IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 October 2022, accepted 24 October 2022, date of publication 1 November 2022, date of current version 9 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3218646

RESEARCH ARTICLE

Effective Natural Language Processing and Interpretable Machine Learning for Structuring CT Liver-Tumor Reports

YI-HSUAN CHUANG[1], JA-HWUNG SU[2], DING-HONG HAN[3], YI-WEN LIAO[4], YEONG-CHYI LEE[5], YU-FAN CHENG[1], TZUNG-PEI HONG[2,3], (Senior Member, IEEE), KATHERINE SHU-MIN LI[3], (Senior Member, IEEE), HSIN-YOU OU[1], YI LU[1], AND CHIH-CHI WANG[6]

[1]Liver Transplantation Program, Department of Diagnostic Radiology and Surgery, Kaohsiung Chang Gung Memorial Hospital, Chang Gung University College of Medicine, Niao-Sung, Kaohsiung 833401, Taiwan
[2]Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811726, Taiwan
[3]Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan
[4]Department of Intelligent Commerce, National Kaohsiung University of Science and Technology, Kaohsiung 824004, Taiwan
[5]Department of Information Management, Cheng Shiu University, Kaohsiung 833301, Taiwan
[6]Liver Transplantation Center and Department of Surgery, Kaohsiung Chang Gung Memorial Hospital, Niao-Sung, Kaohsiung 833401, Taiwan

Corresponding authors: Ja-Hwung Su (bb0820@ms22.hinet.net) and Yu-Fan Cheng (prof.chengyufan@gmail.com)

This work was supported by the Ministry of Science and Technology, Republic of China, under Grant MOST 110-2321-B-182A-003 (NZRPG8L0031) and Grant MOST 111-2321-B-182A-003 (NZRPG8L0032).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board, Chang Gung Medical Foundation under IRB No. 202100262B0.

ABSTRACT In the past, the liver tumors were reported manually in an unstructured format. There actually exists much valuable knowledge in these reports for further disease risk assessment, disease recognition and treatment recommendation. Yet, it is not easy to read and mine knowledge from the unstructured reports. Hence, how to extract the knowledge from these biomedical reports effectively and efficiently has been a challenging issue in the past decades. Although a set of Natural Language Processing techniques were proposed for Bio-medical information retrieval, few related works were made on transforming the unstructured CT liver-tumor reports into structured ones. To aim at this issue, in this paper, we propose a two-stage report structuring method by integrating effective Natural Language Processing (NLP) and interpretable machine learning. For the first stage, the candidate keywords in unstructured reports are extracted. Next, the feature keywords are determined by the feature-selection technique. For the second stage, the well-known multi-classifiers are performed, and finally the reports are labeled in a refined structure format. Further, the factor keywords in the classification model are filtered to interpret the performance. In overall, the proposed report structuring method generates a hierarchical data structure, including the common features and refined features in the 1st and 2nd levels/stages, respectively. To reveal the performance of proposed method, a set of evaluations were conducted and the results show that, the proposed method is more promising than the fashion neural networks such as Bert (Bidirectional Encoder Representations from Transformers) in terms of effectiveness and efficiency.

INDEX TERMS Structured reports, natural language processing, interpretable machine learning, CT liver-tumors, biomedical science.

I. INTRODUCTION

Knowledge discovery from unknown multimedia data has been investigated for a long time. Especially, in the past few

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

years, advanced biomedical sciences enable a great increased demand of knowledge discovery. In the field of biomedical science, the major attention is focused on three main paradigms, namely disease risk assessment, disease recognition and treatment recommendation. Whatever the paradigm is, knowledge discovery plays a crucial role. Therefore, how to effectively retrieve the high-value knowledge from multimedia data has been a hot topic recently. Basically, the biomedical multimedia data can be categorized into visual and textual types: biomedical images, biomedical reports and gene expressions. For visual types, the most popular recent research is to recognize the biomedical images through Deep Learning, such as CT Liver tumor recognition. After the visual tumor recognition, the next step is to generate the reports. In the past, most recognition results were reported manually in an unstructured format. The radiologists reported the results in a set of linguistic sentences, as shown in Figure 1-(a). It incurs problems that, the sentences are not easy to quickly catch and the patterns are not easy to extract also. Actually, the patterns can be viewed as valuable knowledge, which are very helpful to the further biomedical predictions and recommendations. For example, the associations between visual, textual and numerical patterns facilitate automated predictions and reports. Hence, how to define the patterns is a critical point. In fact, the definitions of patterns can be viewed as a problem for structuring the reports. To deal with this problem, in this paper, we propose a two-stage structuring method, including Natural Language Processing (NLP) and Machine Learning (ML). In the first stage, the NLP is performed to retrieve the feature keywords (also called patterns in this paper). That is, all reports will be transformed in this formatted structure, while ML serves as a multi-classifier to tag the reports in the second stage, as shown in Figure 1-(b). Finally, the factor keywords are filtered by SHAP [21]. In detail, the feature keywords determined by NLP can be used as the feature dimensions for training, and the factor keywords can be used for optimizing the model. Moreover, there are some complications hidden in the feature keywords. Once the complications are discovered, the related liver symptoms can be bridged.

In the past, there actually have been a number of related researches devoted on the biomedical information retrieval. Yet, few studies paid their attention on structuring the CT-tumor reports. This inspires us to propose a practical method for an effective transformation from unstructured to structured formats. On the whole, the major contributions in this paper fall into the following folds.

- From technical viewpoint, a hierarchical structure is generated by a progressive learning called two-stage structuring in this paper. The uniqueness in contrast to the Bert neural networks is that, the embedding results and the performance factors in the 1st and 2nd stages are readable, respectively.
- From performance viewpoint, it is not easy to keep the trade-off between effectiveness and efficiency for traditional MLs. To aim at this issue, without a high-priced cost, the proposed method will be more effective than the Bert-based neural networks.
- From practical viewpoint, this work is motivated by demands of radiologists in Kaohsiung Branch of Chang Gung Hospital, Taiwan. The major intents behind the demand are: first, there exist rich patterns in the past reports to mine. Second, the automated image recognition needs a huge amount of image data. The huge image data can be tagged automatically by the proposed method. Third, although the department of diagnostic radiology is currently making attempts to conduct a structuring report system, the old un-structuring report system is working still at this transition time.
- From extension viewpoint, in our next work, valuable patterns and associations will be extracted from the structured reports, which further support the pattern recognition in automated predictions and reports.

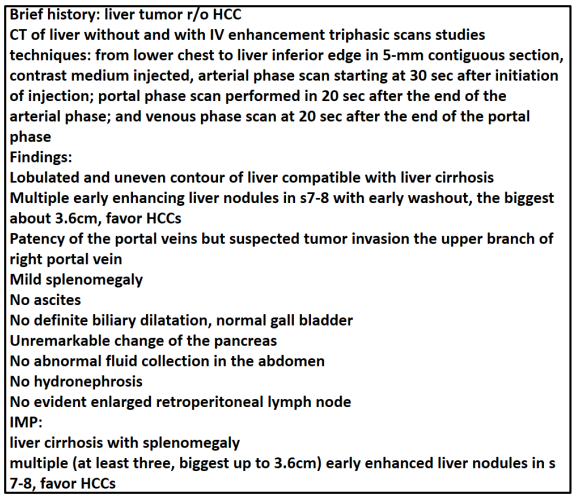


FIGURE 1. Examples of unstructured and structured reports.

To capture the performance of proposed method, a number of evaluations were conducted on a real dataset. The experimental results reveal that, the proposed method is more effective in contrast to the compared methods Bert neural networks. Also, the training cost is much less than that of compared methods. In detail, the feature-keywords in the first stage is sensitive to the ML in the second stage. By integrating these two stages, high-quality structures can be achieved. The rest of this paper is structured as follows: In Section 2, a comprehensive study for previous works will be presented. Next, the details of proposed method with two-stage structuring will be illustrated in Section 3. Then, in Section 4, the evaluation results will be lifted. Finally, conclusions and future works will be shown in Section 5.

II. RELATED WORKS

Biomedical multimedia information retrieval has been a hot topic due to advanced Artificial Intelligence (AI) in recent years. In this study, the main intent is to transform the unstructured reports into structured ones in a regular form. To achieve this purpose, NLP and ML are two core components adopted in the proposed method. In this section, a number of previous studies are reviewed by categories, namely natural language processing, biomedical text information retrieval and explainable machine learning.

A. NATURAL LANGUAGE PROCESSING

Data engineering has been the critical process in recent artificial intelligence techniques. Without effective data engineering, it is not easy to infer the good result. For textual data, Natural Language Processing [43] is the mainstay in the field of data engineering. In traditional, NLP is composed of several components, including tokenizing, stopwords removal, stemming, Term Frequency (TF) calculation, Inverse-Document Frequency (IDF) calculation and n-gram modeling. After these processes, the document features can be extracted for the further applications such as recognition, retrieval and so on. Bojanowski et al. [4] employed subword information to extend the continuous skipgram model. Xue [41] took advantages of Latent Dirichlet Allocation and Word2Vec to calculate the similarities between topics and documents. Ma and Zhang [25] integrated skip-gram and bag-of-words for big data engineering. In addition to the general NLP, Deep Learning-based NLP [28], [35], [36] has been proposed for textual mining. Zhang and Rao [47] fused n-gram models and Bi-LSTM to approximate better classification results. Wang et al. [37] attempted to enhance the skip gram models by Bert. Wang et al. [38] performed Multi-Grained Cascade Forest to classify the texts. Whatever the NLP paradigm is, a number of researches are made on document classification [13], [19]. TFIDF was adopted to classify the texts by Chen [5]. Fast-text oriented researches were studied by Amalia et al. [2] and Yao et al. [44].

B. BIOMEDICAL TEXT INFORMATION RETRIEVAL

Information Retrieval is a popular technique [1], [6], [16], [20] widely used in multimedia applications, such as social media retrieval, search engine, product recommendation and so on. Generally speaking, the related applications for biomedical document/text classifications can be divided into two types, namely traditional classifications [33] and deep learning-based classifications [14]. In traditional classifications, Jamaluddin and Wibawa [18] made use of Support Vector Machine (SVM) for diagnosis classifications. Nguyen et al. [27] compared several traditional classifiers for dutch breast cancer radiology reports, including SVM, Logistic Regression, Ridge Classifier, Gradient Boosted Trees, Random Forest, (RF) K Nearest Neighbors (KNN) and Multinomial Naive Bayes. Xu et al. [42] merged the idea of learning-to-rank into the textual information retrieval for query expansion. Wang et al. [40] proposed Rel-TNG and Type-TNG models, which were combined with Rel-LDA and Type-LDA for semantic relation retrieval from biomedical documents. In deep learning-based methods, the basic ideas include transformer [10], Convolutional Neural Networks (CNN) [23], transfer learning [23], [29], Graph Convolutional Network [39] and so on. Allada et al. [3] made attempts to approximate better word embeddings for a better deep learning-based classification, which can reach the accuracy 79.76% for BioBERT. Bi-LSTM with attention was used for the biomedical text classification [48]. Further, Zhang and Jin [45] merged Bert and graph attention networks to recognize clinical reports, which can reach the accuracy 83.27%. Spandorfer et al. [32] classified the sentences in the CT reports by deep learning. Qiu et al. [30] compared TFIDF-based traditional classifiers and CNN for cancer pathology reports. Chen et al. [8] associated the textual with visual medical-data by using Bert, BiGRU and ResNet.

C. CONTEXT OF CT IMAGE INFORMATION RETRIEVAL

In addition to text information retrieval, a number of past contributions were proposed for context of CT image information retrieval. Dorn et al. [9] proposed a context-sensitive CT imaging scheme which comprised a prior spatial resolution, display and dual energy evaluation. Huynh et al. [12] presented a learning method that predicted the CT image from its corresponding MR image for the same object, and then enhanced the prediction by an auto context model. Jin et al. [17] fused the content and context information for wide areas of medical image retrieval. Li et al. [22] segmented the prostate in a CT image by learning the patient-specific information from the location-adaptive image context. Ma et al. [24] aggregated the context and content similarities for content based medical image retrieval by a weighted graph structure. Nie et al. [26] made attempts to synthesize the CT images for a better connection between CT and MRI images by using a context-aware generative adversarial network. Safaei [31] mined the correlations in medical images based on the users' query logs and conducted a text-based multi-dimensional index for effective medical image retrieval. By using the local context, Zheng et al. [46] combined deep learning and marginal space learning for kidney detection and segmentation.

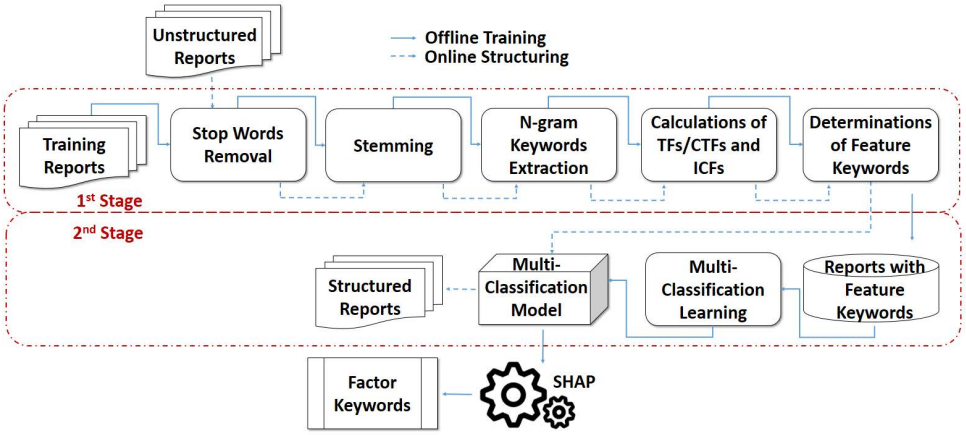


FIGURE 2. Overview of proposed method.

D. EXPLAINABLE MACHINE LEARNING

Although recent machine learning has been approved to be effective for biomedical data, an issue for investigating whys of effectiveness has attracted more and more attention recently. In 2010, Strumbelj and Kononenko [34] proposed an idea for explaining the classification model according to Game Theory. Then, in 2017, Lundberg et al. [21] proposed SHapley Additive exPlanations (SHAP) to explain the predictions. Later, in 2020, Jansen et al. [11] combined Agnostic Explanations (LIME) and SHapley Additive exPlanations to explain the performances of learning models. In 2022, Islam et al. [15] provided explainable classifiers for visual brain states. These works provide us with the idea for exploiting the factors of structuring models.

III. THE PROPOSED METHOD

A. ARCHITECTURE OVERVIEW

As mentioned above, the goal of this paper is to convert the linguistic-sentence format into the regular format, including feature patterns and diagnosis labels. To reach this goal, as shown in Figure 2, the proposed method is decomposed into two phases, namely offline training and online structuring. Further, in each phase, two-stage operations with respect to Natural Language Processing and Machine Learning are performed. Finally, the factor keywords and structured reports will be generated.

1) OFFLINE TRAINING PHASE

Basically, this phase can be divided into two stages. In the first stage, NLP is performed to determine the feature keywords. These feature keywords can be regarded as the main structured patterns. To this end, in this stage, the stop words of each training report are removed first. Next, n-gram keywords are extracted and stemmed. Then, the Term Frequency (named TF) and Inverse Class Frequency (named ICF) for each keyword are calculated. Based on the TFs and ICFs, the top-z keywords are selected as the feature keywords, which are also used as the feature vectors for training the multi-classifiers. In the second stage, the classification model will be constructed and the factor keywords to the classification model will be further extracted by SHAP [21].

2) ONLINE STRUCTURING PHASE

This phase works online for structuring the unstructured reports, consisting of two stages. The goal of the first stage is to transform an unstructured report into a preliminarily structured report called NLP-structured report, while that of the second stage is to classify the NLP-structured report into multiple high-level diagnosis labels. To these ends, in the first stage, each unstructured report is processed by word stemming and stop word removing. Next, on the basis of the determined feature keywords, the related TFs and ICFs are calculated as the features. Therefore, each unstructured report is represented by a feature keyword vector. In the second stage, the NLP-structured reports are further classified into a set of diagnosis labels by the training model. Finally, it will be formatted a regular structure defined by the doctors, as shown in Figure 1-(b).

B. CORE CONCEPT

Before presenting the method details in the succeeding section, in this section, the main concept is shown to make the proposed method easy to catch. As recalled from Figure 1-(b), the expected output is a regular structure including 4 primary items: tumor number/location, tumor size, tumor characteristics and associated liver features. Because the tumor number/location and tumor size can be achieved by basic text parsing, it is not the aimed problem to resolve in this paper. Except the first and second items, the third and fourth items are composed of 12 patterns with respect to {Early arterial enhancement, Early washout, Enhancing capsule, Threshold growth} and {Vascular invasion, Portal vein tumor thrombus, Extrahepatic spread, Splenomegaly, Liver cirrhosis, Ascites, Portosystemic collateral vessel, Portal vein thrombosis}, respectively. If these 12 patterns are flattened, it can be regarded as a

multi-classification problem. This is the first concept to show. Second, after the problem is defined, the next challenge is how to purse the near-optimal features for a better multi-classification. Indeed, recent popular Bert-type neural networks were considerable methods for this work. However, two weak points triggered us to create a better solution shown in this paper. First, it is not easy to extend the results to our next intent for automated predictions and reports, for example, associations among visual, textual and numerical patterns. Second, the model needs a high-priced training cost. To attack these problems, the two-stage multi-classification is proposed in this paper. For the first problem, the first-stage structure yields a set of feature keywords with TFs and ICFs, where TF indicates the representative features and ICF indicates the discriminative features. Moreover, these keywords could be used as the patterns to describe the symptoms in the second stage, even linking the referred visual and numerical data in the biomedical database. For the second problem, the adopted NLP and ML are more simple and fast in contrast to Bert-type neural networks. The details of proposed method are demonstrated in the following section.

C. OFFLINE TRAINING PHASE

The whole process in this phase can be regarded as a NLP-based method, including three steps: *n*-gram keywords extraction, determinations of feature keywords and construction of the multi-classification model. In the first step, as shown in Lines 1-12 of Figure 3, one-gram, bi-gram and tri-gram keywords are extracted from the unstructured reports. Next, the features keywords are determined via feature selection measures. Finally, the multi-classification model is conducted by the training data. In the following subsections, the calculations of TF (Term-Frequency), CTF (Class-Term-Frequency), ICF (Inverse-Class-Frequency), determinations of feature keywords and construction of multi-classification model are presented in detail.

1) CALCULATIONS OF TF, CTF AND ICF

In this operation, the main idea is to calculate the term-frequency (TF), class-term-frequency (CTF) and inverse-class-frequency (ICF) for each *n*-gram keyword. In terms of TF, it indicates the term occurrence rate in a report, which can be defined as:

$$TF_{kwd_i}^{rp_x} = \frac{O_{kwd_i}^{rp_x}}{\sum_{j=1}^{k} O_{kwd_j}^{rp_x}}, \tag{1}$$

where k denotes the number of unique keywords in the training data, $TF_{kwd_i}^{rp_x}$ denotes the term frequency of the ith keyword kwd_i in the xth report rp_x, $O_{kwd_i}^{rp_x}$ and $O_{kwd_j}^{rp_x}$ stand for the occurrence counts of the ith and jth keywords in the xth report.

In addition to TF, the referred Class-Term-Frequencies (CTFs) for keywords in each class is computed then. Note that, the class here indicates the diagnosis labels in Figure 1-(b). To calculate the CTF, the training reports are grouped into the classes, as shown in Figure 4. In this paper, CTF can be defined as:

$$CTF_{kwd_i}^{cs_c} = \frac{\sum_{x=1}^{y} O_{kwd_i}^{rp_x}}{\sum_{j=1}^{k}\sum_{x=1}^{y} O_{kwd_j}^{rp_x}}, \tag{2}$$

where $CTF_{kwd_i}^{cs_c}$ indicates the ith keyword frequency in the cth class cs_c, $O_{kwd_i}^{rp_x}$ and $O_{kwd_j}^{rp_x}$ indicate the occurrence counts of the ith and jth keywords, respectively in the xth report, and y indicates the number of reports in the cth class. The intents of TF and CTF are to reveal the representativeness for a keyword. This is because a highly frequent keyword in a report/class represents its high relevance to this report/class. In addition to TF/CTF, another concern is the discrimination of a keyword. In this paper, Inverse Class Frequency (ICF) is

Input: A set of unstructured CT-tumor reports $CR=\cup rp_x$, a set of labels $CS=\{cs_1, cs_2, \ldots, cs_c, \ldots, cs_m\}$; a stop-word set SW, and thresholds nr and rf;
Output: A multi-classification model;
Algorithm *Multi-classification training*
//----------n-gram keywords extraction----------
1. **for** each report rp in CR **do**
2. **for** each sentence in rp **do**
3. extract keywords from the report into the sequential set $K=\{kwd_1, kwd_2, \ldots, kwd_{|K|}\}$;
4. let $K=K\backslash SW$;
5. **for** each sequential $kwd \in K$ **do**
6. let kwd=stem(kwd); // stem(kwd) indicates the function for stemming the keyword kwd
7. **for** n=1 to 3 **do**
8. generate the n-gram terms as the set nt;
9. let $n\text{-}NT=n\text{-}NT\cup nt$; // n-NT indicates the term set including n-gram keywords where n=1, 2 and3
10. **end do**
11. **end do**
12. **end do**
//----------Determinations of feature keywords----------
13. **for** c=1 to $|D|$ **do**
14. **for** n=1 to 3 **do**
15. **for** each n-gram keyword kwd_i in n-NT **do**
16. calculate $TF_{kwd_i}^{rp_x}$, $CTF_{kwd_i}^{cs_c}$, ICF_{kwd_i}, $NR_{kwd_i}^{cs_c}$ and $RF_{kwd_i}^{cs_c}$ based on Equations (1)-(5);
17. **if** $NR_{kwd_i}^{cs_c} \geq nr$ and $RF_{kwd_i}^{cs_c} \geq rf$ **then**
18. let $n\_FK = n\_FK \cup kwd_i$;
19. **end do**
20. select the top-z feature unique keywords into the set n_FK by feature selection measures;
21. **end do**
22. **end do**
//----------Training by feature keywords----------
23. **for** each report rp in CR **do**
24. let SF_{rp} as the feature vector where the attributes are the n_FK;
25. **for** n=1 to 3 **do**
26. **for** each n-gram $kwd_i \in n\_FK$ **do**
27. calculate the sf_{kwd_i} based on Equation (7);
28. input SF_{rp} into the multi-classifier mc;
29. **end do**
30. **end do**
31. **return** mc;

FIGURE 3. Algorithm of training for multi-classification.

used for this concern, which can be defined as:

$$ICF_{kwd_i} = \frac{|D|}{\sum_{cs_c \in D} f_{kwd_i}^{cs_c}}, \tag{3}$$

where ICF_{kwd_i} indicates the inverse class frequency of the ith term, D indicates the training dataset containing a set of unique classes, and

$$f_{kwd_i}^{cs_c} = \begin{cases} 1, & if\ cs_c contains\ kwd_i \\ 0, & otherwise. \end{cases} \tag{4}$$

That is, if the keyword appears in lots of classes, the related distinctness is weak. In summary, the TF and ICF are used to construct the learning model, while the CTF is used to determine the feature keywords.

2) DETERMINATIONS OF FEATURE KEYWORDS

After the previous operation, the CTFs are derived. Based on CTFs, the representative terms for each class are selected. As shown in Lines 13-22 of Figure 3, the top-z class-terms are filtered into the n-gram feature-keyword set named n_FK by the feature selection measure such as CTF, ANOVA F-value, Mutual information and Chi-square. Next, a set of negative terms of each class are also generated by thresholding Negative Rates (NR) and Report Frequencies (RF). That is, if both NR and RF of a keyword exceed the thresholds, respectively, this keyword will be added into the negative-term set n_FK. Finally, sets of negative-terms and class-terms are combined into a feature-keyword set. Here, the Negative Rate and Report Frequency are defined in Definitions 1 and 2.

Definition 1: Given a database, assume the cth class contains a positive report set P and the other reports not in the cth class are regarded as the negative set N. Accordingly, the Negative Rate for the ith keyword kwd_i in the cth class cs_c is defined as:

$$NR_{kwd_i}^{cs_c} = \frac{O_{kwd_i}^{rp_x \in N}}{O_{kwd_i}^{rp_x \in P}}, \tag{5}$$

where rp_x indicates the xth report, $O_{kwd_i}^{rp_x \in N}$ and $O_{kwd_i}^{rp_x \in P}$ indicate the occurrence counts of rp_x in sets N and P, respectively.

Definition 2: By referring to Definition 1, the Report Frequency for the ith keyword kwd_i in the cth class cs_c is defined as:

$$RF_{kwd_i}^{cs_c} = \frac{O_{kwd_i}^{rp_x \in P}}{O_{kwd_i}^{rp_x \in P \cup N}}, \tag{6}$$

where rp_x indicates the xth report, $O_{kwd_i}^{rp_x \in P}$ and $O_{kwd_i}^{rp_x \in P \cup N}$ indicate the occurrence counts of rp_x in sets N and $\{P \cup N\}$, respectively.

After the negative terms are set, the final feature keywords are composed of the feature-selection terms (class terms) and negative terms, as shown in Figure 4.

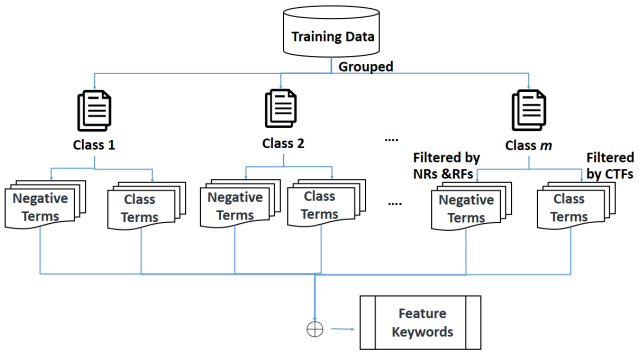


FIGURE 4. Determination of feature keywords.

3) CONSTRUCTION OF THE LEARNING MODEL

In this step, the above feature keywords are viewed as the semantic feature attributes. Accordingly, each training report is transformed into a semantic feature vector which is defined as:

$$SF_{rp_x} = \{sf_1, sf_2, \ldots, sf_z, \ldots, sf_{|n\_FK|}, \tag{7}$$

where n_FK indicates the nth gram feature-keyword set and

$$sf_z = TF_{kwd_z}^{rp_x} * ICF_{kwd_z}. \tag{8}$$

Based on the semantic feature vectors, the learning model is trained, as shown in Lines 23-31 of Figure 3. In this paper, the candidate learning models include Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Random Forest (RF) and Neural Network (NN). In the next section, the comparative evaluation results will be presented.

D. ONLINE STRUCTURING PHASE

As recalled from Figure 2, the online structuring phase is triggered with an unstructured report. Thereupon, three main steps are performed, including n-gram feature keywords extraction, generation of feature vectors and multi-labels classification. Finally, multiple labels will be returned for the regular form. To know the factors in the model, SHAP [21] will be adopted to analyze the feature keywords.

IV. EXPERIMENTS

In the above section, a detailed presentation for the proposed method has been shown. In overall, the proposed method consists of two main stages with several components. Each of them plays a critical role. To catch the role performances, a set of evaluations were made based on four points: 1) effectiveness of feature selections, 2) effectiveness of multi-classifiers, 3) parameter settings and 4) comparisons between proposed and compared methods. Through the empirical analysis, the technical contributions for effectiveness and efficiency will be clear. Note that, all experiments were conducted in python, running on a server with Intel(R) Core(TM) i7-10700K CPU @ 3.80GHz 3.79 GHz and 32GB RAM.

A. EXPERIMENTAL SETTINGS

1) EXPERIMENTAL PROGRAM

By considering the proposed method in Figure 2, the experiment can also be divided into two phases, namely offline training and online testing. In the offline training phase, the experimental data is randomly split into 5 folds. One fold is used for testing and the others are used for training. Next, the feature keywords were determined by selectors and each known report is processed into a feature-keyword vector by NLP operations. Based on the processed training data, 4 candidate multi-classifiers {SVM, LDA, RF, NN} are trained into recognition models. In the testing phase, each testing report is recognized into multiple labels by recognition models. Then, the related evaluation measures are calculated and the comparisons are generated. Finally, the factor keywords are filtered from the training model. In the following, the experimental data and evaluation measures are presented in details.

2) EXPERIMENTAL DATA

The experimental data was gathered from Departments of Diagnostic Radiology, and Surgery, Kaohsiung Chang Gung Memorial Hospital, including 192 CT liver-tumor unstructured reports. To generate the ground-truth, a real annotation system was implemented and the doctors were invited to tag the reports. As shown in Figure 1-(b), the final report expected is a hierarchical structure, consisting of 12 diagnosis labels. (Note that, because no experimental report is labeled as ''Threshold growth'', the number of labels in the experiments is 11.) Hence, this paper is identified as a multi-labeling research. That is, the reports are classified into several labels finally. To make the experiment more solid, 5-cross validations were conducted without using stratified K-folds cross-validator. That is, the experimental data was randomly split into 5 folds, with respect to numbers of 39, 39, 38, 38 and 38, respectively. One fold is used as the testing and the others are used as training.

3) EVALUATION MEASURES

In the experiments, 5 evaluation measures were employed, namely *Accuracy*, *Precision*, *Recall*, *F-measure* and *AUC* (Area Under Curve). These measures are basically inferred by a well-known confusion matrix, which contains four outcomes, True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). The Recall, Precision, Accuracy, F-measure and AUC measures are defined as follows:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{9}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{10}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{11}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{12}$$

TABLE 1. Accuracies of feature selection measures for different multi-classifiers using one-gram feature keywords.

feature measure \ classifier	RF	SVM	LDA	NN
ANOVA	0.901399	0.897644	0.690443	0.901865
MI	0.9	0.898147	0.68064	0.899019
CH	0.893719	0.89762	0.873905	0.880555
CTF	0.907067*	0.900012	0.863514	0.902331

Note that, * denotes the best accuracy.

and AUC stands for the area under the ROC curve, where ROC indicates the receiver operating characteristic. In general machine learning evaluations, 80% is the baseline for AUC. Additionally, the main idea of accuracy is to consider the rates of true positives and true negatives simultaneously in the confusion matrix, while the recall reveals the TP sensitivity against the TP and FN. F-measure is an overall metric which balances the precision and recall. Based on these measures, the evaluation results will be comprehensive and objective. In the following evaluations, all model parameters were approximated by iteratively testing.

B. EFFECTIVENESS OF FEATURE SELECTIONS AND MULTI-CLASSIFIERS

In the proposed method, feature selection is the main component which is sensitive to the classification results. Without good features, the better results are not easy to derive. This issue can further be decomposed of three sub-issues, namely feature selection measures, number of features and n values of grams. As stated in sub-Section 3-C-2, the candidate measures are CTF, ANOVA F-value (termed ANOVA), Mutual information (termed MI) and Chi-square (termed CH), while the candidate multi-classifiers are SVM, LDA, RF and NN, as shown in sub-Section 3-C-3. Table 1 provides the evidence for how to select the feature selection measures and multi-classifiers for the succeeding experiments.

In this table, 4 multi-classifiers were examined by 4 feature selection measures with 32 one-gram feature keywords, which shows the best accuracy is of using RF and CTF. Therefore, RF with CTF features was selected as the main settings of proposed method for the following evaluations. In addition to performances of feature selection measures and multi-classifiers, the next concerns are the impacts of feature quantities and n values of grams.

Table 2 show the one-gram model is better than bi-gram and tri-gram models. Also, 32 features are the best in contrast to the other numbers of features. This is because fewer features cannot reveal the distinctness, while more features contain too many noises. Hence, the 32 n-gram features are the bases that combine different numbers of the other grams features as bi-combinations. Table 3 shows the best accuracy is the setting of combining 32 one-gram and 16 bi-gram features. Then, the tri-combinations of n-gram models are

TABLE 2. Accuracies of single n-gram models for different number of feature keywords.

#feature \ *n*-gram	one	bi	tri
16	0.903263	0.902809	0.887621
32	0.907067*	0.903742	0.89378
64	0.904245	0.901386	0.894283
128	0.901448	0.90568	0.897129

Note that, * denotes the best accuracy.

TABLE 3. Accuracies of bi-combinations of n-gram models for different number of feature keywords.

#features \ *n*-grams	one+bi	one+tri	bi+tri
32+16	0.909901*	0.907987	0.898013
32+32	0.907103	0.90276	0.901374
32+64	0.904245	0.901411	0.899975
32+128	0.902392	0.90092	0.90092

Note that, * denotes the best accuracy.

TABLE 4. Accuracies of tri-combinations of n-gram models for different number of feature keywords.

#features \ *n*-grams	one+bi+tri
32+32+32	0.90611*
32+32+128	0.904233
32+64+128	0.903766
32+128+128	0.90373

Note that, * denotes the best accuracy.

tested as shown in Table 4. Whatever the tri-combination is, the related accuracy cannot be improved. In summary, in this experimental result, the best accuracy is of using RF multi-classifier with 32 one-gram and 16 bi-gram features filtered by CTF measures. The potential reason is that, the more the features, the more the noises, the lower the accuracy. Also, one-gram model performs better than the other gram models.

C. COMPARISONS BETWEEN THE PROPOSED METHOD AND BERT-BASED NEURAL NETWORKS

After the evaluations for approximating the best settings of proposed method, the next issue to clarify is how effective the proposed method is in comparison with recent popular Bert-based neural networks. In this evaluation, four Bert-based neural networks were compared with the proposed method, namely Tiny Bert (termed TBert), Small Bert (termed SBert), Bio Bert (termed BBert) and Clinical Bert (termed CBert). Table 5 shows the confusion matrixes of

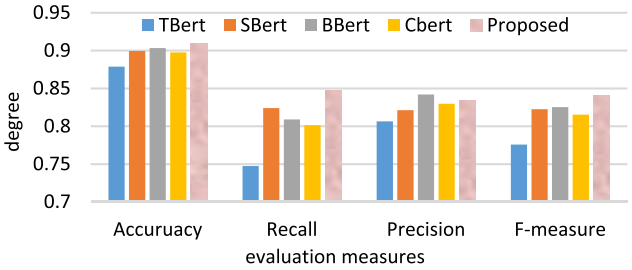


FIGURE 5. Performances of compared methods in terms of accuracy, recall, precision and F-measure.

compared methods. From this matrix, the averaged standard deviations of cross-validations for TBert, SBert, BBert, CBert, Proposed are calculated as 6.829, 7.895, 6.23, 7.963, 6.224, respectively, which indicate the proposed method performs stably in contrast to the compared methods. Based on Table 5, Figure 5 further shows the summarized performances of compared methods in terms of accuracy, recall, precision and F-measure. The results in Figure 5 can be summarized into several points. First, for the concern of imbalanced data, the accuracies of compared methods are close, where the TBert is slightly worse and the proposed method is slightly better. Second, for the positive prediction value, the precision differences of all methods are small. Third, from sensitivity point of view, the recall of the proposed method is better than those of the others, in contrast to the precision. This is because the mined feature keywords are more sensitive for predicting the negatives than predicting the positives. That is, for the proposed method, the false negative is lower than those of the compared methods, in contrast to the false positive. Fourth, for balancing of precisions and recalls, the proposed method is better than the compared methods in terms of F-measure. In summary, the best multi-classifier is the proposed method while considering accuracy and F-measure. Further, the additional evidence for this summary is shown in Figure 6 depicting the AUCs of compared methods. It says that, first, all compared methods achieve around 0.93 of AUCs exceeding the baseline 0.8. Second, the proposed method is slightly better than the compared methods. In conclusion, the results of using measures accuracy, F-measure and AUC show that, the proposed method integrating NLP and ML can bring out a satisfactory structure.

D. EFFICIENCY EVALUATIONS OF THE COMPARED METHODS

In fact, the effectiveness differences of all compared methods are not significant although the proposed method performs better than the compared method. To further clarify the proposed contribution, in this sub-section, an efficiency evaluation with respect to training time is demonstrated. Figure 7 shows the proposed method is much more efficient than the compared methods in terms of training time. This result can be viewed as an echo of contributions mentioned in Section 1,

TABLE 5. Confusion matrixes for compared methods.

	TBert				SBert				BBert				CBert				Proposed			
	TP	FP	FN	TN	TP	FP	FN	TN	TP	FP	FN	TN	TP	FP	FN	TN	TP	FP	FN	TN
Fold 1	91	18	35	285	104	16	22	287	101	16	25	287	101	18	25	285	110	17	16	286
Fold 2	75	23	31	300	87	19	19	304	85	13	21	310	86	13	20	310	92	19	14	304
Fold 3	88	15	32	283	103	22	17	276	98	21	22	277	106	24	14	274	103	23	17	275
Fold 4	99	29	24	266	104	34	19	261	100	25	23	270	88	22	35	273	103	25	20	270
Fold 5	90	22	27	279	90	17	27	284	95	16	22	285	93	21	24	280	94	16	23	285

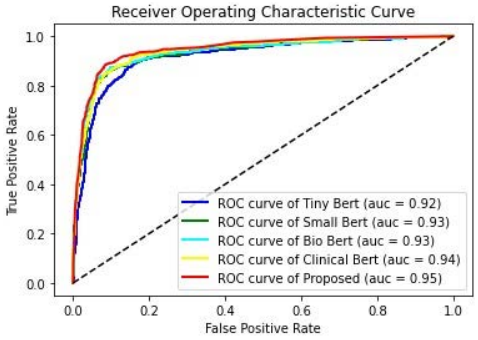


FIGURE 6. AUCs of compared methods.

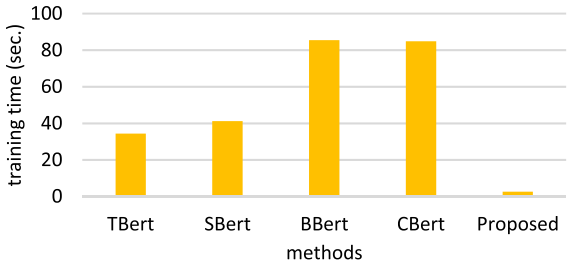


FIGURE 7. Training time of compared methods.

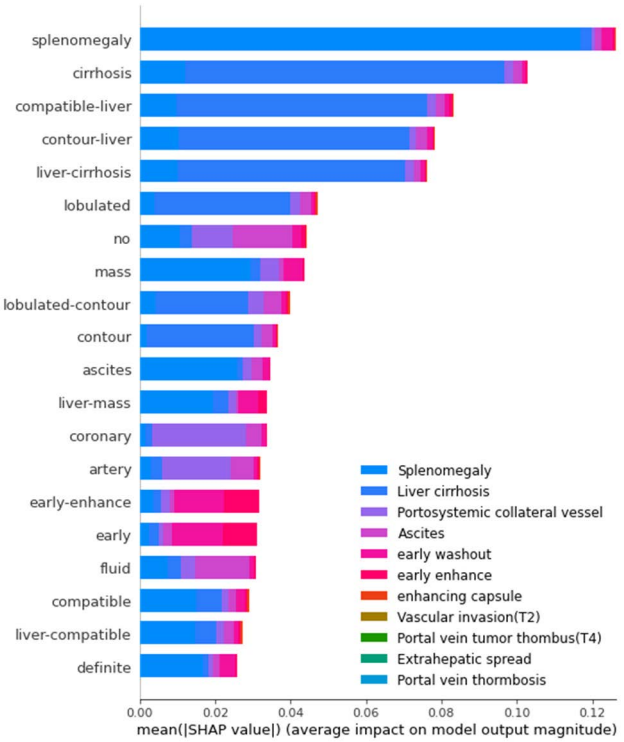


FIGURE 8. Impacts of top 20 keywords for all labels.

which indicates the proposed method is simple and light for the report multi-classifications.

E. EMPIRICAL DISCUSSIONS

In above, the evaluations are presented in a logistic form that, first, how to determine the multi-classifiers with n-gram feature keywords is shown. Next, the evaluations for effectiveness and efficiency are demonstrated. However, some critical issues need to be clarified further. In this sub-section, an insightful discussion for experimental results are listed as follows.

- In the experimental results, the proposed method is evaluated from effectiveness and efficiency points of view, in comparison with 4 Bert-based neural networks. Although the effectiveness is close, the training time improvement, on the contrary, is very obvious. In detail, for effectiveness, Figures 6 and 7 illustrate that, on average, the improvements of the accuracies, F-measures and AUCs are 1.7%, 3.9% and 2.2%, respectively. However, for efficiency, as shown in Figure 7, the proposed method just needs 2.66 seconds to train the learning model in contrast to 61.45 seconds for compared methods on average. This indicates the improvement of training time can reach 2210.1%, which makes the proposed contribution clear. This contribution is important especially for dealing with big data in the field of biomedical AI.
- In traditional ML methods, the feature-selection is a crucial component for the classification quality. In this work, the best model is derived by the CTF-based selection. However, there remains a question: what if using the features determined by experts (doctors)? To answer

TABLE 6. Accuracies of different models using expert-defined features.

n-grams \ classifier	RF	SVM	LDA	NN
one	0.907999	0.899571	0.711925	0.899988
bi	0.904674	0.898123	0.65181	0.901865
tri	0.895215	0.891976	0.773255	0.89243
one+bi	0.908036*	0.900012	0.797816	0.904282
one+tri	0.905693	0.900969	0.705742	0.900945
bi+tri	0.900454	0.896234	0.610477	0.906613
one+bi+tri	0.902822	0.899558	0.822733	0.906134

Note that, * denotes the best accuracy.

this question, all multi-classifiers based on different n-gram models were examined. Table 6 illustrates that, the best accuracy is around 0.908 which is almost the same as that of proposed method. An important point to show here is that, the proposed feature-selection performs as well as the expert-defined. That is, without high-priced manual cost, the proposed method works well also.

- In addition to effectiveness and efficiency, another contribution to clarify here is the interpretability. This idea is motivated by the problem that, it is not easy to catch whys of the learning models work well in traditional. To aim at this issue, SHAP was performed after testing in the experiments. Figure 8 shows the impacts of top 20 keywords for all labels in the RF model, including 13 one-gram keywords and 7 bi-gram keywords. The top-1 and top-2 keywords are ''splenomegaly'' and ''compatible liver'', respectively, which deliver two keywords-to-label relations with respect to ''splenomegaly-to-Splenomegaly'' and ''compatible liver-to-Liver cirrhosis''. This is an understandable result which indicates the factors for recognizing the labels. The other relations can be discovered, and the results will be the bases for approximating a better model in future incremental training.
- Following to above issue, another concern is: is there any other compared classifiers in addition to Bert? For this concern, the past study [7] can be the potential solution. In this study, the primary intent is to classify the sentences by fusing NLP and CNN, which is somewhat different from that of proposed method. However, to clarify the concern, we extended this study as an additional compared method to classify the report (document) instead of classifying sentences. In this extension, the report was transformed into a parts-of-speech to feature-keywords matrix, where the feature-keywords with TFs were selected by doctors. Next, the 2-Dimension CNN was performed to recognize the report as a set of potential labels. Table 7 is the evaluation result in comparison with that of the proposed method. The results deliver an aspect that, although the POS and CNN are fused by the manual feature-keywords, the performances for all metrics are not better than those of proposed method. The potential interpretations are: first, the POS is not sensitive in the 2D CNN for multi-classifying the report. Second, the CNN are too complicated for the 2D-POS matrix.

TABLE 7. Comparisons with the extended reference [7].

	Accuracy	Recall	Precision	F-measure	AUC
Reference [7]	0.89	0.788	0.814	0.801	0.93
Proposed	0.91	0.848	0.834	0.841	0.95

TABLE 8. Accuracies of NN models by different settings.

n-grams \ parameter	epochs	batch size	learning rate	#nodes in layer1	#nodes in layer2	accuracy
one	200	48	0.002	256		0.90233
one+bi	250	64	0.001	512		0.90467
one+bi+tri	250	64	0.001	512	512	0.90519*

Note that, * denotes the best accuracy.

- The other concern to clarify is the performance of NN. In this paper, another core role is the multi-classifier besides feature selection. Actually, NN is a popular solution in the field of Artificial Intelligence. However, the related performance in this paper is not outstanding in contrast to the other candidate MLs shown in Table 1. It elicited a further interest: what is the best performance based on the optimal settings for NN? To aim at this interest, a number of settings were investigated to approximate the nearly optimal result, which indicates 0.90519, as summarized in Table 8. In particular, the best result is generated by using one-, bi- and tri-gram models with 2 layers. Insightfully, because the main advantage of NN is to deal with more complicated data, the best result is derived by more complicated data structures, more epochs, more neurons and deeper networks in this evaluation.
- In summary, the performance of proposed method relies on two core components, namely feature keyword determination and multi-classifier. In terms of feature keyword determination, the feature selection measure and n-gram model play critical roles. For this concern, Tables 2-4 reveal the best settings. In terms of multi-classifier, the impact of the classifier is shown in Tables 1 and 5. Whatever the n-gram model is, LDA performs much worse than the other testing classifiers. On one hand, the classifier is still a considerable factor. On the other hand, based on the robust feature keywords selected, the differences of reliable classifiers are not significant.

F. CASE STUDY

To make the proposed method easier to understand, an illustrative case is lifted here based on Figure 1. First, in the

TABLE 9. Example of multi-labeling results for figure 1.

	Prediction Result	Ground Truth
Early arterial enhancement	1	1
Early washout	1	0
Enhancing capsule	0	0
Vascular invasion(T2)	0	0
Portal vein tumor thrombus(T4)	0	0
Extrahepatic spread	0	0
Splenomegaly	1	1
Liver cirrhosis	1	1
Ascites	0	0
Portosystemic collateral vessel	0	0
Portal vein thrombosis	0	1

TABLE 10. Example of the confusion matrix and evaluation results for Table 9.

	Prediction Result
Tue Positive	3
False Positive	1
False Negative	1
True Negative	6
Accuracy	0.818
Precision	0.75
Recall	0.75
F-measure	0.75

first stage, there are overall 87 one-gram and 72 bi-gram feature keywords are determined from 153 training reports offline, which is called *n_FK* in Figure 3. Based on these feature keywords, the RF model is trained in the second stage offline. Figure 8 shows the top 20 keywords in the training model. For the testing example shown in Figure 1-(a), 65 one-gram and 73 bi-gram feature keywords with TFs are extracted in the first stage online. For example, in this case, ones of extracted one-gram and bi-gram feature keywords are ''liver'' and ''favor-hcc'', respectively, where the referred TFs are 0.069 and 0.028, respectively. The other feature keywords not existing in this example is with zero TFs. These feature keywords can be referred to the vector SF_{rp} in Figure 3. After this stage, the first-stage structuring has been completed, and the report is therefore represented by these feature keywords. In other words, these feature keywords are viewed as the patterns with different representativeness. With this feature keyword vector, the report is multi-classified into a set of labels. Table 9 shows the prediction results and ground truths, and thereupon the confusion matrix and evaluation results are generated as shown in Table 10.

V. INSIGHTFUL DISCUSSIONS FOR CONTRIBUTIONS

After presenting the proposed method and the evaluation results, an insightful discussion for the contribution is lifted here. As we can recall from above, there exist a massive amount of knowledge in the past unstructured CT liver image reports. However, it is not easy to discover the knowledge from these unstructured reports. Therefore, the major intent of this paper is to propose an effective and efficient method to transform the unstructured reports into structured ones. On the whole, the main contributions can be summarized into 4 points. First, the proposed method is readable because the first stage extracts the feature keywords and the impact keywords for the classification model are further discovered by SHAP. Actually, the impact keywords can be used for optimizing the recognition model. Second, once the feature keywords are extracted, the potential complications can be bridged to the liver symptoms in the future. Third, because the automated CT tumor recognition needs a huge amount of training image data, it needs autonomous tagging. Hence, if the report can be organized into a readable structure, the autonomous visual tagging can be achieved. According to the structured report, the ground truth of training tumors can be generated automatically. For example, the tumor location is noted in the report. Based on the location, the tumor can be segmented automatically. Therefore, the ground truth can be derived. This is why we propose this paper. Further, it is easier to conduct the visual recognition report by the regular format instead of linguistic sentences. Fourth, if the image can be recognized, the candidate treatments can be recommended.

VI. RESEARCH LIMITATIONS

In this paper, we have provided the method details and evaluation analysis. Yet, there remain some limitations needing to be declared. First, the parameters in the proposed method were approximated for the experimental data. Second, the final structure was defined by the radiologists in Kaohsiung Chang Gung Memorial Hospital, Taiwan. Third, the ensemble learning, meta learning and federated learning were not used in this research. Fourth, the problem of data imbalance was not solved. Fifth, the stratified K-folds cross-validator was not adopted in the experiments. More discussions for future works will be listed in the Section 7.

VII. CONCLUSION AND FUTURE WORKS

In principle, knowledge discovery refers to a set of mechanisms retrieving the valuable patterns from massive data through effective data engineering. Good knowledge is very helpful to further prediction, retrieval and recommendation. Up to the present, there have been lots of researches approved to be effective on knowledge discovery in the field of biomedical science. Nevertheless, few previous literatures focused their attention on structuring CT liver-tumor reports. It leads a high manual cost to organize a readable report. To address this issue, in this paper, a two-stage structuring method is proposed from effectiveness, efficiency and interpretability

points of view. In the first stage, a creative feature selection, named ''features filtered by class-term-frequency'', is proposed for determining the feature keywords. Therefore, the unstructured reports are regulated in a useful structure formatted by these feature keywords. Based on these keywords, in the second stage, an effective multi-classifier is performed to structure the reports as an advanced form. To present the effectiveness, efficiency and interpretability of proposed method, a number of robust evaluations were conducted on a real dataset. The experimental results show that, the proposed method is slightly more effective but much more effective than the modern Bert-based neural networks. Moreover, an insightful analysis and discussion are lifted to make the contribution clearer.

Although the goal of this paper is to structure the linguistic reports, it can be recognized as a multi-labeling problem. After structuring, the linguistic reports are transformed as two-level structures. The final format is actually a biomedical ontology defined by doctors. From the results, the patterns in the 1st stage can be viewed as the features to calculate the similarity between labels in the 2nd stages in the future. Also, the referred CT images can be clustered by these patterns and labels. And so on, this paper can be viewed as a data engineering study providing an extensible result for future explorations of knowledge. That is, this research is just a beginning for biomedical information retrieval. In the future, a number of investigations will be carried further. First, from technical point of view, the ensemble learning will be tested for a better prediction. Further, the meta learning will be adopted for problem of data imbalance. Second, from the extension point of view, the patterns will be used to mine the associations among numerical, textual and visual data for further disease risk assessment, disease recognition and treatment recommendation. Third, from practical point of view, it will be implemented into the existing bioinformatic system. Fourth, from application point of view, the proposed idea will be applied to the other clinic reports.

ACKNOWLEDGMENTS

The experimental data was approved by Kaohsiung Chang Gung Memorial Hospital, Taiwan, and all operations in this paper were executed according to the ethical standards of the Institutional Review Board, Taiwan.

REFERENCES

[1] O. A. Abass and O. A. Arowolo, ''Information retrieval models, techniques and applications,'' *Int. Res. J. Adv. Eng. Sci.*, vol. 2, no. 2, pp. 197–202, 2017.
[2] A. Amalia, O. S. Sitompul, E. B. Nababan, and T. Mantoro, ''An efficient text classification using fastText for Bahasa Indonesia documents classification,'' in *Proc. Int. Conf. Data Sci., Artif. Intell., Bus. Anal. (DATABIA)*, Jul. 2020, pp. 69–75.
[3] A. K. Allada, Y. Wang, V. Jindal, M. Babee, H. R. Tizhoosh, and M. Crowley, ''Analysis of language embeddings for classification of unstructured pathology reports,'' in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 2378–2381.
[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, ''Enriching word vectors with subword information,'' *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.
[5] Z. Chen, ''Short text classification based on word2vec and improved TDFIDF merge weighting,'' in *Proc. 3rd Int. Conf. Electron. Inf. Technol. Comput. Eng. (EITCE)*, Oct. 2019, pp. 1719–1722.
[6] Y. Chouni, M. Erritali, Y. Madani, and H. Ezzikouri, ''Information retrieval system based semantique and big data,'' *Proc. Comput. Sci.*, vol. 151, pp. 1108–1113, Jan. 2019.
[7] S. Chotirat and P. Meesad, ''Part-of-speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning,'' *Heliyon*, vol. 7, no. 10, Oct. 2021, Art. no. e08216.
[8] Y. Chen, X. Zhang, and T. Li, ''Medical records classification model based on text-image dual-mode fusion,'' in *Proc. 4th Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2021, pp. 432–436.
[9] S. Dorn, S. Chen, S. Sawall, J. Maier, M. Knaup, M. Uhrig, H.-P. Schlemmer, A. Maier, M. Lell, and M. Kachelrieß, ''Towards context-sensitive CT imaging–organ-specific image formation for single (SECT) and dual energy computed tomography (DECT),'' *Med. Phys.*, vol. 45, no. 10, pp. 4541–4557, Oct. 2018.
[10] S. Gao, M. Alawad, M. T. Young, J. Gounley, N. Schaefferkoetter, H. J. Yoon, X.-C. Wu, E. B. Durbin, J. Doherty, A. Stroup, L. Coyle, and G. Tourassi, ''Limitations of transformers on clinical text classification,'' *IEEE J. Biomed. Health Informat.*, vol. 25, no. 9, pp. 3596–3607, Sep. 2021.
[11] T. Jansen, G. Geleijnse, M. Van Maaren, M. P. Hendriks, A. T. Teije, and A. Moncada-Torres, ''Machine learning explainability in breast cancer survival,'' *Stud Health Technol Inf.*, vol. 270, pp. 307–311, Jun. 2020.
[12] T. Huynh, Y. Gao, J. Kang, L. Wang, P. Zhang, J. Lian, and D. Shen, ''Estimating CT image from MRI data using structured random forest and auto-context model,'' *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 174–183, Jan. 2016.
[13] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, ''Comparing automated text classification methods,'' *Int. J. Res. Marketing*, vol. 36, no. 1, pp. 20–38, 2019.
[14] E. H. Houssein, R. E. Mohamed, and A. A. Ali, ''Machine learning techniques for biomedical natural language processing: A comprehensive review,'' *IEEE Access*, vol. 9, pp. 140628–140653, 2021.
[15] R. Islam, A. V. Andreev, N. N. Shusharina, and A. E. Hramov, ''Explainable machine learning methods for classification of brain states during visual perception,'' *Mathematics*, vol. 10, no. 15, p. 2819, Aug. 2022.
[16] S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir, ''A review on recent research in information retrieval,'' *Proc. Comput. Sci.*, vol. 201, pp. 777–782, Jan. 2022.
[17] H. Jin, A. Sun, R. Zheng, R. He, Q. Zhang, Y. Shi, and W. Yang, ''Content and semantic context based image retrieval for medical image grid,'' in *Proc. 8th IEEE/ACM Int. Conf. Grid Comput.*, Sep. 2007, pp. 105–112.
[18] M. Jamaluddin and A. D. Wibawa, ''Patient diagnosis classification based on electronic medical record using text mining and support vector machine,'' in *Proc. Int. Seminar Appl. Technol. Inf. Commun. (iSemantic)*, Sep. 2021, pp. 243–248.
[19] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, ''Text classification algorithms: A survey,'' *Information*, vol. 10, no. 4, p. 150, Apr. 2019.
[20] S. F. N. B. S. Kamaruddin, F. Mohd, M. P. Hamzah, F. Harun, N. R. Zainol, and N. I. M. Daud, ''Information retrieval for Malay text: A decade review of research (2008–2019),'' in *Proc. 5th Int. Conf. Inf. Retr. Knowl. Manage. (CAMP)*, Jun. 2021, pp. 2–7.
[21] S. M. Lundberg and S.-I. Lee, ''A unified approach to interpreting model predictions,'' in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 4765–4774.
[22] W. Li, S. Liao, Q. Feng, W. Chen, and D. Shen, ''Learning image context for segmentation of the prostate in CT-guided radiotherapy,'' *Phys. Med. Biol.*, vol. 57, no. 5, pp. 1283–1308, Feb. 2012.
[23] S. Mohan, N. Fiorini, S. Kim, and Z. Lu, ''A fast deep learning model for textual relevance in biomedical information retrieval,'' in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 77–86.
[24] L. Ma, X. Liu, Y. Gao, Y. Zhao, X. Zhao, and C. Zhou, ''A new method of content based medical image retrieval and its applications to CT imaging sign retrieval,'' *J. Biomed. Informat.*, vol. 66, pp. 148–158, Feb. 2017.
[25] L. Ma and Y. Zhang, ''Using Word2 Vec to process big text data,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 2895–2897.
[26] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, ''Medical image synthesis with context-aware generative adversarial networks,'' in *Medical Image Computing and Computer Assisted Intervention*, vol. 10435. U.S. National Institutes of Health's National Library of Medicine (NIH/NLM), Sep. 2017, pp. 417–425. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6044459/, doi: 10.1007/978-3-319-66179-7_48.

[27] E. Nguyen, D. Theodorakopoulos, S. Pathak, J. Geerdink, O. Vijlbrief, M. van Keulen, and C. Seifert, "A hybrid text classification and language generation model for automated summarization of Dutch breast cancer radiology reports," in *Proc. IEEE 2nd Int. Conf. Cognit. Mach. Intell. (CogMI)*, Oct. 2020, pp. 72–81.

[28] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604–624, Feb. 2021.

[29] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," in *Proc. 18th BioNLP Workshop Shared Task*, 2019, pp. 58–65.

[30] J. X. Qiu, H.-J. Yoon, P. A. Fearn, and G. D. Tourassi, "Deep learning for automated extraction of primary sites from cancer pathology reports," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 244–251, Jan. 2018.

[31] A. Safaei, "Text-based multi-dimensional medical images retrieval according to the features-usage correlation," *Med. Biol. Eng. Comput.*, vol. 59, no. 10, pp. 1993–2017, Aug. 2021.

[32] A. Spandorfer, C. Branch, P. Sharma, P. Sahbaee, U. J. Schoepf, J. G. Ravenel, and J. W. Nance, "Deep learning to convert unstructured CT pulmonary angiography reports into structured reports," *Eur. Radiol. Experim.*, vol. 3, no. 1, pp. 1–8, Sep. 2019.

[33] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: A review," *J. Healthcare Eng.*, vol. 2018, pp. 1–9, Apr. 2018.

[34] E. Strumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," *J. Mach. Learn. Res.*, vol. 11, pp. 1–18, Jan. 2010.

[35] S. Sivakumar, L. S. Videla, T. R. Kumar, J. Nagaraj, S. Itnal, and D. Haritha, "Review on Word2 Vec word embedding neural net," in *Proc. Int. Conf. Smart Electron. Commun. (ICOSEC)*, Sep. 2020, pp. 282–290.

[36] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," 2020, *arXiv:2003.01200*.

[37] Y. Wang, L. Cui, and Y. Zhang, "Improving skip-gram embeddings using BERT," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1318–1328, 2021.

[38] W. Wang, G. He, and X. Liu, "Text multi-classification based on word embedding and multi-grained cascade forest," in *Proc. IEEE 5th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2019, pp. 13–17.

[39] J. Wu, K. Tang, H. Zhang, C. Wang, and C. Li, "Structured information extraction of pathology reports with attention-based graph convolutional network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 2395–2402.

[40] Z. Wang, S. Xu, and L. Zhu, "Semantic relation extraction aware of N-gram features from unstructured biomedical text," *J. Biomed. Inform.*, vol. 86, pp. 59–70, Oct. 2018.

[41] M. Xue, "A text retrieval algorithm based on the hybrid LDA and Word2 Vec model," in *Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS)*, Jan. 2019, pp. 373–376.

[42] B. Xu, H. Lin, and Y. Lin, "Learning to refine expansion terms for biomedical information retrieval using semantic resources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 954–966, May 2019.

[43] K. Yalcin, I. Cicekli, and G. Ercan, "An external plagiarism detection system based on part-of-speech (POS) tag n-grams and word embedding," *Expert Syst. Appl.*, vol. 197, Jul. 2022, Art. no. 116677.

[44] T. Yao, Z. Zhai, and B. Gao, "Text classification model based on fast text," in *Proc. IEEE Int. Conf. Artif. Intell. Inf. Syst. (ICAIIS)*, Dalian, China, Mar. 2020, pp. 154–157.

[45] Z. Zhang and L. Jin, "Clinical short text classification method based on Albert and GAT," in *Proc. 7th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2022, pp. 401–404.

[46] Y. Zheng, D. Liu, B. Georgescu, D. Xu, and D. Comaniciu, "Deep learning based automatic segmentation of pathological kidney in CT: Local versus global image context," in *Deep Learning and Convolutional Neural Networks for Medical Image Computing* (Advances in Computer Vision and Pattern Recognition), L. Lu, Y. Zheng, G. Carneiro, and L. Yang, Eds. Cham, Switzerland: Springer, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-42999-1_14#citeas, doi: 10.1007/978-3-319-42999-1_14.

[47] Y. Zhang and Z. Rao, "N-BiLSTM: BiLSTM with n-gram features for text classification," in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Jun. 2020, pp. 1056–1059.

[48] B. Zhou, D. Su, and Z. Qu, "Medical text classification system based on deep learning," in *Proc. Int. Conf. Intell. Comput., Autom. Appl. (ICAA)*, Jun. 2021, pp. 388–392.

YI-HSUAN CHUANG received the M.D. degree from China Medical University and the Radiology Residency degree from Kaohsiung Chang Kung Memorial Hospital, with experience in diagnostic imaging including sonography, MRI, and CT scan, where she has been a Diagnostic Radiologist since December 2018. She specializes in liver and breast imaging. She is committed to achieving an improved technique for cancer detection and diagnosis.

JA-HWUNG SU received the Ph.D. degree from the Department of Computer Science and Information Engineering, National Cheng Kung University, in 2010. He is currently an Assistant Professor at the Department of Computer Science and Information Engineering, National University of Kaohsiung. He has held 17 patents in USA and Republic of China and published more than 80 research papers in some premier journals and international conferences, such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE INTELLIGENT SYSTEMS, and ICME. His research interests include machine learning and multimedia information retrieval. He served as a Program Committees and a Reviewers in these journals and international conferences, such as SIGKDD, ICDM, CIKM, DASFAA, TKDD, KNOSYS, CIM, and JBHI.

DING-HONG HAN received the B.S. degree from the Department of Computer Science and Information Engineering, National University of Kaohsiung, in 2021. He is currently pursuing the M.S. degree with the Department of Computer Science and Engineering, National Sun Yat-sen University. His research interests include biomedical data mining and information retrieval.

YI-WEN LIAO received the Ph.D. degree in information management from the National Sun Yat-sen University, Taiwan, in 2010. She is an Associate Professor at the Department of Intelligent Commerce, National Kaohsiung University of Science and Technology, Taiwan. She has held 16 patents in Republic of China and published more than 60 research papers in some premier journals and international conferences, such as *International Journal of Information Management*, *Internet Research*, *Computers in Human Behavior*, *Journal of Educational Computing Research*, and *Journal of Internet Technology*. Her research interests include technology education, programming education, data mining application, artificial intelligence in education, and the Internet of Things application. She also served as an important director and a supervisor at several important domestic societies and devoted herself to information volunteering for eight years.

YEONG-CHYI LEE received the B.S. degree in information management in 1999, and the M.S. and Ph.D. degrees in computer science and information engineering from I-Shou University, in 2001 and 2007, respectively. He is an Assistant Professor at the Department of Information Management, Cheng-Shiu University. He is also the Vice Director of Office of Library and Information, Cheng-Shiu University, in 2021. His research interests include data mining, genetic algorithms, and fuzzy theory.

YU-FAN CHENG graduated from China Medical University. He completed the Radiology Training at Chang Gung Memorial Hospital, Taiwan, Mallinckrodt Institute of Radiology, Washington, and University School of Medicine, St. Louis, USA. He is a Professor at the Department of Radiology and the Liver Transplantation Center, Kaohsiung Chang Gung Memorial Hospital. He is a pioneer in liver cancer treatment, liver transplantation image, and interventional radiology. He has been supported by the National Research Grants for more than 20 years. After becoming a Professor of Radiology in 2003, he was subsequently bestowed with the title of an Honorary Professor by several international universities. He has published more than 400 scientific articles and has lectured in more than 100 international congresses. He is a well-known scientific physician. He is currently one of the global leading experts in interventional radiology and liver transplant imaging.

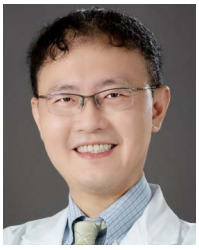
TZUNG-PEI HONG (Senior Member, IEEE) received the B.S. degree in chemical engineering from the National Taiwan University, in 1985, and the Ph.D. degree in computer science and information engineering from the National Chiao-Tung University, in 1992.

He served at the Department of Computer Science, Chung-Hua Polytechnic Institute, from 1992 to 1994; and the Department of Information Management, I-Shou University, from 1994 to 2001. He was in charge of the whole computerization and library planning for the National University of Kaohsiung in Preparation, from 1997 to 2000. He served as the first Director of the Library and Computer Center, National University of Kaohsiung, from 2000 to 2001, the Dean of Academic Affairs, from 2003 to 2006, the Administrative Vice President, from 2007 to 2008, and the Academic Vice President in 2010. He is currently a Distinguished Professor and the Chair Professor at the Department of Computer Science and Information Engineering, the Department of Electrical Engineering, and the Director of AI Research Center, National University of Kaohsiung, Taiwan. He is also a Joint Professor at the Department of Computer Science and Engineering, National Sun Yat-sen University, Taiwan. He has published more than 600 research papers in international/national journals and conferences and has planned more than 50 information systems. He is also the board member of more than 40 journals and the program committee member of more than 1000 conferences. His current research interests include knowledge engineering, data mining, soft computing, management information systems, and www applications. He got the first National Flexible Wage Award from the Ministry of Education in Taiwan.

KATHERINE SHU-MIN LI (Senior Member, IEEE) received the B.S. degree in computer science, Rutgers University, New Brunswick, NJ, USA, and the M.S. degree in computer science and the Ph.D. degree in electrical engineering from the National Chiao-Tung University, Taiwan, in 2001 and 2006, respectively.

She is currently a Full Professor with the Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan. She was a Visiting Fellow with Electrical and Computer Engineering, Princeton University, from September 2021 to 2022. She has been a Visiting Research Collaborator at the Department of Electrical and Computer Engineering, Princeton University, since January 2022. Her current research interests include AI's multiple applications, including software/hardware security (AIS), AI in routing (AIR), AI in microfluidic placement and routing with hardware security (AIMPR&HS), AI and automation in financial engineering (AIAFE), AI in weather/climate prediction (AIWC), AI in medical imaging (AIM), AI chip design and test (AIDT), wafer bin map (WBM) pattern recognition by machine learning techniques (AIW) in addition to 3DIC power/thermal modeling and test automation (3DIC PM/TM and TA), design for yield (DfY), foundry automation (FA), interposer test (IT), 2.5D/3D/SiP IC test, microfluidic chip synthesis and test/diagnosis/fault tolerance, hardware security (hardware trojan, physical unclonable function, and side channel attack), design for security (DfS), oscillation ring test schemes (ORT), and 5G/RF Test.

Dr. Li is a member of IEEE Education and IEEE Circuits and Systems Society, Association for Computing Machinery (ACM), ACM Special Interest Group on Design Automation, and IEEE Women in Engineering (WIE), since October.



HSIN-YOU OU received the M.D. degree from the School of Medicine, Kaohsiung Medical University. He has completed the Postdoctoral Fellow and a Radiology Training at the Magnetic Resonance Research, Johns Hopkins University, USA. He is an Associate Professor at the Department of Radiology, Kaohsiung Chang Gung Memorial Hospital. He has published more than 75 medical articles, 71 of which in SCI journal. His research interests include radiology, hepatobiliary images, angiography, and MRI. He received the Best Annual Article Award of The Radiological Society of the Republic of China, the Outstanding Award of Scientific Article, and the National Biotechnology Medical Quality Award.

YI LU received the bachelor's degree from the Department of Physical Therapy, Kaohsiung Medical University, in 2009. She is a Research Assistant at the Department of Diagnostic Radiology, Kaohsiung Chang Gung Memorial Hospital.

CHIH-CHI WANG graduated from China Medical University. He completed the Surgical Residency Training at the Chang Gung Memorial Hospital, Taiwan. He also completed the Postdoctoral Fellowship at the Harbor UCLA, Liver Support Unit, Department of Surgery, Cedars-Sinai Medical Center, UCLA School of Medicine, Los Angeles. He is a Superintendent at the Kaohsiung Chang Gung Memorial Hospital and a Professor at the Liver Transplantation Center and the Department of Surgery, Kaohsiung Chang Gung Memorial Hospital. His research interests include liver transplantation, bioartificial liver, liver failure, gastrointestinal surgery, endoscopic surgery, and general surgery.