Building a trustworthy AI differential diagnosis application for Crohn's disease and intestinal tuberculosis

Keming Lu
Tsinghua University
Yuanren Tong
Peking Union Medical College Hospital
Si Yu
Peking Union Medical College Hospital
Yucong Lin
Tsinghua University
Yingyun Yang
Peking Union Medical College Hospital
Hui Xu
Peking Union Medical College Hospital
Yue Li (yuelee76@gmail.com)
Peking Union Medical College Hospital https://orcid.org/0000-0001-6799-1812
Sheng Yu
Tsinghua University

Research article

Keywords: Neural network, integrated gradients, knowledge distillation, Crohn's disease, intestinal tuberculosis

Posted Date: June 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1625845/v1

Building a trustworthy AI differential diagnosis application for Crohn's disease and intestinal tuberculosis

Author list: Keming Lu[1†];Yuanren Tong[2†]; Si Yu[2]; Yucong Lin[3,4]; Yingyun Yang[2]; Hui Xu[2]; Yue Li[2*]; Sheng Yu[3,4*]

1.Department of Automation, Tsinghua University, Beijing, China, 100084

2. Department of gastroenterology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, 100730

3. Center for Statistical Science, Tsinghua University, Beijing, China, Beijing, China, 100084

4. Department of Industrial Engineering, Tsinghua University, Beijing, China, 100084

† These authors contributed equally to this article.

Mr. Yuanren Tong: tongyr14@mails.tsinghua.edu.cn

Mr. Keming Lu: lkm16@mails.tsinghua.edu.cn

Ms. Si Yu: yusi0518@126.com

Dr. Yucong Lin: linyc16@mails.tsinghua.edu.cn

Dr. Yingyun Yang: yingyunyang@126.com

Dr. Hui Xu: heronxu@aliyun.com

Dr. Yue Li: yuelee76@gmail.com

Dr. Sheng Yu: syu@tsinghua.edu.cn

***Corresponding authors:**

Yue Li, M.D.

yuelee76@gmail.com

Tel/Fax: +86-10-69155751

Department of Gastroenterology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, 100730

Sheng Yu, Ph.D.

syu@tsinghua.edu.cn

Tel/Fax: +86-10-62783842

Center for Statistical Science& Department of Industrial Engineering& Institute for Data Science, Tsinghua University, Beijing, China, 100084

Word Count: 3864

Abstract

Background: Differentiating between Crohn's disease (CD) and intestinal tuberculosis (ITB) with endoscopy is challenging. We aim to perform more accurate endoscopic diagnosis between CD and ITB by building a trustworthy AI differential diagnosis application.

Methods: A total of 1271 electronic health record (EHR) patients who had undergone colonoscopies at Peking Union Medical College Hospital (PUMCH) and were clinically diagnosed with CD (n=875) or ITB (n=396) were used in this study. We build a workflow to make diagnoses with EHRs and mine differential diagnosis features; this involves finetuning the pretrained language models, distilling them into a light and efficient TextCNN model, interpreting the neural network and selecting differential attribution features, and then adopting manual feature checking and carrying out debias training.

Results: The accuracy of debiased TextCNN on differential diagnosis between CD and ITB is 0.83 (CR F1: 0.87, ITB F1: 0.77), which is the best among the baselines. On the noisy validation set, its accuracy was 0.70 (CR F1: 0.87, ITB: 0.69), which was significantly higher than that of models without debias. We also find that the debiased model more easily mines the diagnostically significant features. The debiased TextCNN unearthed 39 diagnostic features in the form of phrases, 17 of which were key diagnostic features recognized by the guidelines.

Conclusion: We build a trustworthy AI differential diagnosis application for differentiating between CD and ITB focusing on accuracy, interpretability and robustness. The classifiers perform well, and the features which had statistical significance were in agreement with clinical guidelines.

Key words: Neural network, integrated gradients, knowledge distillation, Crohn's disease, intestinal tuberculosis

1 Background

Crohn's disease (CD) is a chronic and idiopathic inflammatory disease that usually has a disease course with repeating remission-relapses. Intestinal tuberculosis (ITB) is an infectious intestinal disease caused by *Mycobacterium tuberculosis*. The treatment, progression, and prognosis of CD and ITB are different, and the initial correct diagnosis and differentiation between CD and ITB are of critical importance.

Endoscopy is an essential examination for a timely and accurate diagnosis and is always conducted first[16]. However, the differential diagnosis between CD and ITB can be challenging because the two diseases have a very similar endoscopic appearance. Therefore, diagnosis relies heavily on the experience of the clinician who conducts the examination. This situation often causes incorrect endoscopic diagnosis and results in delayed treatment.

This study aims to facilitate correct interpretation of endoscopic reports and differentiation between CD and ITB using natural language processing. Furthermore, we aim to provide a workflow for obtaining trustworthy neural

network classifiers using texts, particularly unstructured texts, such as electronic health records (EHRs). We define a trustworthy neural network as a neural network that can be explained with human understandable phrase features that allow doctors to understand how the model reaches a certain conclusion.

Artificial intelligence (AI) is widely used in the medical field and has been applied to differentiate CD and ITB. However, as the model becomes increasingly complex, the inability of AI users to interpret the decision process has become problematic. Classical AI models, such as support vector machines, random forests and neural networks, are commonly described as "black boxes" due to the lack of interpretability. The interpretability of the AI model in the medical field is an important metric for the following reasons: 1) clinicians should be able to judge if the prediction of the model is reasonable; 2) new interpretable features found by the model can be further verified through clinical studies so that guidelines of the disease can be updated; and 3) clinicians are professionally conservative, and an interpretable model will be more readily accepted than a black-box model.

Recently, research on explanation methods in deep learning has emerged. The integrated gradient (IG) method has the property of being model agnostic and can be derived everywhere for the model parameters. Compared with other methods, the computational cost of IG is relatively small, and therefore it is selected as the interpretation method in our work. Sundararajan et al.[12] show the explanatory effect of IG in the fields of text classification and question answering. In addition, because IG has a small computational cost and derivability in all cases, it is also used

to integrate prior knowledge or to correct bias as described by Liu et al.[8] The attribution method represented by IG often means that it can obtain interpretability at the token level, which is still challenging to understand. Chen et al.[1] and Singh et al.[11] proposed a hierarchical interpretation method based on contextual decomposition to solve this problem. They obtained the interpretability of the model for features of different scales. All of these works inspire us to build an interpretable deep learning AI diagnosis system. However, all of the results in the previous works are based on a corpus in English. Few methods and experiments focus on interpreting neural networks with IG in Chinese corpora.

Several works also use neural networks to explain or obtain medical concepts in the medical image processing field. Graziani et al.[3] propose a framework that shifts the attribution focus from pixel values to user-defined images. Experts can explain and trust the network output by checking whether specific diagnostic measures are present in the learned representations. Hu et al.[5] construct a diagnosis model for COVID-19 with CT images and weakly supervised lesion localization with IG. Preuer et al. [9] employed IG to identify the most relevant components of a compound for network prediction of molecular properties and bioactivities. Lauritsen et al.[7] present the Xai EWS - an explainable AI early warning score (EWS) system for predicting acute critical illness using EHRs. Sayres et al.[10] investigate the effect of 2 types of visualization models to indicate diabetic retinopathy scores and expansion heatmaps on the accuracy, speed, and confidence of readers. However, there are few works on building a trustworthy diagnosis application with text data.

Present work. We introduce a workflow to build a trustworthy AI differential diagnosis system for Crohn's disease and intestinal tuberculosis. And we also analyze significant diagnostic features we mined. Figure 1 illustrates the whole process of the proposed workflow. From our perspective, a trustworthy AI diagnosis system should have the properties of correctness, interpretability, and robustness. More specifically, correctness means that the classifier is expected to have acceptable accuracy in differential diagnosis; interpretability indicates that doctors know how the classifier works to achieve the diagnosis; robustness indicates that the classifier should not overfit meaningless features in the data and is expected to be mining features with medical significance. This work proposes a 6-step workflow to build a trustworthy differential diagnosis system for Crohn's disease and intestinal tuberculosis:

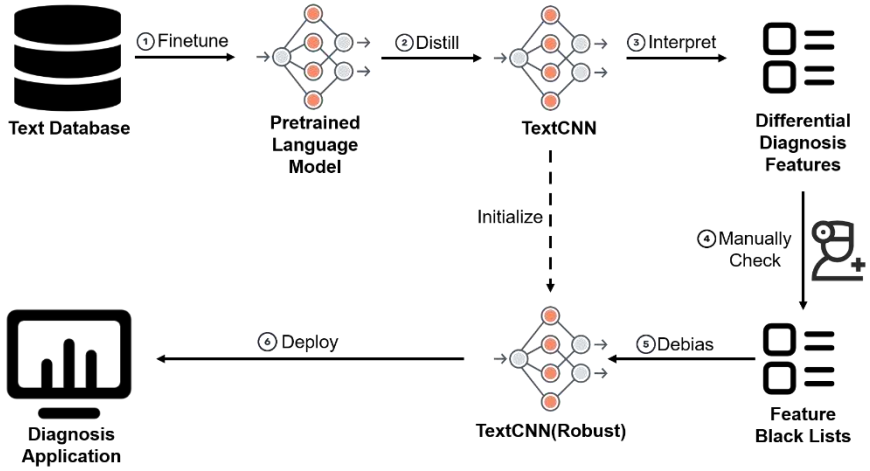

Figure 1. The workflow of building a text-based trustworthy diagnosis model.

1. **Finetune.** In the first step, we finetune a pretrained language model with text description as a classification problem.
2. **Distill.** We distill the finetuned pretrained language model into a TextCNN model.

3. **Interpret.** We use Integrated Gradients method to obtain local interpretation of all samples. Then, hierarchical phrase features are selected and filtered by statistical significance as differential diagnosis features.
4. **Manually Check.** Medical doctors label the differential diagnosis features with medical guidelines and professional knowledge. A set of features that are meaningless or apparent are selected into a blacklist.
5. **Debias.** We do a debias training by adding an attribution penalty to the loss function. After debias training, the TextCNN model has zero attributes on meaningless features in the blacklist.
6. **Deploy.** Finally, we deploy this model as a web service. Doctors can query with text descriptions and obtain classification results and visualization of attribution.

To summarize, this study aims to make endoscopic diagnosis of CD and ITB more accurate with the help of natural language processing (NLP) and statistical analysis and builds a trustworthy diagnosis application. The novelty of this workflow is that it employs high precision neural networks and cutting-edge interpretation methods to significantly reduce workloads of clinicians in human-in-loop data mining. Clinicians can only check features instead of predictions to debias the model and make it provide trustworthy results. The workflow can improve the diagnostic accuracy between CD and ITB with fewer risks in clinical application. The codes used in this work are provided on Github.[1]

[1] https://github.com/Lukeming-tsinghua/Interpretable-NN-for-IBD-diagnosis

2 Methods

2.1 Notations

We define D as a labeled text dataset with N samples: $D = \{(\boldsymbol{t}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{t}_i$ is the token sequence of the i-th endoscopy report. The granularity and the tokenization method are determined by the downstream model. In the pretrained model, the granularity of the token is character-level; y_i and $\hat{y}_i$ are the actual and predicted d-dimensional one-hot vectors, where d is the number of categories. The model aims to predict $\hat{y}_i$ from $\boldsymbol{t}_i$ and further obtain a sequence $FT = \{(t_{i,b_k}, t_{i,b_k+1}, \dots, t_{i,e_k})\}_{k=1}^{K}$ that represents the features used by the model when conducting the classification task, and b_k and e_k are the start and end indices of the k-th feature. The FT set is important for the differential diagnosis between CD and ITB.

2.3 Methods

This section introduces the development steps of our system. The PTM is first finetuned with labeled training data to obtain a classification model with good diagnostic performance. Then, this large model is distilled into a light TextCNN model. After that, we interpret the distilled TextCNN model with IG and design an analysis method to extract differential attribution features, including hierarchical feature set extraction and a feature selection pipeline.

2.3.1 Finetuning pretrained language model

Language model pretraining is an effective approach for improving many natural language processing tasks. RoBERTa-wwm-ext[2] is a state-of-the-art model for conducting text classification in Chinese. This model was trained on Chinese texts with the same architecture of RoBERTa using the whole word masking (wwm) strategy that replaced tokens with mask labels after Chinese tokenization when conducting the masking strategy used in BERT[14]. We chose RoBERTa-wwm-ext for its excellent effect on multiclassification tasks on Chinese text. RoBERTa-wwm-ext can be replaced by other BERT-like models; thus, we refer to RoBERTa-wwm-ext as the pretrained model (PTM) in this article.

The input text is segmented to tokens t_i by the Chinese word segmentation tool LAC[13]. Special markers are added to t_i for the PTM, and the input tokens become $\hat{\boldsymbol{t}}_i = \{[CLS], t_{i,1}, t_{i,2}, \dots, t_{i,n}, [SEP]\}$, where $[CLS]$ and $[SEP]$ are the reserved special tokens for identifying the beginning and end of sentences. For each input text, we use the hidden vector of $[CLS]$ as the embedding of the input. The softmax result after the linear layer was used as the probability for classification:

$$\boldsymbol{h}_i^{CLS} = PTM(\hat{\boldsymbol{t}}_i),$$

$$\boldsymbol{p}_i = softmax(\boldsymbol{W}\boldsymbol{h}_i^{CLS} + \boldsymbol{b}),$$

where $\boldsymbol{h}_i^{CLS} \in R^{d_h}$ is the representation of the output of the PTM; d_h is the dimension of the hidden layer; and $\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable parameters of the linear

layer. $\boldsymbol{p}_i \in R^d$ is the probability for classification. Due to the extremely unbalanced samples in the research, we used the focal loss[15] as the loss function:

$$L = \sum_{i \in D} \sum_{k=1}^{d} -\alpha_k y_{i,k} (1 - p_{i,k})^{\gamma} log(p_{i,k}),$$

where α_k is the weight of each classification, γ is the balance factor, and $y_{i,k}$ is the true label.

2.3.2 Distilling the PTM into TextCNN

TextCNN is a convolutional neural network for text classification proposed by Kim et al.[6] The input of the model is a sentence, represented as a sequence of word vectors. Let $\boldsymbol{x}_i$ be the word vector corresponding to the i-th word in the sentence with length n. The input can be defined as the concatenation of all of the word vectors:

$$\boldsymbol{x}_{1:n} = \bigcup_{i=1}^{n} \boldsymbol{x}_i,$$

where the union symbol denotes vector concatenation and $\boldsymbol{x}_{1:n}$ denotes the concatenation of the word vectors between the 1st word and the n-th word. A convolution filter matrix $\boldsymbol{w}$ is applied to a window of h words to obtain the new feature:

$$c_i = f(\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h-1} + b),$$

where c_i is a new feature, b is a bias term and f is a nonlinear activation function. This filter is applied to all possible windows in the sentences to obtain a feature list $\boldsymbol{c} = [c_1, c_2, \dots, c_{n-h+1}]$. Then, a max pooling operation is employed on this feature list

to obtain the feature corresponding to this filter $\hat{c} = \max(\boldsymbol{c})$. All max features of various filters are combined as $\boldsymbol{h}$, and the logit is obtained with a linear layer:

$$z = f(\boldsymbol{W} \cdot \boldsymbol{h} + \boldsymbol{b})$$

We distill the finetuned model into TextCNN for two purposes. First, prediction in RoBERTa-wwm-ext is time-consuming and will result in low efficiency. A helpful method is to distill RoBERTa-wwm-ext into TextCNN, which is a significantly faster model. Second, TextCNN is a neural network with word-level features that is easier to interpret. The distillation procedure in our methods follows Hinton et al.[4]. We use f_t and f_s to denote the PTM model and the TextCNN model, respectively. Logits of each sample are first calculated according to:

$$\boldsymbol{z}_t = f_t(\boldsymbol{x}_i)\,, \boldsymbol{z}_s = f_s(\boldsymbol{x}_i),$$

where $\boldsymbol{x}_i$ denotes the i-th sample in the training set. Then, a Kullback–Leibler divergence between the softmax logits of the teacher and student models is calculated as the distillation loss function:

$$L = T^2 \sum_i \frac{exp(\boldsymbol{z}_t/T)}{\sum_j exp(\boldsymbol{z}_t/T)} log(\frac{exp(\boldsymbol{z}_s/T)}{\sum_j exp(\boldsymbol{z}_s/T)}),$$

where T is a temperature constant. The training loss may also include classification loss as the hard labels, but we only use the distillation loss since this knowledge distillation loss achieves better performance in our work.

2.3.3 Differential attribution analysis

Differential attribution analysis aims to identify understandable N-gram features that have significant differences in attribution between different diseases. These differential attribution features are the differential diagnosis features of the neural network models.

Algorithm 1. Differential attribution analysis.

Input: The training set D_T, the trained neural network M, D_T size N

Output: A candidate N-gram phrase set S_k, A N-gram feature setS_o

$S_k = \emptyset$

For i **in** 1:N

$a_i = LocalExplanation(D_T[i])$//See **Local explanation with LG**

$S_k = S_k \cup FeatureSetExtraction(D_T[i], a_i)$//See **Hierarchical feature set extraction**

$S_k = Unique(S_k), K = size(S_k)$

$\boldsymbol{A} = \mathbf{0}, \boldsymbol{A} \epsilon \boldsymbol{R}^{N,K}$

For i **in** 1:N

For k **in** 1:K

$\boldsymbol{A}[i,k] = CalculateAttribution(D_T[i],\ S_k[j])$// See **Hierarchical feature set extraction**

$S_o = \text{FeatureSelection}(S_k, \boldsymbol{A})$// See **Statistical feature selection**

Local explanation with IG. We calculate the attribution of input with the IG method to identify the most important features for classification. IG is an attribution method for neural networks. Attributions are contributions of inputs to the prediction. Formally, suppose a function $F: R^n \rightarrow [0,1]$ represents the classification function of the PTM, and the token embedding of the input is denoted as $H_i = (h_{i,1}, \dots, h_{i,l}) \in R^{l \times d_e}$. d_e is the dim of token embeddings. An attribution of the prediction at input H_i relative to a baseline input H' is a vector $A_F(H_i, H') = (a_{i,1}, \dots, a_{i,l}) \in R^{l \times d_e}$, where $a_{i,k}$ is the contribution of $h_{i,k}$ to the prediction $F(x)$. In our work, we use token embedding of the padding token as the reference baseline input. The IG method conforms to the two axioms of attribution methods namely sensitivity and implementation invariance of the gradient, requires no modification on the neural network architecture and is simple to implement. Therefore, we choose IG as the attribution method in this work.

Hierarchical feature set extraction. Words and N-gram phrases are more explainable to humans than individual Chinese characters. Therefore, after obtaining the attributions of the input character tokens, we further derive a hierarchical feature set of words and phrases along with their attributions. Denoting

the sample as a Chinese character sequence $t = \{t_1, \dots, t_l\}$ with attributions $a = \{a_1, \dots, a_l\}$, we can segment this sequence with Chinese word segmentation and obtain word-level tokens $w = \{w_1, \dots, w_m\}$ with attributions $a_w = \{a_{w1}, \dots, a_{wm}\}$, which are calculated by $a_{wi} = \sum_{t_j \in w_i} a_j$. Then, we form a set of phrases $p = \{\{w_{p1_1}, \dots, w_{p1_k}\}, \dots, \{w_{pn_1}, \dots, w_{pn_k}\}\}$ that are successive words whose attribution w_{pi_j} is larger than the 0.9 quantile of a_w. Then, N-grams (up to 3 words) are generated from each phrase in p. The feature set will be the union of N-gram sets obtained from each sample. This N-gram feature set is the set of candidates for differential features.

After collecting the set of N-gram candidate features, the attributions of each feature in all training samples are calculated and arranged as an attribution matrix $A \in R^{N \times K}$, where N is the size of the training set and K is the number of candidate features (Figure 2).

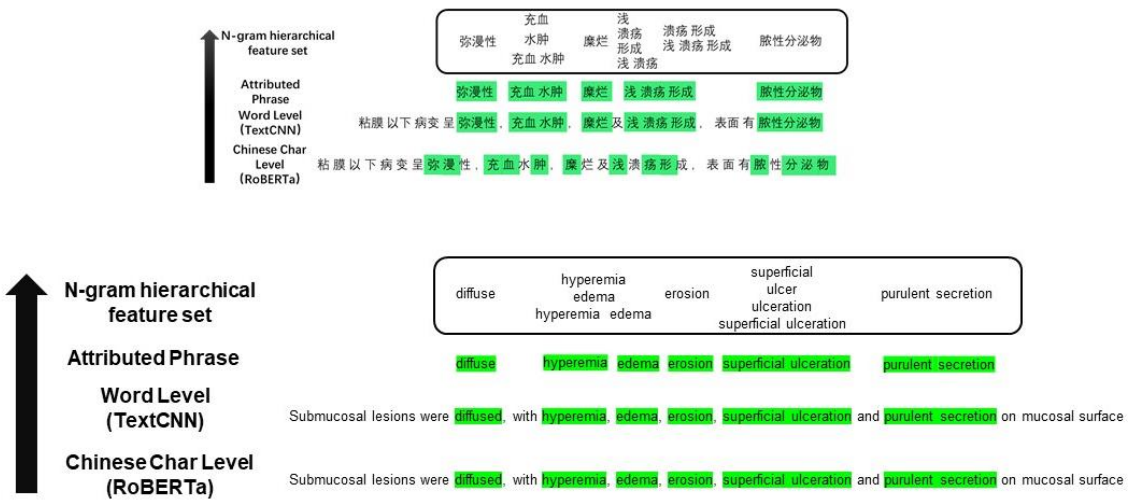

Figure 2. A case demonstration of hierarchical feature set extraction: words or characters in the sentence with positive attribution scores are highlighted with a green background. The extraction process constructs an N-gram hierarchical feature set from bottom (word or character level) to top.

Statistical feature selection. After obtaining the hierarchical feature set and calculating the attribution matrix $\boldsymbol{A}$, we further analyze this matrix and perform feature selection to obtain the differential diagnosis features. A feature can be represented by an attribution vector $\boldsymbol{a}_k \in R^{1\times N}$ in the attribution matrix. The i-th value in $\boldsymbol{a}_k$ is the attribution of feature k in the i-th sample. We rank the variance of $\{\boldsymbol{a}_k\}_{k=1}^{K}$ and select 50 features with the largest variance.

Then, we use a t-test to further select the features with significantly different attributes between the classes. We denote the class labels as C. When selecting a feature that is highly attributed in the samples of class c and shows relevantly low attribution in other classes, a t-test is employed to calculate the statistical significance. Let $I(c)$ represent the index set of samples with class c. The t statistic is calculated as

$$t_k = \frac{\frac{1}{|I(c)|}\sum_{j\in I(c)} a_j - \frac{1}{n-|I(c)|}\sum_{j\notin I(c)} a_j}{\sqrt{\frac{\sum_{j\in I(c)}(a_j - \frac{1}{|I(c)|}\sum_{j\in I(c)} a_j)^2}{|I(c)|^2} + \frac{\sum_{j\notin I(c)}(a_j - \frac{1}{n-|I(c)|}\sum_{j\notin I(c)} a_j)^2}{n-|I(c)|^2}}}$$

The p value $p_k, k = 1,2, \dots, K$ can be obtained for each feature, and we rank p values as $p_1 \leq p_2 \leq \cdots \leq p_K$. Since this is a multiple comparison, we employed the Benjamini-Hochberg method to control the false discovery rate (FDR) at 0.01. This adjustment begins with $q_K = p_K$ and sequentially calculates q_k from the largest index by the following rules:

$$q_k = \begin{cases} p_k \times \frac{m}{k} & , \quad p_k \times \frac{m}{k} \leq q_{k+1} \\ q_{k+1} & , \quad p_k \times \frac{m}{k} \geq q_{k+1} \\ 1 & , \quad p_k \times \frac{m}{k} \geq 1 \end{cases}, \mathrm{k} = 1, 2, \dots, \mathrm{K}$$

The features with $q_k \leq 0.01$ will be selected and ranked by difference $d = \frac{1}{|I(c)|}\sum_{j \in I(c)} a_j - \frac{1}{n-|I(c)|}\sum_{j \notin I(c)} a_j$. These features are differential attribution features of class c.

2.3.6 Debias finetuning by attribution penalty

Differential attribution analysis proposes a method to identify readable diagnosis features that the neural networks rely on. However, we find that the features extracted by the above methods indicate that neural networks make classifications with inappropriate and unwanted features. Therefore, we include a debias finetuning processing in our workflow that allows clinical doctors to adjust model performance using their professional knowledge.

First, a blacklist of unwanted features of each disease is manually selected from differential attribution features. For example, a blacklist containing disease names themselves is as follows.

$$Blacklist = \{CD: [Crohn's\ disease], ITB: [intestinal\ tuberculosis]\}$$

The above blacklist means that when the model classifies a real CD sample in fine tuning, the word "Crohn's disease" is expected to be a neutral feature. To achieve that, we add an attribution penalty to the classification loss and fine tune the model.

$$Loss = \frac{1}{n}\sum_{i=1}^{n} \text{FocalLoss}(y_i, \hat{y}_i) + \lambda \times \frac{1}{l_i}\sum_{j=1}^{l_i}(a_{ij} - target_{ij})^2$$

λ is a hyperparameter determined by cross-validation, and l_i denotes the length of the i-th sentence in the token. $target_{ij}$ is defined as a tokenwise label. If a token is included in the blacklist, this label equals 0. Otherwise, $target_{ij}$ equals the attribution of this token. The attribution penalty will lead the model to ignore the blacklisted tokens during classification.

$$target_{ij} = \begin{cases} a_{ij} & , \quad token_{ij} \notin Blacklist[y_i] \\ 0 & , \quad token_{ij} \in Blacklist[y_i] \end{cases}$$

3 Data and Experimental setup

3.1 Dataset

A total of 1271 electronic health records (EHRs) of successive patients who had undergone colonoscopies at Peking Union Medical College Hospital (PUMCH) and were clinically diagnosed with CD (n=875) or ITB (n=396) from January 2008 to November 2018 were included in this study. Research approval was obtained from Peking Union Medical College Hospital's Ethics Committee (approval no. S-K894).

All the patients had given informed consent. We separated 80% of the data into the training set and 20% of the data into the test set for training models and analysis.

The clinical diagnoses of CD were made via endoscopic results, medical history, pathological features, and treatment follow-up based on the Chinese consensus of IBD (2018) by IBD specialists in this hospital. The clinical diagnoses of ITB were confirmed by the presence of at least one criterion from the following: 1) positive acid-fast bacilli on histological examination, 2) positive *M. tuberculosis* culture, 3) radiologically or colonoscopically proven TB, and 4) full response to anti-TB therapy. Colonoscopies were performed with Olympus CF-Q260 or H260 colonoscopes and were conducted by well-trained endoscopists at PUMCH. Based on the well-established terminology used by endoscopists to describe colonoscopic images, we extracted descriptions of colonoscopic images of the patients' index colonoscopy in the form of free text. Clinically confirmed diagnoses extracted from the hospital information system (HIS) were used as labels.

Text sample		Diagnosis
Chinese description	钩拉法循腔插镜至回盲部。回盲部巨大不规则溃疡，周边结节样隆起，回盲瓣显示不清，局部活检 6 块，质硬，送病理及抗酸染色；余所见结肠、直肠粘膜光滑，血管纹理清晰，无充血、糜烂、溃疡及新生物。	CD

Translation	The colonoscope was introduced into the rectum and advanced to the terminal ileum using the Pull method. Large irregular ulcer(s) in the terminal ileum, with peripheral nodule(s). The ileocecal valve was not well seen. Biopsy of 6 pieces, which were firm, for pathological investigation and acid-fast stain test. Other findings: smooth colorectal mucosa, normal vascular pattern, no hemorrhage, no erosion and ulcer, no neoplasm.	
Chinese description	肠道准备欠佳循腔进镜至回肠末段约 15 cm,进镜顺利，末段回肠粘膜可见多发溃疡，形态欠规则，约 0.5–1.5 cm 大小，中心凹陷，周边粘膜肿胀隆起，表覆灰白苔，取活检共 3 块，质韧。回盲瓣呈唇形,阑尾开口看不清楚,所见全结肠、直肠粘膜光滑,血管纹理清,半月襞完整,未见糜烂、溃疡及新生物。	ITB
Translation	Poor bowel preparation. The colonoscope was introduced into the rectum and advanced to 15 cm from terminal ileum. Multiple cratered ulcers of 0.5–1.5 cm in the mucosa of terminal ileum, with peripheral edematous mucosa, covered by gray and white fur. Biopsy of 3 pieces, which were tough. Lip-shaped ileocecal valve. The vermix opening was not well seen.	

	Findings: smooth colorectal mucosa, normal vascular pattern, normal semilunar folds, no erosion and ulcer, no neoplasm.	

Table 1. Some examples of the collected and analyzed samples.

4 Results

Dataset	Model	CD			ITB			Overall Accuracy
		precision	recall	F1	Precision	recall	F1	
Standard	TextCNN	**0.92**	0.81	0.86	0.62	0.81	0.70	0.81
	PTM	0.87	0.86	0.87	0.75	0.77	0.76	0.83
	TextCNN(distill)	**0.92**	0.84	**0.88**	0.70	**0.83**	0.76	**0.84**
	TextCNN(Robust)	0.87	**0.87**	0.87	**0.77**	0.77	**0.77**	0.83
Noisy	TextCNN(distill)	0.60	0.61	0.61	0.33	0.32	0.32	0.50
	TextCNN(Robust)	0.82	0.83	0.87	0.83	0.71	0.69	0.70

Table 2. Classification results between CD and ITB

Table 2 displays the classification performance of the various models. The standard dataset refers to the original data. The distilled TextCNN gave the highest overall accuracy of 0.84 and the highest F1 score of CD of 0.88. By contrast, the standard TextCNN obtained the lowest overall accuracy of 0.81, which is 3 percentage points lower than that of distilled TextCNN. The Robust TextCNN gave the highest recall rate of 0.87 in CD and the highest F1 score of 0.77 in ITB. PTM did not show

advantages in any task. In the noisy dataset, the distilled TextCNN performed poorly, with an overall accuracy of 0.50. The Robust TextCNN thoroughly outperformed the distilled TextCNN that gave an overall accuracy of 0.70.

Model	TextCNN	TextCNN (distill)	PTM	TextCNN (Robust)
CD	The colonoscope was introduced into the rectum and advanced to advanced to **ulcer** **linear** anastomosis **findings**	The colonoscope was introduced into the rectum and advanced to **linear** advanced to **anastomosis** Crohn's disease findings	**ulcer** Crohn's disease **anastomosis** after treatment for Crohn's disease **linear ulcer** anus xx cm from anus	**linear** **anastomosis** **linear ulcer** **findings** **hyperemic** diffuse **cobblestone-like** **edematous** **stenosis**

	The colonoscope was introduced into the rectum The colonoscope was introduced advanced to ileum to **hyperemic** **erosion and ulcer** xx cm from anus to terminal ileum	**ulcer** The colonoscope was introduced into the rectum and advanced to xx cm from anus **linear ulcer** sigmoid colon The colonoscope was introduced into the rectum and advanced to	anus	xx cm from anus **moderately hyperemic** reexamination after treatment moderate bowel preparation localized **cobblestone-like** reexamination **small ulcer** sigmoid colon

	linear ulcer Crohn's disease sigmoid colon to ileum anus ulcer and **edematous stenosis**	advanced to ileum **edematous** to terminal ileum anus **hyperemic erosion ulcer** after treatment for Crohn's disease to ileum reexamination after treatment for Crohn's disease		

		descending colon and sigmoid colon diffuse		
ITB	**ileocecal valve** cecum pouch other other **polyps** **ileocecal valve deformity** ileocecal valve biopsy findings	**ileocecal valve** cecum other pouch other **polyps** findings biopsy ileocecal valve was The colonoscope was introduced	piece **protruding lesions** The colonoscope was introduced into the rectum and advanced to lesion **polyps-like protruding lesions** ileocecal valve	ileocecal valve ileocecal valve was biopsy **ileocecal valve deformity** findings cecum **polyps** other **round** other

	vermix opening The colonoscope was introduced into the rectum and advanced to **pouch** tissue submitted was pathological **deformity** normal The colonoscope was introduced	into the rectum and advanced to tissue submitted **ileocecal valve deformity** The colonoscope was introduced into the rectum and advanced to was **smooth folds** pathological fold **deformity**	lesion protruding lesions biopsy of 1 piece 1 piece other to ileocecal valve was to terminal ileum soft vermix opening **smooth erosion** to ileum	biopsy of 4 pieces biopsy of 4 pieces The colonoscope was introduced into the rectum and advanced to **remain opened** 4 pieces **fold** advanced to **scarring**

	into the rectum and advanced to biopsy from **smooth folds** biopsy of 4 pieces	**round** vermix opening **smooth folds** normal **remain** **opened**	soft biopsy of 4 pieces 3 pieces	

Table 3. Differential diagnosis features selected by the attribution analysis and their translation are listed. Features supported by clinical guidance are in bold. A table showing original features in Chinese is presented in appendix Table 1.

Table 3 shows the differential diagnosis features from each model. For CD, all of the classifiers gave *ulcer, linear*, and *anastomosis*. Notably, only robust TextCNN gave the feature *cobblestone-like* that was unique and set as a specific diagnostic feature in CD. Other features found by the classifiers included *hyperemic, edematous*, and *stenosis*. In addition, PTM gave much fewer features than the other three classifiers. For ITB, all four models gave similar features, including *ileocecal valve, polyp*, and *remain opened*. PTM model found *protruding lesions*, while the Robust TextCNN model found *round lesions*. All classifiers found uninformative features, which are those not highlighted in Table 3.

5 Discussion

Corresponding to the definition of trustworthy AI we proposed before, we discuss the contributions of this work from three aspects: accuracy, interpretability and robustness. For each aspect, we analyzed our contributions both from the perspective of the techniques and the perspective of clinical medicine.

5.1 Accuracy of differential diagnosis

For clinical medicine, this research provided a new possible approach for differentiating CD and ITB. Differential diagnosis of CD and ITB has long been a challenging and essential problem. Retrospective Chinese studies show that approximately 65% of CD patients have been misdiagnosed with ITB at least once[17]. At the same time, another study indicated that more than 40% of CD patients had received tentative anti-TB treatments due to ambiguous diagnoses. Traditional histologic or pathologic evidence, such as caseating granuloma or positive acid-fasting staining, was considered to be the gold standard with high specificity. However, these examinations are time-consuming and have a sensitivity lower than 50%. Thus, an immediate differential diagnosis with high sensitivity and specificity is valuable.

The four classifiers all achieved an overall accuracy above 80%, demonstrating that artificial intelligence can provide satisfactory results in clinical practice. This could help clinicians, particularly for inexperienced patients, to make a more accurate

diagnosis. The distilled TextCNN and robust TextCNN provided a balanced precision and recall rate, which was also crucial for clinical practice.

It is important to note that knowledge distillation leveraged the language knowledge of PTM and obtained a higher classification accuracy. As shown in Table 2, the overall accuracy of PTM was 2% higher than that of TextCNN. Then, distilled TextCNN achieved an even higher overall accuracy (by 1%) than PTM, and its F1 scores of both diseases also ranked first. In addition, the student model is significantly lighter than the teacher model. Therefore, knowledge distillation contributed to obtaining a better model while requiring less training and deployment resources. These advantages make diagnosis models more conducive to deployment and adoption.

5.2 Interpretability

Our previous study built a classifier for classifying CD and ITB using a convolutional neuron network (CNN)[18]. However, due to the low interpretability of CNN, the previous classifier could not explain the basis of the diagnosis to doctors, greatly limiting its clinical application. This research solved the previous problem. The front end can clearly show the classification result and the supporting and nonsupporting evidence, based on which clinicians can make further judgments.

5.3 Robustness

Debias training is an essential component of this system. First, it provides an effective method for doctors to customize the diagnosis model with their knowledge. In addition, debias training restricts the model from attributing the classification results to meaningless or unreasonable features in the blacklist and achieves significantly better results than the baseline model on the noisy dataset. Although we restricted the model from learning certain significant features in the standard dataset, it still reached the same level of accuracy as the models without debias training. The optimization of deep neural networks would by default exploit and extract any feature whose distribution in the training data correlates with the class label, and the extracted features are not guaranteed to be informative. Manually labeling the feature blacklist and penalizing it during training adds an additional regularization to the optimization of the neural network, forcing it to avoid unreasonable features in the blacklist to find features that truly differentiate and diagnose the two diseases.

The differential features for classification found by the classifiers were highly consistent with the guidelines. We noticed that Robust TextCNN provided more specific features, such as cobblestone appearance, while TextCNN and distilled TextCNN tended to offer more general features. This may occur because patients with these specific features comprise only a small portion of the total data set. TextCNN and distilled TextCNN tended to ignore these features due to the small sample size and corresponding low statistical power. However, Robust TextCNN

gave these specific features, most likely due to the penalty coefficient of the general features. Therefore, in the noisy dataset, Robust TextCNN strongly outperformed distilled TextCNN. A further discussion of Robust TextCNN is given below. In summary, clinicians can use diagnostic evidence from different classifiers to support their judgment.

However, we should note that some patients may not be distinguished purely by endoscopy and need further examinations due to the similarity of the endoscopic results of CD and ITB. Therefore, additional clinical and biological research on CD and ITB may be conducted to evaluate whether feature extraction by AI can help improve the upper limit value of the accuracy while differentiating CD and ITB.

5.4 Limitations

Our current work is limited in that it only uses the description text of endoscopy reports. It should be noted that a loss of information can occur when inexperienced clinicians describe the endoscopic findings, and there are also CD and ITB cases that are not distinguishable by endoscopy. Therefore, a combination of other clinical lab examinations and the text model can potentially improve the model's classification capability and requires further research. Additionally, language patterns may differ across institutions. Although the extracted differential features appear consistent with clinical experience and guidelines, the portability of the text model at different institutions requires further testing.

6 Conclusion

In this work, we developed a differential diagnosis application using state-of-the-art natural language processing for differentiating between CD and ITB, focusing on the accuracy, interpretability, and robustness aspects of a trustworthy AI. The resulting classifier performed well, and the extracted differential features that met statistical significance conformed with clinical guidelines, proving the effectiveness of our human-in-circle workflow.

Declaration

Ethics approval and consent to participate

Research approval was obtained from Peking Union Medical College Hospital's Ethics Committee (approval no. S-K894). All the patients had given informed consent.

Consent to publish

All the patients had given informed consent.

Availability of data and materials

Clinical data cannot be published due to ethical consideration. Please contact the corresponding author for potential access of clinical data.

Competing interests

None declared.

Funding

This work was supported by the Natural Science Foundation of Beijing Municipality (Grant No. Z190024), the National Natural Science Foundation of China (Grant No. 11801301), the Tsinghua University Initiative Scientific Research Program, the CAMS Innovation Fund for Medical Sciences (CIFMS), 2020-I2M-C&T-B-005 and the Beijing Municipal Natural Science Foundation (7212078).

Author Contributions

Sheng Yu and Yue Li contributed to the conception of this study; Keming Lu and Yuanren Tong designed the model and performed experiments; Keming Lu, Yuanren Tong and Si Yu significantly contributed to result analysis and manuscript preparation; Yucong Lin helped perform the analysis with constructive discussions; Yingyun Yang and Hui Xu helped collect and preprocess data. Each author agreed this version of manuscript to be submitted.

Acknowledgments

We sincerely express our appreciate to all patients who agreed to participate in this research.

References

[1] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. arXiv preprint arXiv:2004.02015, 2020.

[2] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. arXiv preprint arXiv:1906.08101, 2019.

[3] M Graziani, V Andrearczyk, S Marchand-Maillet, and H Müller. Concept attribution: Explaining cnn decisions tophysicians. Computers in biology and medicine, 123:103865, 2020.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network.arXiv preprint arXiv:1503.02531,2015.

[5] Shaoping Hu, Yuan Gao, Zhangming Niu, Yinghui Jiang, Lao Li, Xianglu Xiao, Minhao Wang, Evandro Fei Fang, WadeMenpes-Smith, Jun Xia, et al. Weakly supervised deep learning for covid-19 infection detection and classification from ct images.IEEE Access, 8:118869–118883, 2020.

[6] Yoon Kim. Convolutional neural networks for sentence classification. corr abs/1408.5882 (2014).arXiv preprintarXiv:1408.5882, 2014.

[7] Simon Meyer Lauritsen, Mads Kristensen, Mathias Vassard Olsen, Morten Skaarup Larsen, Katrine Meyer Lauritsen,Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. Explainable artificial intelligence model to predict acute critical illness from electronic health records.Nature communications, 11(1):1–11, 2020.

[8] Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification. arXiv preprintarXiv:1906.08286, 2019.

[9] Kristina Preuer, Günter Klambauer, Friedrich Rippmann, Sepp Hochreiter, and Thomas Unterthiner. Interpretable deep learning in drug discovery. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pages 331–345. Springer,2019.

[10] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology, 126(4):552–564, 2019.

[11] Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. arXiv preprintarXiv:1806.05337, 2018.

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In International Conference on Machine Learning, pages 3319–3328. PMLR, 2017.

[13] Jiao, Zhenyu, Shuqi Sun, and Ke Sun. "Chinese lexical analysis with deep bi-gru-crf network." *arXiv preprint arXiv:1807.01882* (2018).

[14] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[15] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

[16] He, Y., Zhu, Z., Chen, Y., Chen, F., Wang, Y., Ouyang, C., ... & Chen, M. (2019). Development and validation of a novel diagnostic Nomogram to differentiate between intestinal tuberculosis and Crohn's disease: a 6-year prospective multicenter study. *Official journal of the American College of Gastroenterology| ACG*, *114*(3), 490-499.

[17] Lee, Y. J., Yang, S. K., Byeon, J. S., Myung, S. J., Chang, H. S., Hong, S. S., ... & Yu, C. S. (2006). Analysis of colonoscopic findings in the differential diagnosis between intestinal tuberculosis and Crohn's disease. *Endoscopy*, *38*(06), 592-597.

[18] Tong, Y., Lu, K., Yang, Y., Li, J., Lin, Y., Wu, D., ... & Qian, J. (2020). Can natural language processing help differentiate inflammatory intestinal diseases in China? Models applying random forest and convolutional neural network approaches. *BMC medical informatics and decision making*, *20*(1), 1-9.

Figure legend

Figure 1. The workflow of building a text-based trustworthy diagnosis model.

Figure 2. A case demonstration of hierarchical feature set extraction: words or characters in the sentence with positive attribution scores are highlighted with a green background. The extraction process constructs an N-gram hierarchical feature set from bottom (word or character level) to top.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- appendix.docx