2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385827

Improving VTE Identification through Adaptive NLP Model Selection and Clinical Expert Rule-based Classifier from Radiology Reports

Jamie Deng†, Yusen Wu†,*, Hilary Hayssen‡, Brain Englum‡, Aman Kankaria‡, Minerva Mayorga-Carlin‡, Shalini Sahoo‡, John Sorkin‡, Brajesh Lal‡, Yelena Yesha†, Phuong Nguyen†
†*Frost Institute for Data Science & Computing*, University of Miami, FL, USA
‡*School of Medicine*, University of Maryland, MD, USA
{jxd3987, yxw1259, yxy806, pnx208}@miami.edu
{amankankaria}@gmail.com {HHayssen, BEnglum, MCarlin, Shalini.Sahoo, jsorkin, BLal}@som.umaryland.edu

***Abstract*—Rapid and accurate identification of Venous thromboembolism (VTE), a severe cardiovascular condition including deep vein thrombosis (DVT) and pulmonary embolism (PE), is important for effective treatment. Leveraging Natural Language Processing (NLP) on radiology reports, automated methods have shown promising advancements in identifying VTE events from retrospective data cohorts or aiding clinical experts in identifying VTE events from radiology reports. However, effectively training Deep Learning (DL) and the NLP models is challenging due to limited labeled medical text data, the complexity and heterogeneity of radiology reports, and data imbalance. This study proposes novel method combinations of DL methods, along with data augmentation, adaptive pre-trained NLP model selection, and a clinical expert NLP rule-based classifier, to improve the accuracy of VTE identification in unstructured (free-text) radiology reports. Our experimental results demonstrate the model's efficacy, achieving an impressive 97% accuracy and 97% F1 score in predicting DVT, and an outstanding 98.3% accuracy and 98.4% F1 score in predicting PE. These findings emphasize the model's robustness and its potential to significantly contribute to VTE research.**

***Index Terms*—VTE, NLP, Deep Learning, Transfer Learning, BERT, Bi-LSTM**

I. Introduction

Venous thromboembolism (VTE) [1], including deep vein thrombosis (DVT) and pulmonary embolism (PE), is recognized as the third most prevalent cardiovascular disease [2]. DVT occurs when a blood clot forms within a deep vein, typically affecting the lower leg, thigh, or pelvis, while PE arises when a clot dislodges and migrates through the bloodstream to the lungs. VTE not only introduces complications during surgical procedures but also leads to extended hospital stays and heightened mortality rates when left undiagnosed [3]. In fact, the risk of VTE can surge by up to 20 times following surgical interventions [4]. Consequently, the timely detection of VTE assumes a critical role in shaping medical decisions, and the integration of automated methods for identifying VTE diagnosis holds promise for further advancements in healthcare practices.

This work was supported by UMBC/UMD ATIP 2022 and NIH AIM-AHEAD. * Corresponding author

The widespread implementation of electronic health record systems (EHRs) in the US hospitals presents a valuable opportunity to leverage advanced data analytics techniques for post-operative VTE classification. Clinical notes and reports include crucial information regarding postoperative complications [5]. To extract meaningful insights from these unstructured and free-text reports, natural language processing (NLP) utilizes computational linguistics to process and analyze the textual data. The application of NLP has seen a growing trend in the analysis of radiologist reports from medical imaging [6]. Considering that the diagnosis of VTE relies heavily on imaging findings, the application of NLP can assist in automatically identifying patients with VTE using radiology reports. To better understand NLP reports, we show a de-identified Ultrasound report and a partial CT-scan report (partial) format as follows:

Sample Ultrasound Report:
Right: There is persistent occlusive thrombus visualized at right gastrocnemius veins and right soleal veins. The right common femoral, proximal femoral and profunda femoris veins were not visualized due to the ECMO cannula. **Left:** There is persistent thrombus visualized at left posterior tibial veins, left peroneal veins, left gastrocnemius and left soleal veins.

Sample CT Scan Report (partial):
Examination: Contrast enhanced CT of the chest (CT pulmonary angiography)
Clinical History:: The patient is a 56-year-old male with tachycardia and shortness of breath with new oxygen requirements to evaluate for pulmonary embolism. The patient has a prior history of oral tongue malignancy and known pulmonary nodules.
.....
Impression:
After further review of the images, there is a small filling defect demonstrated within the subsegmental branch of the left lower lobe pulmonary artery adjacent to the major fissure that is consistent with pulmonary embolism.

Numerous studies conducted at individual institutions have developed NLP tools to analyze free-text medical reports

and notes. However, there are challenges and limitations that must be addressed in order to fully harness the potential of automated methods in VTE diagnosis. We discuss the limitation of *L1*, *L2* and *L3* as follows:

(*L1*): Achieving a probability of higher accuracy in machine learning tasks requires a well-labeled dataset. However, the availability of adequate numbers of de-identified and labeled VTE data is limited, and this scarcity is exacerbated by the problem of data imbalance. The scarcity of VTE medical data usually arises from various factors, such as privacy concerns (e.g., hospitals may be reluctant to share patient personal data), restricted data accessibility, and the challenges of gathering large-scale labeled datasets in the medical domain. This limited availability of VTE data poses difficulties for effectively training machine learning models since clinical experts have to read and label the data reports.

(*L2*): Transfer learning (TL) and pre-trained models have become popular in the fields of NLP and medical text analysis [7]–[12]. However, there is limited research discussing the use of pre-trained models to enhance model accuracy in VTE prediction. In such scenarios, transfer learning proves valuable by leveraging pre-trained models that have learned generic features from large-scale medical datasets or related tasks. These models can be fine-tuned on the limited VTE data available. However, choosing the best pre-trained model among the many available can be a challenging endeavor due to their significant variations. Each pre-trained model possesses distinct characteristics, making the selection process more complex.

(*L3*): Traditional NLP methods usually involve rule-based systems or statistical machine learning approaches [2], [5], [13]–[15]. Although rule-based approaches offer the advantage of requiring less training data and producing explainable results, the design process for these methods demands a substantial amount of effort by domain experts. Statistical approaches offer the benefit of requiring minimal effort during training. However, they require a large amount of training data to ensure accuracy and provide results based on probabilities.

To address the issue of limited datasets affecting model accuracy, we employed the Data Augmentation (DA) technique [16]. However, in the case of text data, traditional image-based DA techniques are not directly applicable. To address this, textual DA techniques are employed to generate additional text samples by applying word replacement, synonym substitution, sentence shuffling, and contextual augmentation. By leveraging these techniques, we can effectively increase the amount of training data, thereby helping to alleviate the data imbalance problem and slightly improved the model performance.

To discover the optimal pre-trained model, we have developed an Adaptive Pre-train Model Selection (APMS) algorithm. This intelligent algorithm dynamically selects the most appropriate pre-trained model based on the unique attributes of specific downstream tasks and data characteristics. By doing so, our aim is to enhance model performance and efficiency by leveraging the strengths of different pre-trained models to address the challenges posed by limited datasets in the context of VTE. We utilize the pre-trained BERT [17] model, specifically ClinicalBERT [18] selected by AMPS, for word embedding in medical texts. Subsequently, a bi-directional LSTM (Bi-LSTM) network is fine-tuned on the embedded representations to perform the classification task. The Bi-LSTM architecture involves stacking two LSTM layers together. This arrangement effectively enhances the information available to the network, thereby improving its ability to learn from the context. The dataset used consists of free-text medical reports obtained from University of Maryland Medical Center (UMMC) hospitals. These reports are de-identified and have been annotated by medical professionals.

Ultimately, we constructed a rule-based deep-learning model for the purpose of classifying the VTE dataset. This model utilizes a combination of rules and deep learning techniques to accurately categorize VTE and Non-VTE within the dataset based on predefined criteria. The integration of rule-based methods with deep learning enhances the model's ability to capture complex patterns and achieve more effective and accurate VTE classification. Importantly, the NLP model has the capability to automatically generate labels for the VTE dataset. This automated labeling process eliminates the need for manual annotation, significantly reducing human effort and potential errors.

We summarize our contributions as follows:

- The paper presents an automated approach for VTE classification using DL and NLP model, enabling timely detection and improved patient outcomes.
- An adaptive pre-train model selection (APMS) algorithm is proposed to dynamically choose the best pre-trained model for improved VTE classification.
- We applied the rule-based classifier, significantly enhancing the predictive capability of the DL model, especially in cases where the PE dataset is small and exhibits class imbalance. We also introduced Data Augmentation techniques to mitigate the impact of limited PE datasets, slightly enhancing model performance.
- We conducted plenty of experiments and evaluations. The results demonstrated the model's high effectiveness in predicting VTE events, achieving an impressive accuracy rate of 98.3%. These findings highlight the model's robustness and its potential to significantly contribute to VTE research.

II. Related Work

Traditionally, NLP systems for classification involved rule-based methods or statistical machine learning approaches. Rule-based methods necessitated considerable effort from domain experts for manual feature selection, while statistical approaches required a large volume of training data. Despite deep learning (DL) studies showing improved results, it is noteworthy that there are not many works utilizing DL methods for classifying VTE from medical report datasets because of the limited datasets.

Traditional approaches. Nelson et al. [2] combined statistical machine learning and rule-based NLP methods to identify

postoperative VTE among surgical patients treated in VA hospitals. However, their NLP system was proven unsuccessful and failed to adequately identify postoperative VTE events based on clinical notes. Tian et al. [13] randomly sampled radiology reports from a university health network of 5 hospitals in Montreal. The authors trained and utilized rule-based symbolic NLP classifiers from the dataset. They achieved 73% positive predictive value (PPV) on DVT and 80% PPV on PE. Sabra et al. [14] proposed a Semantic Extraction and a Sentiment Assessment of Risk Factors approach to produce feature inputs to a support vector machine classifier for VTE identification. Due to their small dataset of clinical narratives from electronic health records (EHR), the resulting F1 score was only 0.7. Shi et al. [5] extracted clinical notes from 2 independent healthcare systems. Their NLP system broke down a patient's report into sentence tokens. It identified relevant concepts by tokens and aggregated those semantic representations back to the document level, and eventually to the patient level for VTE classification. The results were an AUC of 0.9 for PE and an AUC of 0.92 for DVT. Verma et al. [15] employed an NLP algorithm, based on weighted regular expression rules, to classify radiologist reports of medical images for VTE. However, those rules were hand-picked by domain experts. Their approaches achieved a PPV of 0.90 and an AUC of 0.96 for identifying DVT; for PE, the results were a PPV of 0.89 and an AUC of 0.96.

Deep Learning methods. Many medical text classification tasks have taken advantage of Deep Learning approaches. Mulyar et al. [7] explored several architectures for modeling phenotyping that rely on BERT representations of free-text clinical notes. Olthof et al. [8] also concluded that the deep learning-based BERT model outperformed traditional ML and rule-based methods in radiology reports classification tasks. Goodrum et al. [9] extracted text from EHR and evaluated multiple text classification ML models, including bag-of-words and machine learning methods. The results showed that a deep learning model using ClinicalBERT performed best. They concluded that deep learning methods were effective in identifying clinically-relevant content. Lee et al. [10] found that RNN-based networks had the ability to classify significant findings in radiology reports with high F1 scores. A comparative analysis of text classification methods [11] studied the impact of various word representations, text pre-processing, and classification algorithms on different text classification tasks. Their results showed that the Bi-LSTM algorithm combined with Word2Vec embedding trained on MIMIC performs the best, BioBERT the second. For VTE risk factor identification tasks based on electronic medical records, a hybrid study [12] employed BERT for word embedding, and Bi-LSTM for information extraction. Then they used rule reasoning to judge the risk of PE. Experiment results showed that this method achieved 93.3% and 94.3% of entity and relation F1.

In contrast to their research ideas, we propose a DL model where we employ pre-trained ClinicalBERT for feature selection and a Bi-LSTM network for classification tasks. For the PE dataset, we employ a data augmentation method to generate synthetic data for training. We also enhance the prediction of the DL model with a rule-based classifier.

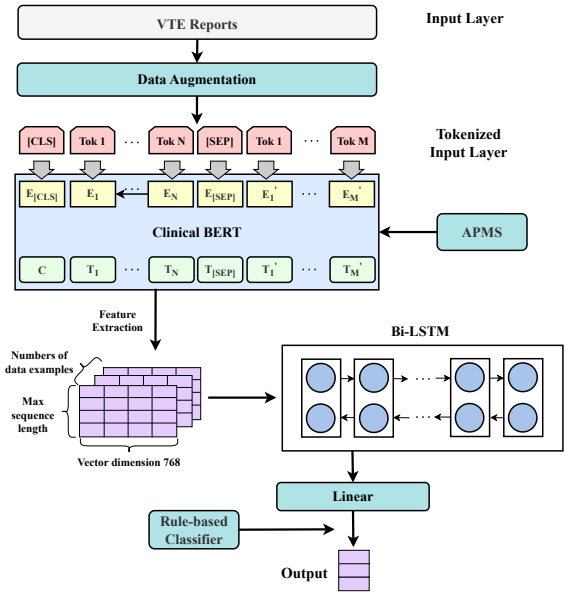

Fig. 1: Model structure. The ClinicalBERT layer transforms input text into word embeddings. [CLS] here is a special classification token and this token is used for classification tasks. [CLS] tokens store the vectors for classification tasks. Those features are fed into Bi-LSTM and linear layers for training.

III. Proposed Methods

We propose a deep learning (DL) model that comprises two main functions: (1) Feature selection: We applied a pre-trained ClinicalBERT model to convert medical texts into numerical representations. (2) Classification task: We applied bi-directional LSTM (Bi-LSTM) to train a model based on the embedded data and use the trained model for prediction tasks.

As shown in Figure 1, for the input text, data augmentations and Adaptive Pre-train Model Selection are performed. Then a tokenizer converts the text into tokens, attaching a [CLS] token to the beginning. The [CLS] tokens store the vectors for classification purposes. The pre-trained ClinicalBERT model processes the tokens and produces word embeddings from the input vectors. The classification layer of the output embeddings is extracted and fed to the Bi-LSTM layer. A linear layer is attached to Bi-LSTM and they are trained together for classification tasks. Each part of the model is described in detail below. After that, a rule-based classifier is attached to enhance the predictions of the DL model.

A. Textual Data Augmentation

Labeled and de-identified VTE data is scarce since it's time-consuming and costly to prepare the data. Also, medical data is sometimes imbalanced, as the positive examples are far fewer than the negative examples. Therefore we can apply the technique of data augmentation to artificially increase the

size and diversity of a textual dataset by generating new examples with slight modifications while preserving the original meaning. Textual data augmentation helps in improving the performance and robustness of NLP models, especially when faced with limited labeled or imbalanced data. An empirical study [19] suggests that for supervised learning, token-level augmentations, specifically word replacement and random swapping, consistently demonstrate the most enhancement in performance.

Algorithm 1: Data Augmentation Algorithm for VTE Text Classification

Require: Training VTE dataset $\mathcal{D}$ with labeled reports, augmentation parameters
Ensure: Augmented training dataset $\mathcal{D}'$
Initialize an empty augmented dataset $\mathcal{D}' = \{\}$
for n iterations **do**
  Randomly select a text-label pair from the minority class
  repeat
    $(aug_{\text{min}}, aug_{\text{max}})$ augments are produced
    Randomly select a sentence s from the text
    Tokenize the sentence into individual tokens
    if Synonym Replacement **then**
      Randomly select 1 token t within the sentence
      Look for the synonyms of t from the database, produce a list of synonyms
      Replace t with a randomly selected synonym from the list with probability p_{replace}
    else
      if Random Swapping **then**
        Randomly select 2 tokens t_1, t_2 within the sentence
        Swap the positions of t_1, t_2 with probability p_{swap}
      end if
    end if
    Add augmented text-label pair (x, y) to $\mathcal{D}'$
  until $(aug_{\text{min}}, aug_{\text{max}})$ augments are produced
end for
return augmented dataset $\mathcal{D}'$

Word replacement data augmentation methods involve replacing specific words in the training data with alternative words or synonyms. These techniques slightly change the wording in the text while preserving the overall meaning. We apply the commonly used synonym replacement in our experiments. This technique replaces a word with one of its synonyms. It helps diversify the vocabulary and introduces alternative expressions while maintaining semantic similarity. The synonym library we use is from the PPDB database [20]. We also experiment with random swapping, which is a data augmentation method used to generate new training samples by swapping words or tokens within a sentence while maintaining the overall sentence structure. The aim is to introduce variations in the data and can help improve model performance and generalization. The DA algorithm is shown in Algorithm 1.

This algorithm outlines the steps to perform data augmentation for the VTE classification task. It includes two types of possible transformation including synonym replacement and random swapping. The parameters p_{replace}, p_{swap}, control the probabilities of applying each transformation, while aug_{min} and aug_{max} determine the minimal and maximal numbers of words will be augmented. If aug_{max} is not given, the number of augmentation is calculated via p_{replace} or p_{swap}. If the calculated result from p is smaller than aug_{max}, will use the calculated result from p. Otherwise, using aug_{max}. Parameter n determines the number of synthetic samples that will be generated. The resulting augmented dataset $\mathcal{D}'$ contains the

Algorithm 2: Adaptive Pre-train Model Selection (APMS) Algorithm for VTE Dataset

Require: VTE dataset, list of pre-trained model options $\mathcal{M} = \{M_1, M_2, \ldots, M_k\}$, evaluation metric(s) $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$
Ensure: Optimal pre-trained model M^* for VTE task
Parameters:
  Number of pre-trained model options, k
  Number of evaluation metrics, m
Split the VTE dataset into training, validation, and test sets: $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}$.
for each pre-trained model $M_i \in \mathcal{M}$ **do**
  Initialize the model M_i with pre-trained weights.
  Fine-tune the model M_i on the training set for VTE task:
  for each evaluation metric $E_j \in \mathcal{E}$ **do**
    Add task-specific layers and loss functions for binary classification (e.g., VTE or non-VTE).
    Fine-tune the model M_i on the VTE-specific data using hyperparameters and optimization techniques.
  end for
  for each evaluation metric $E_j \in \mathcal{E}$ **do**
    Evaluate the fine-tuned model M_i on the validation set for VTE task using evaluation metric E_j.
  end for
end for
Identify the pre-trained model M^* with the best performance on the validation set for VTE task based on the evaluation metrics:

$$M^* = \arg\max_{M_i \in \mathcal{M}} \left(\sum_{E_j \in \mathcal{E}} E_j(M_i, \mathcal{D}_{\text{val}}) \right).$$

Fine-tune and evaluate the selected optimal model M^* on the test set for VTE task to obtain final performance results.
return The optimal pre-trained model M^* for the VTE task.

original images along with their augmented versions, ready for training a robust VTE classification model. We tested both synonym replacement and random swapping only on the CT scan reports (PE classification) dataset since the data contains fewer samples and is imbalanced.

B. Adaptive Pre-train Model Selection (APMS)

The Adaptive Pre-train Model Selection (APMS) algorithm is designed to dynamically and intelligently select the most suitable pre-trained model based on specific downstream tasks and data characteristics. The goal is to optimize model performance and efficiency by leveraging the strengths of different pre-trained models for various tasks. Algorithm 2 illustrates the pseudo-code summary of the APMS. The selection method is inspired by [21].

The selection process shows that the pre-trained ClinicalBERT [18] outperforms others in word embedding. The other two candidate methods are: (1) Original BERT, and (2) Clinical BioBERT, fine-tuned from BioBERT [22] with clinical notes. We select ClinicalBERT because of its superior performance [9] and its relevance to the domain of medical texts. ClinicalBERT is a publicly available word embedding model pre-trained on a large and publicly accessible collection of clinical notes: MIMIC-III v1.4 database, which contains approximately 2 million clinical notes.

C. Word Embedding with Clinical Expert Rule-based Classifier

Following data augmentation, the medical reports undergo tokenization, dividing radiology reports into token vectors limited to a maximum length of 512 tokens. These vectors are then converted into numerical representations using a pre-trained word embedding layer. The [CLS] tokens within these representations encapsulate all the necessary information for the classification task. These features, with a dimension of 768, are used as input to the classifier during training. The Bi-LSTM layer's output consists of both forward and backward sequences, which are concatenated before passing to the linear layer. Both the Bi-LSTM and linear layer are trained together during the fine-tuning process.

Given a limited dataset size and imbalanced classes, deep learning models often overfit on the majority (negative) class. To counter this issue, we leverage the strength of a rule-based expert system [23], which focuses on predicting the positive class. Specifically, we apply the CT-All PE ruleset which was developed by medical experts for identifying PE in CT scan reports [24]. By incorporating this ruleset, we aim to improve the predictions of our DL model on the PE - CT scan reports dataset. The ruleset consists of a series of regular expressions designed to match specific keywords within a CT scan report. Each match is assigned a score of -1, 0, or 1. The rule-based classifier first breaks down a report into sentences and then computes a sentence score by aggregating the scores of each match within that sentence. For example, if a sentence contains the keywords [segmental] and [filling], the sentence score is 1. However, if the sentence also contains keyword [no] OR [negative] OR [without] OR [question] OR [unchanged], the algorithm ignores previously assigned score 1, the sentence score remains 0. All sentence scores are then summed to produce a total score for the entire report. If the final score is greater than 0, the output prediction is positive for PE; otherwise, it is negative.

For the PE dataset, we combine the outcomes of both the DL classifier and the rule classifier. In cases where the DL classifier predicts a negative label but the rule classifier predicts a positive one, we prioritize the output of the rule classifier. However, there is an exception to this rule. If the DL classifier assigns a high probability (more than 95%) of the negative class and the report score is lower than 2, the final prediction remains negative.

IV. Experiment Results

A. VTE Datasets

We possess two datasets of medical imaging reports for VTE classification (DVT and PE). These datasets comprise de-identified and labeled medical reports. They were sourced from the University of Maryland Medical Center (UMD). The de-identification and labeling of datasets were done by medical experts from UMD.

The first dataset includes 1,000 free-text duplex ultrasound imaging reports. The reports were classified into 3 categories by a Radiologist: Class 0 - No acute DVT, Class 1 - Upper extremity acute DVT, and Class 2 - Lower extremity acute DVT. A total of 78% of data samples fall into the category of class 0, and 11% for class 1 and 2 respectively. The dataset consists primarily of structured reports containing concise texts, with the majority of them being less than 170 words in length.

The second dataset includes 900 free-text chest computed tomography (CT) angiography scan reports. It has fewer samples than the first dataset and is more imbalanced. The reports were classified into 2 categories: class 0 - No PE (88%), class 1 - PE (12%). These CT scan reports contain mostly unstructured texts and are longer in length. Most of them are around 200 words. Some reports exceed 600 words. The input size of a BERT model is limited to 512 tokens since high-dimensional vectors require larger computational power. Therefore longer text will be truncated to fit into the model and some of the information in the text will be lost. The reports also contain many special symbols, numbers, and punctuation. All of these increase the complexity of the CT scan reports dataset.

B. Experimental Settings

The experiment was run on a GPU-accelerated high-performance computing (HPC) system, built using IBM Power Systems AC922 servers. This system was designed to maximize data movement between the IBM POWER9 CPU and attached accelerators like GPUs. The GPU was an NVIDIA Tesla V100 GPU with a memory size of 16 GB. The experiments were run on the IBM Watson machine learning environment.

TABLE I: Performance of different techniques on the DVT dataset of Ultrasound reports.

Algorithm	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
ClinicalBERT + LSTM	0.885	0.885	0.88	0.92	0.885	0.887
ClinicalBERT + Linear	0.87	0.87	0.79	0.87	0.87	0.86
BioBERT + Bi-LSTM	0.955	0.955	0.94	0.957	0.955	0.955
BioBERT + LSTM	0.885	0.885	0.89	0.90	0.885	0.89
BioBERT + Linear	0.895	0.895	0.899	0.897	0.895	0.89
base BERT + Bi-LSTM	0.94	0.94	0.975	0.948	0.94	0.94
ClinicalBERT + Bi-LSTM (Ours)	**0.97**	0.97	0.93	0.97	0.97	0.97

TABLE II: Effectiveness of Data Augmentations on the PE dataset of CT scan reports.

Method	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
No Augmentation	0.89	0.89	0.45	0.88	0.89	0.885
Random Swapping	0.9	0.9	0.33	0.88	0.9	0.877
Synonym Replacement (Ours)	**0.911**	0.911	0.41	0.90	0.911	0.895

TABLE III: DL classifier and rule classifier results on PE - CT scan reports dataset

Method	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
Rule	0.972	0.972	0.955	0.975	0.972	0.973
DL	0.911	0.911	0.41	0.9	0.911	0.895
DL + Rule (Ours)	**0.983**	0.983	0.956	0.984	0.983	0.984

To evaluate the effectiveness of Transfer Learning, Data Augmentation, and Rule-based system, we conducted three sets of experiments. The first set utilized the DVT dataset, which consists of shorter and well-structured text from Ultrasound reports. We tested the ClinicalBERT and Bi-LSTM models proposed in this study, along with several baseline algorithms. In the second set of experiments, we focused on the PE dataset, which contains longer and more intricate text from CT scan reports. This dataset is limited in size and imbalanced. The third experiment aims to assess the effectiveness of integrating the capabilities of both a DL classifier and a rule-based classifier when dealing with the PE dataset.

We split the datasets into 80% training set and 20% test set. The training sets are further split into 90% train sets and 10% validation sets. For the DVT dataset that contains mostly short texts, the input texts are limited to a maximum of 170 tokens. Any input longer than that will be truncated to the right, shorter texts are padded. For the PE dataset, input texts are limited to 512 tokens, which is the maximum input size of ClinicalBERT. Longer texts are truncated to the left since we notice that some important information such as conclusions usually appear by the end of the CT scan reports. The Bi-LSTM network's input size is 768, which is the dimension of BERT's output [CLS] tokens. It is comprised of two layers, each having a hidden size of 256. A linear layer is appended to the Bi-LSTM network to form a classifier.

C. Model Performance

We compare the proposed method with some baseline contextual embedding techniques and classification methods. The baseline Transfer Learning methods for word embedding include:

- Original (base) BERT: this contextual word embedding network was trained on Wikipedia 2.5 billion words and Books Corpus 0.8 billion words. It's a general-purpose language representation model that can then be fine-tuned on small-data NLP tasks. BERT improves upon previous models by introducing deep bidirectionality and unsupervised learning. Unlike its predecessors, BERT is the first language model to be pre-trained solely on a plain text corpus.
- BioBERT fine-tuned on clinical notes: BioBERT is a domain-specific language representation model pre-trained on large-scale biomedical corpora of biomedical research articles: PubMed article abstracts and PubMed Central article full texts. It's designed for biomedical text-mining tasks. Alsentzer et al. [18] fine-tuned BioBERT on the MIMIC-III v1.4 database. Note that both ClinicalBERT and BioBERT were initialized with base BERT and then fine-tuned on other domain-specific databases.

The baselines of classification methods are the LSTM network and linear classifier. The LSTM network only consists of unidirectional layers, making it a more basic variant compared to the Bi-LSTM. In order to perform classification tasks, a linear layer is added to the LSTM network, similar to the Bi-LSTM approach, and both components are trained in conjunction. The linear classifier consists of two linear layers with 256 hidden sizes.

Table I shows the experiment results in terms of common metrics of weighted precision, recall, and F1 scores, as well as accuracy, sensitivity, and specificity. Our purposed method of ClinicalBERT and Bi-LSTM performs the best, with the high-

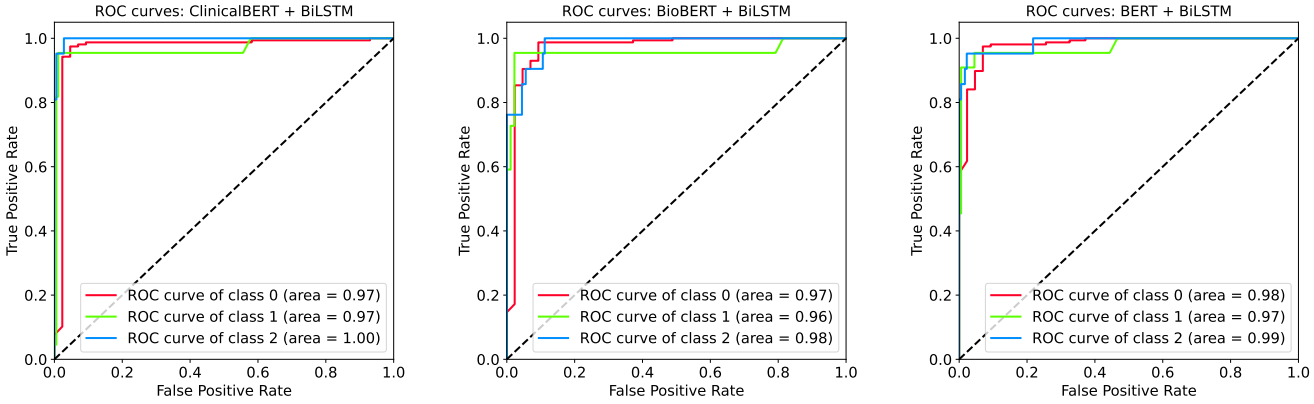

Fig. 2: ROC curves of different methods on the DVT dataset of Ultrasound reports. (Class 0 - No acute DVT, Class 1 - Upper extremity acute DVT, and Class 2 - Lower extremity acute DVT.)

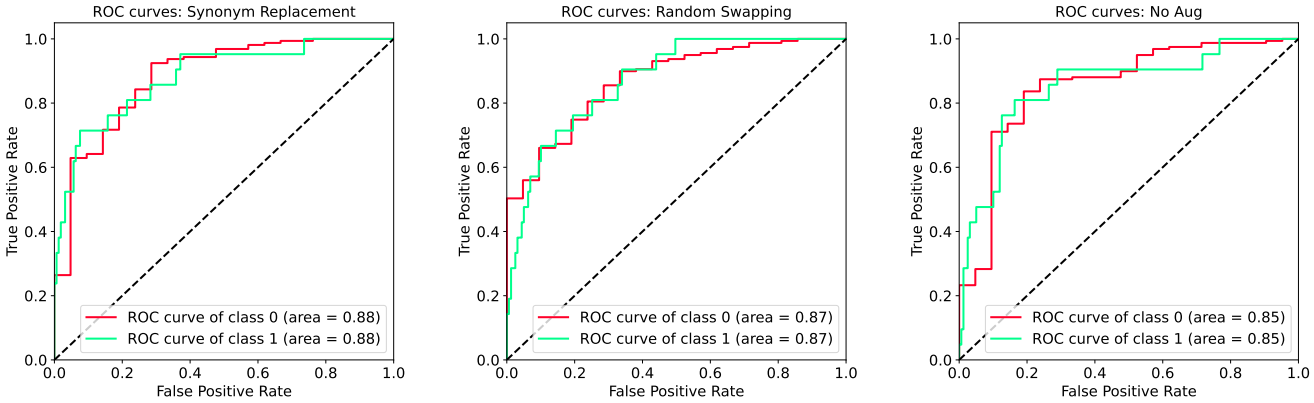

Fig. 3: ROC curves of different Data Augmentation methods on the PE dataset of CT scan reports. (class 0 - No PE, class 1 - PE.)

est values across all performance measures. Both BioBERT and BERT demonstrate strong performance, when combined with Bi-LSTM, yielding results that are close to the top-performing methods. Although BERTs exhibit good performance across all variants, their effectiveness starts to decrease when the classification methods are switched to LSTM or linear classifiers. The results indicate that the power of domain-specific ClinicalBERT embeddings effectively transfers to the VTE dataset, and the Bi-LSTM network performs better than the basic LSTM and linear classifiers. The ROC curves of the three best models are shown in Figure 2. They consist of different BERT embedding with Bi-LSTM classifiers. Their resulting AUCs are very similar.

Data augmentation techniques are only applied to the PE dataset, which is characterized by its limited size and imbalanced classes. The dataset consists of 900 CT scan reports, with 88% of them falling into the majority class labeled as negative for pulmonary embolism (PE). Before word embedding, the text undergoes two types of data augmentation: Synonym Replacement and Random Swapping. Each of these techniques has a few adjustable parameters. Both methods generate 200 synthetic instances specifically for the minority class. Incorporating more supplementary data would lead the model to rapidly overfit the training set. The influence of augmentation probability is significant in determining the outcomes. In the case of Synonym Replacement, a higher probability of 0.8 is preferred to improve performance. Conversely, for Random Swapping, a lower probability of 0.2 is more inclined to yield favorable results. The minimal number of augmentations is 30 for both methods.

In Table II, the outcomes of two different Data Augmentation methods implemented on the PE dataset are displayed. Synonym Replacement demonstrates slightly superior performance when compared to Random Swapping. Both techniques exhibit better results than not employing any augmentation. However, both augmentation methods still tend to overfit the training data, causing the trained models to correctly predict more samples of the majority class, but perform slightly worse when predicting the minority class. Hence No Augmentation method produces a higher specificity score. The ROC curves of different data augmentation methods are shown in Figure 3. The two augmentation methods produce slightly better AUCs than no augmentation approach.

Table III presents the performances of DL and rule-based

classifiers on the PE dataset. The integration of the rule classifier's predictions into our proposed DL model leads to a substantial enhancement in the results. Notably, all evaluation metrics show improvement, with specificity experiencing a remarkable increase from 0.41 to 0.956. The incorporation of rule-based systems plays a crucial role in enhancing the DL model's predictive capacity, especially for the rare class. This integration effectively addresses the challenges posed by imbalanced datasets and significantly improves the model's ability to accurately classify instances of the rare class on the PE dataset.

V. Conclusion

In this study, we have successfully utilized Deep Learning (DL) and NLP techniques to effectively identify venous thromboembolism (VTE) based on freetext clinical reports obtained from medical imaging. Our approach incorporates advanced NLP methods, including ClinicalBERT for word embedding and Bi-LSTM networks for model training, leading to the transformation of textual data into numerical features. To optimize our model's performance, ClinicalBERT was fine-tuned on corpora of radiology reports, making it particularly efficient at handling Natural Language Processing (NLP) tasks in identifying VTE events. Additionally, we addressed the challenges posed by the complexity and data imbalance of classifying PE through the application of a textual Domain Adaptation (DA) method and an APMS pre-trained model section algorithm. To further enhance accuracy, a clinical expert rule-based approach was introduced, which showed notable improvements in the DL model's performance. As a result, our model achieved impressive results, boasting a remarkable 97% accuracy and F1 score on the DVT dataset and an exceptional 98.3% accuracy and 98.4% F1 score on the PE dataset. The experimental findings substantiate the efficacy of NLP Transfer Learning approaches and NLP rule-based methods for medical text classification tasks.

References

[1] A. T. Cohen, G. Agnelli, F. A. Anderson, J. I. Arcelus, D. Bergqvist, J. G. Brecht, I. A. Greer, J. A. Heit, J. L. Hutchinson, A. K. Kakkar, *et al.*, "Venous thromboembolism (VTE) in Europe," *Thrombosis and haemostasis*, vol. 98, no. 10, pp. 756–764, 2007.

[2] R. E. Nelson, S. D. Grosse, N. J. Waitzman, J. Lin, S. L. DuVall, O. Patterson, J. Tsai, and N. Reyes, "Using multiple sources of data for surveillance of postoperative venous thromboembolism among surgical patients treated in department of veterans affairs hospitals, 2005–2010," *Thrombosis research*, vol. 135, no. 4, pp. 636–642, 2015.

[3] B. Woller, A. Daw, V. Aston, J. Lloyd, G. Snow, S. M. Stevens, S. C. Woller, P. Jones, and J. Bledsoe, "Natural language processing performance for the identification of venous thromboembolism in an integrated healthcare system," *Clinical and Applied Thrombosis/Hemostasis*, vol. 27, p. 10760296211013108, 2021.

[4] R. H. White and M. C. Henderson, "Risk factors for venous thromboembolism after total hip and knee replacement surgery," *Current opinion in pulmonary medicine*, vol. 8, no. 5, pp. 365–371, 2002.

[5] J. Shi, J. F. Hurdle, S. A. Johnson, J. P. Ferraro, D. E. Skarda, S. R. Finlayson, M. H. Samore, and B. T. Bucher, "Natural language processing for the surveillance of postoperative venous thromboembolism," *Surgery*, vol. 170, no. 4, pp. 1175–1182, 2021.

[6] 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.

[7] A. Mulyar, E. Schumacher, M. Rouhizadeh, and M. Dredze, "Phenotyping of clinical notes with improved document classification models using contextualized neural language models," *arXiv preprint arXiv:1910.13664*, 2019.

[8] A. W. Olthof, P. Shouche, E. Fennema, F. IJpma, R. Koolstra, V. Stirler, P. M. van Ooijen, and L. J. Cornelissen, "Machine learning based natural language processing of radiology reports in orthopaedic trauma," *Computer methods and programs in biomedicine*, vol. 208, p. 106304, 2021.

[9] H. Goodrum, K. Roberts, and E. V. Bernstam, "Automatic classification of scanned electronic health record documents," *International journal of medical informatics*, vol. 144, p. 104302, 2020.

[10] C. Lee, Y. Kim, Y. S. Kim, and J. Jang, "Automatic disease annotation from radiology reports using artificial intelligence implemented by a recurrent neural network," *American Journal of Roentgenology*, vol. 212, no. 4, pp. 734–740, 2019.

[11] A. Mascio, Z. Kraljevic, D. Bean, R. Dobson, R. Stewart, R. Bendayan, and A. Roberts, "Comparative analysis of text classification approaches in electronic health records," *arXiv preprint arXiv:2005.06624*, 2020.

[12] J. Chen, J. Yang, and J. He, "Prediction of venous thrombosis chinese electronic medical records based on deep learning and rule reasoning," *Applied Sciences*, vol. 12, no. 21, p. 10824, 2022.

[13] Z. Tian, S. Sun, T. Eguale, and C. M. Rochefort, "Automated extraction of vte events from narrative radiology reports in electronic health records: a validation study," *Medical care*, vol. 55, no. 10, p. e73, 2017.

[14] S. Sabra, K. M. Malik, and M. Alobaidi, "Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives," *Computers in biology and medicine*, vol. 94, pp. 1–10, 2018.

[15] A. A. Verma, H. Masoom, C. Pou-Prom, S. Shin, M. Guerzhoy, M. Fralick, M. Mamdani, and F. Razak, "Developing and validating natural language processing algorithms for radiology reports compared to icd-10 codes for identifying venous thromboembolism in hospitalized medical patients," *Thrombosis Research*, vol. 209, pp. 51–58, 2022.

[16] 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.

[17] 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.

[18] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[19] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An empirical survey of data augmentation for limited data learning in nlp," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 191–211, 2023.

[20] E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, (Beijing, China), pp. 425–430, Association for Computational Linguistics, July 2015.

[21] K. You, Y. Liu, J. Wang, and M. Long, "Logme: Practical assessment of pre-trained models for transfer learning," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, (Virtually), pp. 12133–12143, PMLR, Jul 2021.

[22] 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.

[23] J. Villena Román, S. Collada Pérez, S. Lana Serrano, and J. C. González Cristóbal, "Hybrid approach combining machine learning and a rule-based expert system for text categorization," in *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, (Palm Beach, Florida), pp. 323–328, AAAI, 2011.

[24] A. A. Verma, H. Masoom, C. Pou-Prom, S. Shin, M. Guerzhoy, M. Fralick, M. Mamdani, and F. Razak, "Developing and validating natural language processing algorithms for radiology reports compared to icd-10 codes for identifying venous thromboembolism in hospitalized medical patients," *Thrombosis Research*, vol. 209, pp. 51–58, 2022.