [28] is a popular pre-training language model, which can bring improvement in cross-domain tasks. We follow the fine-tune method in [53], and try to adapt BERT to the special field of liver cancer, with the following parameter configurations: learning rate is 1e-5, and the selected pre-training model is bert-base-Chinese[1].
- BioBERT [54] is a biomedical version of BERT and provides domain-specific representation vectors for biomedical tasks. We fine-tune the BioBERT to adapt our dataset by configuring the following parameters: learning rate is 1e-5, and the selected pre-training model is biobert-v1.1[2].
- DualCL [55] takes BERT as a basic encoder for generating representations, and introduces external label information to enrich views of input data in low-resource scenarios. The pre-training language model is bert-base-Chinese[1], and other default parameters are not changed.
- KPT [56] tunes the pre-training language model by constructing task-specific prompts in domain-specific low-resource scenarios. The pre-training language model is roberta-chinese-base[3], and other default parameters are not changed.
- ChatGPT [57] leverages the GPT architecture for open-domain conversational AI. Unlike task-specific models, it excels without fine-tuning, offering flexibility and fluency in diverse conversations. Its robustness and scalability make it suitable for medical diagnosis.
- HuatuoGPT-II [58] is a large language model[4] pre-trained on extensive Chinese medical knowledge, and is designed for advanced applications in healthcare.

In this section, the performance evaluation of the proposed model is conducted on the liver cancer dataset. We compare our proposed method with a wide range of the baselines models and analyze the experimental results.

From Table V, it can be seen that compared to the basic machine learning models (i.e. RF, DT, and SVM), SVM performs better. However, their reliance on handcrafted features makes them perform worse than the deep models (i.e. LSTM, HAN, BERT, BioBERT, DualCL, and KPT). This is because static handcrafted features lack contextual information, making it difficult for the model to learn deep and separable text features during the staging process. Therefore, traditional models cannot be well applied to imbalanced datasets.

Besides, we attempt to use various pre-training language models combined with fine-tuning methods (i.e. BERT, BioBERT, DualCL, and KPT) for the CSoLC task, among which DualCL performs better. Table. V reveals that the accuracy of BERT is similar to that of LSTM, while the performance of BioBERT after pre-training with medical corpus is even worse. This is because BERT's tokenizer fragments out-of-vocabulary words into subwords [32], and the difference between the source and target domains will cause the model to lose decisive information during the encoding process. Although BioBERT was trained in the biomedical domain, its corpus consists of English literature, which is fundamentally different from our Chinese dataset. Such pre-

[1]https://huggingface.co/bert-base-chinese
[2]https://huggingface.co/dmis-lab/biobert-v1.1
[3]https://huggingface.co/clue/roberta_chinese_base
[4]https://github.com/FreedomIntelligence/HuatuoGPT-II

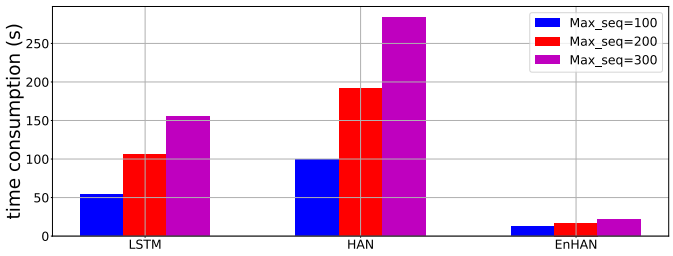
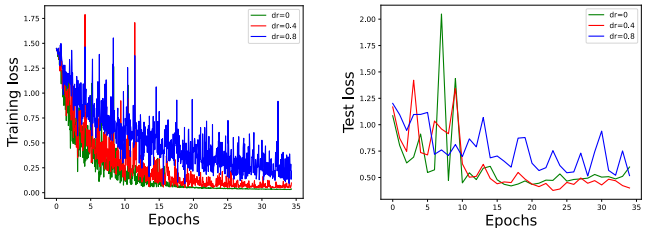


Fig. 9. Complexity analysis.



(a) Training loss (b) Test loss

Fig. 10. The loss curve under the different dropout settings.

training language models will lead to catastrophic negative transfer. KPT also has similar problems. It is useful to provide prompt templates to prompt the pre-training models, but their relevant domain knowledge is scarce when facing special domains such as biomedical. KPT's ability to identify patients with stages IIb or IIIb is weak, meaning that every patient will be missed (Sens=0.00 and Spec=1.00). On the contrary, KPT exhibits high sensitivity but low specificity and precision in identifying stage IIa, indicating that a majority of samples from other stages are incorrectly classified as stage IIa (Sens = 1.00 and Spec = 0.50). Additionally, the model's ability to accurately identify stage IIa is also inadequate, leading to catastrophic misdiagnoses (Pr = 0.11). Although dualCL introduces additional label information, its ability to diagnose patients in stage IIa remains inferior to BERT (Pr = 0.50).

Furthermore, LLMs combined with few-shot learning are adopted for staging performance comparison. Compared to the method utilizing pre-trained language models with few-shot prompts (i.e., KPT), large language models have a wider range of knowledge and more optimized language comprehension abilities. However, the performance of the universal large model ChatGPT on the staging task is slightly better than KPT, and the medical-specific large language model HuatuoGPT-II performs similarly to KPT. The knowledge within the LLMs is insufficient to support them in medical decision-making for the complex staging task.

For the issues that occur when applying the above models to CSoLC, we aim to propose a method that can alleviate these adverse effects and identify precise staging results. Compared with the baseline models, the classification accuracy of LKAN has achieved the best results with 90.3% Accuracy, 90.0% Macro_F1 score, and 90.0% Macro_Recall.

D. Complexity analysis

This section presents the computational complexity. Due to the difficulty of deep models in calculating time complexity, overall time consumption is provided by us to demonstrate that LKAN is rapid and effective.

To compare the various algorithms' timings, we conduct every experiments with same parameter settings. Fig. 9 depicts the time of all models on our dataset. This clearly shows that HAN is the most time-consuming due to its hierarchical attention calculation, followed by LSTM's step-by-step computation. Besides, as the length of the sequence increases, the calculation time will also increase significantly. When the model is applied to medical decision-making, shorter computation time is needed. In contrast, LKAN retains the advantages of hierarchical models, but the running time is not excessive affected by sequence length. And under the premise of the same sequence length, the training time of LKAN is even much smaller than the simplest model LSTM.

E. Overfitting analysis

In this section, we investigate the impact of various dropout ratios dr on overfitting.

As the results are shown in Fig. 10, we find that the training loss and the test loss can not achieve converge when setting dr = 0, and the test loss curve oscillates back and forth between 0.5 and 1.0. This is because removing the dropout layer will reduce the diversity of the model, which also leads to weak generalization ability and difficulty in adapting to the test dataset. From Fig. 10 (a), it can be seen that compared to dr = 0, the use of dropout can accelerate the model's learning process.

Specifically, when dr is set to 0.8, the training loss curve converges earlier than the others and there is no significant fluctuation during the training process, but it shows poor convergence performance on the test set. This means that more neurons are discarded, making the training parameters sparse and leading to the model difficult to stabilize. Therefore, setting a proper dropout value can significantly improve the performance of LKAN. Experiments indicate that LKAN achieves best performance when setting dr = 0.4.

F. Ablation study

A consensus is that oversampling is a useful strategy for overcoming the difficulty of data scarcity. To demonstrate the effectiveness of AugUIE, three oversamplers are applied for comparison (i.e., RANDOM, KMEANS, and AugGPT).

- RANDOM: The minority categories are randomly sampled until all categories are evenly distributed.
- KMeans [59]: A simple and usable oversampler for generating more synthetic samples in sparse regions, which effectively alleviates category imbalance and reduces noise generation.
- AugGPT [35]: ChatGPT with superior comprehension ability and rich knowledge storage, is an excellent tool for generating augmented samples even in very specific and low-resource domains.

Table V shows that our proposed LKAN performs well in recognizing and staging four-class radiology reports. In par-

TABLE VI
ABLATION STUDY: COMPARISON BETWEEN LKAN AND ITS VARIANTS

Method	IIa			IIb			IIIa			IIIb			Acc	F1	Recall
	Pr	Sens	Spec	Pr	Sens	Spec	Pr	Sens	Spec	Pr	Sens	Spec			
LKAN-OS	0.67	0.85	0.97	0.66	0.83	0.84	0.88	0.62	0.96	0.80	0.79	0.87	0.76	0.75	0.78
LKAN-OS+RANDOM	0.50	0.71	0.95	0.81	0.70	0.94	0.74	0.74	0.88	0.80	0.81	0.87	0.75	0.72	0.74
LKAN-OS+KMEANS	0.50	0.85	0.94	**0.96**	0.80	**0.98**	0.78	**0.88**	0.88	0.90	0.79	0.94	0.82	0.83	0.83
LKAN-OS+AugGPT	0.60	0.85	0.96	0.86	0.80	0.95	0.77	0.77	0.90	0.79	0.77	0.87	0.78	0.79	0.78
LKAN-DK+BERT	0.50	0.71	0.95	0.84	0.90	0.94	0.93	0.77	0.97	0.87	0.88	0.87	**0.91**	0.85	0.85
LKAN-GAP-GMP	0.46	0.85	0.93	0.64	0.76	0.84	0.52	0.37	0.85	0.81	0.77	0.88	0.65	0.65	0.66
LKAN-CE+FL	0.83	**1.00**	0.95	0.85	**0.93**	0.92	**0.95**	0.66	**0.99**	**0.91**	0.88	**0.96**	0.88	0.87	0.87
LKAN	**0.86**	0.85	**0.99**	0.90	**0.93**	0.97	0.88	0.80	0.95	**0.91**	**0.95**	0.94	0.90	**0.90**	**0.90**

ticular, LKAN is the first model to process free-text radiology reports and implement CSoLC.

Ablation study is conducted to validate the effectiveness of different methods in our LKAN. In particular, we compared the proposed LKAN with seven variant methods:

- LKAN-OS: We remove the part of over sampling method AugUIE on the basis of LKAN to examine its effectiveness of balanced categories.
- LKAN-OS+RANDOM: We remove the part of over sampling method AugUIE and add RANDOM sampling method on the basis of LKAN for comparison.
- LKAN-OS+KMEANS: We remove the part of over sampling method AugUIE and add KMEANS sampling method on the basis of LKAN for comparison.
- LKAN-OS+AugGPT: We remove the part of over sampling method AugUIE and apply AugGPT as the sampler on the basis of LKAN for comparison.
- LKAN-DK+BERT: We remove the part of domain-specific word embedding and utilize BERT as knowledge encoder on the basis of LKAN to examine its effectiveness of domain knowledge.
- LKAN-GAP-GMP: We remove the GAP and GMP of the classifier on the basis of LKAN to examine its effectiveness.
- LKAN-CE+FL: We remove the cross-entropy and apply focal loss as the loss function.

From Table VI, we first evaluate the effectiveness of AugUIE by the comparison with four variants. Its results indicate that the simple RANDOM method can balance the class distribution, but it deepens overfitting and leads to a decrease in classification performance. KMEANS generates minority categories based on cluster density from different clusters, which effectively alleviates noise in oversampling process and helps the classifier to distinguish the imbalanced categories. However, traditional methods for measuring the similarity between two nodes are still limited, as they are difficult to ensure the semantic consistency of the original sentences and cannot bring rich features to the minority data. With the development of technology, LLMs trained from rich knowledge can effectively alleviate the above problems through appropriate fine-tuning. Compared to the other methods, AugGPT can bring excellent improvement to the least number of categories (IIa). Due to the domain sensitivity of the medical domain, AugGPT finds it difficult to bridge the gap between general knowledge and medical knowledge solely by relying on input training text to prompt ChatGPT to generate auxiliary samples. Although current LLMs can generate a large number of diverse sentences for various routine tasks, their relevant knowledge in clinical staging problems is not sufficient to support their mastery of internal medical rules and the generation of convincing training data. Therefore, there is no significant difference in overall classification performance between AugGPT and RANDOM.

The variants (i.e., LKAN-OS, LKAN-OS+RANDOM, LKAN-OS+KMEANS, LKAN-OS+AugGPT, LKAN-DK+BERT, and LKAN-GAP-GMP) demonstrate lower precision in identifying stage IIa. This indicates that the model will exhibit a significantly high misdiagnosis rate when encountering patients with stage IIa, which hinders the effective identification of the true staging of patients. Focal loss is considered beneficial for imbalanced datasets. However, LKAN-CE+FL exhibits inadequate performance in identifying stage IIIa, with only a 66.0% sensitivity rate. This implies that a significant portion of patients with stage IIIa will be misdiagnosed. Pre-training word weights can effectively alleviate the problem of imbalanced data, but when there is a significant difference between the pre-training corpus and downstream tasks, even larger computational parameters cannot bring positive effects. LKAN-DK+BERT shows that the problem of bias towards most categories still exists in model prediction.

After ablation study, Table VI shows the effectiveness of LKAN.

IV. DISCUSSION AND VISUAL ANALYSIS

We discuss the three aspects of domain-specific word embedding, attention, and the limitations of LKAN.

A. Discussion of domain-specific word embedding

In this part, we try to discuss why constructing domain-specific word embeddings plays an important role in medical tasks. Therefore, we use t-distributed stochastic neighbor embedding (t-SNE) [60] algorithm to visualize and analyze the distribution and representation of words in the radiology reports by mapping word embeddings onto 3-dimensional space. We aim to show that embeddings trained on the radiology-specific corpus could offer better performance than the general corpus.

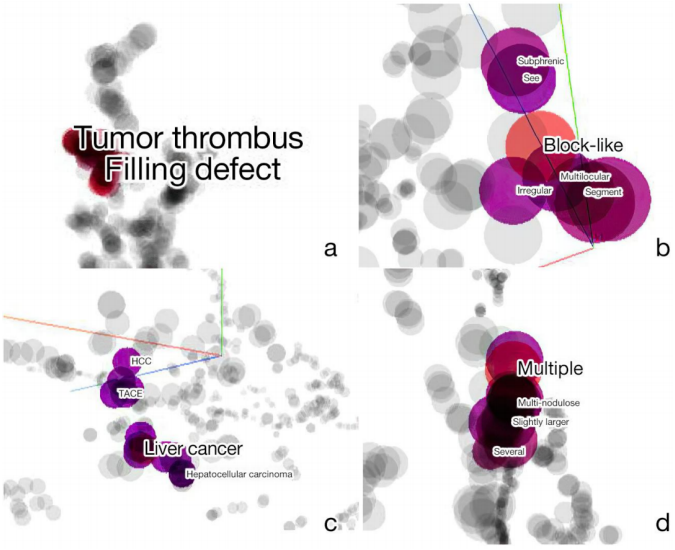


Fig. 11. Visualization of t-SNE distribution of word embedding.

A huge mass exists in the right lobe of the liver , which is considered primary liver cancer . There are multiple lesions in the liver, combined with the formation of cancer thrombi in the main, left, and right branches of the portal vein . Having portal varicose veins (cavernous transformation of the hepatic portal vein) . There is lymph node metastasis in the hepatic hilum area . The greater omentum and mesenteric space are turbid , and multiple lymph nodes are enlarged . The retroperitoneal lymph nodes are enlarged , which is considered the possibility of metastasis . Having cirrhosis with small amounts of ascites ; Varicocele in the lower esophagus and around the stomach fundus. cholecystitis exists . Multiple fibrous lesions were found in the upper and lower tongue segments of the lung's right middle lobe and left upper lobe . Several small nodules in both lungs suggest the possibility of inflammatory lesions . It is recommended to have a follow-up examination .

Fig. 12. Visualization of attention weights. The English translation is for illustration. Visualization of the attention of the corresponding Chinese radiology reports in Fig. 4, in Supplemental Material.

Fig. 11 clearly shows the distribution of pre-training word2vec embeddings, in which we can see the correlation between words. According to the context, words with similar meanings should be close to each other. Fig. 11 (a) and (c) show radiology terms such as "Tumor thrombus" and "Filling defect", both of which are used to describe whether liver cancer is already in stage IIIa, so they are relatively close together. From Fig. 11 (b), we can see that "Block-like" is tight to "Irregular", which also conforms to the fact that a massive hepatocellular carcinoma is irregular. In radiology terminology, quantitative terms such as "Multiple", "Multi-nodulose", and "Several" have the same meaning, and they all often express that hepatocellular carcinoma has metastasized and spread within the liver, which is the symptom of stage IIb. Fig. 11 (d) shows that they are adjacent to each other.

B. Discussion of attention

To validate our model's ability to select critical sentences and words in documents, we visualize the attention layers of several documents in the liver cancer dataset. From Fig. 12 can be seen that each colored block represents a word in the training sentence. Different shades of red represent attention weights. As the color deepens, this word's influence on model decision-making becomes greater.

Fig. 12 shows that our model can choose medical terms with clear indications, such as "cancer thrombi", "metastasis", "huge mass", "lymph nodes", and "enlarged". The selection of punctuation will not be paid too much attention. Words that are not very important or unrelated to CSoLC, such as "cavernous", "stomach" and "cholecystitis", will be paid little weight by LKAN. By visualizing attention weights, we also gain a rough understanding of the model's ability to focus on the important information.

The attention mechanism can make the model focus on important words, but in the case of the imbalanced dataset, it is difficult for the model to find discriminative of the minority data. Therefore, using the preprocessing method of regular expressions to reduce noise in text is beneficial for guiding the model to learn the important knowledge. As shown in Fig. 13, LKAN without preprocessing cannot distinguish important words and applies higher weights to the words unrelated to CSOLC, such as "the portal vein", "transformation", and "cirrhosis".

Table VII compares the results obtained from LKAN, where we achieve data augmentation by concatenation of information from GAP, GMP, and attention. A single attention mechanism can focus on important information in the context, while GAP and GMP retain vital features while reducing dimensions. This find shows that the concatenation of information from attention, GAP, and GMP plays a significant role in data augmentation. Aggregation of multiple types of information can improve model classification performance.

A huge mass exists in the right lobe of the liver , which is considered primary liver cancer . There are multiple lesions in the liver, combined with the formation of cancer thrombi in the main, left, and right branches of the portal vein . Having portal varicose veins (cavernous transformation of the hepatic portal vein) . There is lymph node metastasis in the hepatic hilum area . The greater omentum and mesenteric space are turbid , and multiple lymph nodes are enlarged . The retroperitoneal lymph nodes are enlarged , which is considered the possibility of metastasis . Having cirrhosis with small amounts of ascites ; Varicocele in the lower esophagus and around the stomach fundus. cholecystitis exists . Multiple fibrous lesions were found in the upper and lower tongue segments of the lung's right middle lobe and left upper lobe . Several small nodules in both lungs suggest the possibility of inflammatory lesions . It is recommended to have a follow-up examination .

Fig. 13. Visualization of the influence of redundant information on the classification results.

TABLE VII
COMPARISON OF GAP, GMP, AND ATTENTION ON FEATURE ENHANCEMENT

Methods	F1	Acc
ATT	0.66	0.66
GAP, GMP	0.88	0.87
GAP, GMP, ATT	0.90	0.90

C. Limitations of LKAN

Our experiment focuses on the radiology reports, mining diagnostic information from them for CSoLC, which provides a solution for the dependence on LROIs in medical images. However, at present, due to the difficulty in obtaining medical data, there are few studies on mining text information from cancer images and applying it to real-world problems, which also brings many limitations to our research. 1) There is no relevant public dataset, making it difficult to verify the generalization of our method. 2) There are few studies related to cancer staging, and studies using the free-text radiology reports are especially scarce, thus leading to fewer comparative experiments. 3) In terms of staging, CNLC divides liver cancer into 6 stages. However, due to practical reasons, most patients are already in the end-stage when symptoms appear. Therefore, it is difficult for us to collect data on the early-stage of liver cancer, which is also a major regret for us.

The shortcomings of the method caused by the above issues will be tackled in our future work. Finally, the accuracy of our LKAN in the dataset of the First Affiliated Hospital of Sun Yat-sen University has reached 90.3%, which still leaves much room for improvement.

V. Conclusion and Future Work

In this paper, we propose a deep learning method to solve the CSoLC task. Faced with the imbalanced categories, the over sampling method AugUIE is proposed for alleviating the problem of the skewed majority in the prediction of CSoLC. The Chinese radiology reports of liver cancer have strong domain knowledge. Domain-specific vocabulary and pre-trained word embedding are constructed to relieve out-of-vocabulary and increase the association relationships between various medical terms, providing better semantic relationships for subsequent model classification. The attention block of HAN is improved by incorporating both global and local features to substantially improve the ability of the deep model to identify precise CSoLC. The experimental results indicate that the proposed method achieves the state-of-the-art performance in CSoLC using the dataset provided by the hospital, which still leaves much room for improvement. In our future work, we will enhance the correlation between the indicators of CNLC and achieve the classification of stage I tumors on the limited dataset.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] H. Rumgay, M. Arnold, J. Ferlay, O. Lesi, C. J. Cabasag, J. Vignat, M. Laversanne, K. A. McGlynn, and I. Soerjomataram, "Global burden of primary liver cancer in 2020 and predictions to 2040," *Journal of Hepatology*, vol. 77, pp. 1598–1606, 2022.

[3] L. Sun, Y. Yang, Y. Li, B. Zhang, and R. Shi, "The past, present, and future of liver cancer research in China," *Cancer Letters*, vol. 574, p. 216334, 2023.

[4] S. Katyal, J. H. Oliver III, M. S. Peterson, J. V. Ferris, B. S. Carr, and R. L. Baron, "Extrahepatic metastases of hepatocellular carcinoma," *Radiology*, vol. 216, no. 3, pp. 698–703, 2000.

[5] C. Yang, H. Zhang, L. Zhang, A. X. Zhu, R. Bernards, W. Qin, and C. Wang, "Evolving therapeutic landscape of advanced hepatocellular carcinoma," *Nature Reviews Gastroenterology & Hepatology*, vol. 20, no. 4, pp. 203–222, 2023.

[6] S. Q. Sin, C. D. Mohan, R. M. W.-J. Goh, M. You, S. C. Nayak, L. Chen, G. Sethi, K. S. Rangappa, and L. Wang, "Hypoxia signaling in hepatocellular carcinoma: Challenges and therapeutic opportunities," *Cancer and Metastasis Reviews*, vol. 42, no. 3, pp. 741–764, 2023.

[7] J. Zhou, H. Sun, Z. Wang, W. Cong, J. Wang, M. Zeng, W. Zhou, P. Bie, L. Liu, T. Wen *et al.*, "Guidelines for the diagnosis and treatment of hepatocellular carcinoma (2019 edition)," *Liver Cancer*, vol. 9, no. 6, pp. 682–720, 2020.

[8] S. Lobanov-Rostovsky, Q. He, Y. Chen, Y. Liu, Y. Wu, Y. Liu, T. Venkatraman, E. French, N. Curry, N. Hemmings, P. Bandosz, W. K. Chan, L. Jing, and E. J. Brunner, "Growing old in china in socioeconomic and epidemiological context: systematic review of social care policy for older people," *BMC Public Health*, vol. 23, no. 1, pp. 1272–1294, 2023.

[9] J. Ye, L. He, and M. Beestrum, "Implications for implementation and adoption of telehealth in developing countries: a systematic review of china's practices and experiences," *NPJ Digital Medicine*, vol. 6, no. 1, pp. 174–186, 2023.

[10] Y. Pan, W. Huang, Z. Lin, W. Zhu, J. Zhou, J. Wong, and Z. Ding, "Brain tumor grading based on neural networks and convolutional neural networks," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2015, pp. 699–702.

[11] R. Patil and S. Bellary, "Machine learning approach in melanoma cancer stage detection," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3285–3293, 2022.

[12] C. De Vente, P. Vos, M. Hosseinzadeh, J. Pluim, and M. Veta, "Deep learning regression for prostate cancer detection and grading in bi-parametric MRI," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 2, pp. 374–383, 2020.

[13] P. Huang, X. Tan, X. Zhou, S. Liu, F. Mercaldo, and A. Santone, "FABNet: fusion attention block and transfer learning for laryngeal cancer tumor grading in P63 IHC histopathology images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1696–1707, 2021.

[14] M. Fan, W. Yuan, W. Zhao, M. Xu, S. Wang, X. Gao, and L. Li, "Joint prediction of breast cancer histological grade and Ki-67 expression level based on DCE-MRI and DWI radiomics," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1632–1642, 2019.

[15] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, R. Yang *et al.*, "Annotation-efficient deep learning for automatic medical image segmentation," *Nature Communications*, vol. 12, no. 1, pp. 5915–5927, 2021.

[16] V. M. D'Anniballe, F. I. Tushar, K. Faryna, S. Han, M. A. Mazurowski, G. D. Rubin, and J. Y. Lo, "Multi-label annotation of text reports from computed tomography of the chest, abdomen, and pelvis using deep learning," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–12, 2022.

[17] B. Felfeliyan, N. D. Forkert, A. Hareendranathan, D. Cornel, Y. Zhou, G. Kuntze, J. L. Jaremko, and J. L. Ronsky, "Self-supervised-RCNN for medical image segmentation with limited data annotation," *Computerized Medical Imaging and Graphics*, vol. 109, p. 102297, 2023.

[18] E. Pons, L. M. Braun, M. M. Hunink, and J. A. Kors, "Natural language processing in radiology: a systematic review," *Radiology*, vol. 279, no. 2, pp. 329–343, 2016.

[19] K. Yan, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers, "Holistic and comprehensive annotation of clinically significant findings on diverse CT images: learning from radiology reports and label ontology," in *CVPR*, 2019, pp. 8523–8532.

[20] H.-Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, and Y. Yu, "Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports," *Nature Machine Intelligence*, vol. 4, no. 1, pp. 32–40, 2022.

[21] M. A. Fink, V. L. Mayer, T. Schneider, C. Seibold, R. Stiefelhagen, J. Kleesiek, T. F. Weber, and H.-U. Kauczor, "CT angiography clot burden score from data mining of structured reports for pulmonary embolism," *Radiology*, vol. 302, no. 1, pp. 175–184, 2022.

[22] M. A. Fink, K. Kades, A. Bischoff, M. Moll, M. Schnell, M. Küchler, G. Köhler, J. Sellner, C. P. Heussel, H.-U. Kauczor *et al.*, "Deep learning–based assessment of oncologic outcomes from natural language processing of structured radiology reports," *Radiology: Artificial Intelligence*, vol. 4, no. 5, p. e220055, 2022.

[23] S. Eskreis-Winkler, E. J. Sutton, D. D'Alessio, K. Gallagher, N. Saphier, J. Stember, D. F. Martinez, E. A. Morris, and K. Pinker, "Breast MRI background parenchymal enhancement categorization using deep

learning: outperforming the radiologist," *Journal of Magnetic Resonance Imaging*, vol. 56, no. 4, pp. 1068–1076, 2022.

[24] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *ACL*, 2016, pp. 1480–1489.

[25] K. Uchino, R. Tateishi, S. Shiina, M. Kanda, R. Masuzaki, Y. Kondo, T. Goto, M. Omata, H. Yoshida, and K. Koike, "Hepatocellular carcinoma with extrahepatic metastasis: clinical features and prognostic factors," *Cancer*, vol. 117, no. 19, pp. 4475–4483, 2011.

[26] H.-X. Gu, X.-S. Huang, J.-X. Xu, P. Zhu, J.-F. Xu, and S.-F. Fan, "Diagnostic value of MRI features in dual-phenotype hepatocellular carcinoma: A preliminary study," *Journal of Digital Imaging*, vol. 36, no. 6, pp. 2554–2566, 2023.

[27] X. Li, Y. Meng, X. Sun, Q. Han, A. Yuan, and J. Li, "Is word segmentation necessary for deep learning of Chinese representations?" in *ACL*, 2019, pp. 3242–3252.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[30] W. Liao, Z. Liu, H. Dai, Z. Wu, Y. Zhang, X. Huang, Y. Chen, X. Jiang, D. Liu, D. Zhu, S. Li, W. Liu, T. Liu, Q. Li, H. Cai, and X. Li, "Mask-guided BERT for few-shot text classification," *Neurocomputing*, vol. 610, p. 128576, 2024.

[31] A. Gupta, K. Thadani, and N. O'Hare, "Effective few-shot classification with transfer learning," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1061–1066.

[32] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2021.

[33] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[34] V. Kumar, H. Glaude, C. de Lichy, and W. Campbell, "A closer look at feature space data augmentation for few-shot intent classification," *arXiv preprint arXiv:1910.04176*, 2019.

[35] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, and X. Li, "AugGPT: Leveraging ChatGPT for text data augmentation," 2023.

[36] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," *arXiv preprint arXiv:2105.03075*, 2021.

[37] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–39, 2022.

[38] Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, and H. Wu, "Unified structure generation for universal information extraction," in *ACL*, 2022, pp. 5755–5772.

[39] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu *et al.*, "ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv:2107.02137*, 2021.

[40] Y. Li, H. Guo, Q. Zhang, M. Gu, and J. Yang, "Imbalanced text sentiment classification using universal and domain-specific knowledge," *Knowledge-Based Systems*, vol. 160, pp. 1–15, 2018.

[41] B. Wang, Q. Xie, J. Pei, Z. Chen, P. Tiwari, Z. Li, and J. Fu, "Pre-trained language models in biomedical domain: A systematic survey," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–52, 2023.

[42] V. Kalra, I. Kashyap, and H. Kaur, "Generation of domain-specific vocabulary set and classification of documents: weight-inclusion approach," *International Journal of Information Technology*, pp. 1–11, 2022.

[43] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas, "Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario," *arXiv preprint arXiv:2109.03570*, 2021.

[44] Z. Li and M. Sun, "Punctuation as implicit annotations for Chinese word segmentation," *Computational Linguistics*, vol. 35, no. 4, pp. 505–512, 2009.

[45] Y. Peng, K. Yan, V. Sandfort, R. M. Summers, and Z. Lu, "A self-attention based deep learning method for lesion attribute detection from CT reports," in *2019 IEEE International Conference on Healthcare Informatics*, 2019, pp. 1–5.

[46] Y. Liu, D. Yuan, H. Fan, T. Jin, and M. A. Mohamed, "A multidimensional feature-driven ensemble model for accurate classification of complex power quality disturbance," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[47] A.-S. Mohammad, M. M. Hammad, A. Sa'ad, A.-T. Saja, and E. Cambria, "Gated recurrent unit with multilingual universal sentence encoder for arabic aspect-based sentiment analysis," *Knowledge-Based Systems*, p. 107540, 2021.

[48] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 5, pp. 1–32, 2021.

[49] Y. Xu, Z. Yu, W. Cao, and C. P. Chen, "Adaptive dense ensemble model for text classification," *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 7513–7526, 2022.

[50] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[51] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, pp. 62:1–62:40, 2022.

[52] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[53] L. Yao, Z. Jin, C. Mao, Y. Zhang, and Y. Luo, "Traditional Chinese medicine clinical records classification with BERT and domain specific corpora," *Journal of the American Medical Informatics Association*, vol. 26, no. 12, pp. 1632–1636, 2019.

[54] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[55] Q. Chen, R. Zhang, Y. Zheng, and Y. Mao, "Dual contrastive learning: Text classification via label-aware data augmentation," *CoRR*, vol. abs/2201.08702, 2022.

[56] S. Hu, N. Ding, H. Wang, Z. Liu, J. Wang, J. Li, W. Wu, and M. Sun, "Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification," in *ACL*, 2022, pp. 2225–2240.

[57] OpenAI, "ChatGPT," https://openai.com/blog/chatgpt, 2023.

[58] J. Chen, X. Wang, A. Gao, F. Jiang, S. Chen, H. Zhang, D. Song, W. Xie, C. Kong, J. Li, X. Wan, H. Li, and B. Wang, "HuatuoGPT-II, one-stage training for medical adaption of LLMs," *CoRR*, vol. abs/2311.09774, 2023.

[59] F. Last, G. Douzas, and F. Bação, "Oversampling for imbalanced learning based on K-Means and SMOTE," *CoRR*, vol. abs/1711.00837, 2017.

[60] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.