

RESEARCH

Open Access



Exploring the performance and explainability of fine-tuned BERT models for neuroradiology protocol assignment

Salmonn Talebi^{1†}, Elizabeth Tong^{2†}, Anna Li², Ghiam Yamin², Greg Zaharchuk² and Mohammad R. K. Mofrad^{1*}

Abstract

Background Deep learning has demonstrated significant advancements across various domains. However, its implementation in specialized areas, such as medical settings, remains approached with caution. In these high-stake environments, understanding the model's decision-making process is critical. This study assesses the performance of different pretrained Bidirectional Encoder Representations from Transformers (BERT) models and delves into understanding its decision-making within the context of medical image protocol assignment.

Methods Four different pre-trained BERT models (BERT, BioBERT, ClinicalBERT, RoBERTa) were fine-tuned for the medical image protocol classification task. Word importance was measured by attributing the classification output to every word using a gradient-based method. Subsequently, a trained radiologist reviewed the resulting word importance scores to assess the model's decision-making process relative to human reasoning.

Results The BERT model came close to human performance on our test set. The BERT model successfully identified relevant words indicative of the target protocol. Analysis of important words in misclassifications revealed potential systematic errors in the model.

Conclusions The BERT model shows promise in medical image protocol assignment by reaching near human level performance and identifying key words effectively. The detection of systematic errors paves the way for further refinements to enhance its safety and utility in clinical settings.

Keywords Healthcare, Machine learning, Interpretability, Explanations, BERT

Background

Machine learning systems are being rapidly adopted for many applications including high-stakes settings such as medical applications [1–5]. Recent progress with self-attention techniques, and specifically Transformers, have

dominated the field of text processing and classification tasks. Large pretrained Transformers have outperformed humans on language understanding tasks such as SuperGLUE [6], a suite of challenging NLP tasks designed to evaluate a system's proficiency in understanding and generating human language. These tasks encompass a range of complex language scenarios, from question answering to sentiment analysis. However, despite these advancements, many specialized text analysis tasks do not make use of modern machine learning methods [7]. It remains questionable how well existing pretrained models will transfer to large, specialized texts.

[†]Salmonn Talebi and Elizabeth Tong contributed equally to this work.

*Correspondence:

Mohammad R. K. Mofrad
mofrad@berkeley.edu

¹ University of California, 208A Stanley Hall #1762, Berkeley, CA 94720-1762, USA

² Stanford University, Stanford, CA, USA



In high-stakes fields like medicine, law, and security, where specialized human expertise is crucial, the effective deployment of machine learning algorithms hinges on not only achieving human-level performance but also providing clear, trustworthy explanations to the user [8, 9]. Furthermore, model explainability is being driven by laws and regulations which state that decisions from machine learning algorithms must provide information about the logic behind those decisions [10]. In fact, the lack of explainability of ML models often plagues medical artificial intelligence (AI) [11]. For these reasons, in high-stake settings, explainability should be a priority for researchers.

In this study, we focus on the specialized task of identifying medical imaging protocols within text descriptions. Clinicians often order radiologic studies, such as magnetic resonance imaging (MRI) or computed tomography (CT), to help answer clinical questions and guide treatment decisions [12–14]. Typically, when a physician orders an imaging study, he/she will provide a description with the patient's symptoms and history, which radiologists then review to recommend the most suitable radiologic protocol.

Traditionally, protocol assignment to each radiologic order is done manually by the radiologists or radiology technologists. This can incur substantial costs to the healthcare system. This tedious task may take up to at least 6% of the radiologists' time [15]. With increasing radiology orders, an automated process with high throughput and accuracy is desirable to ensure patient care and to avoid radiologists' burnout. However, given the high stakes of medical tasks, machine learning models must be evaluated for any systematic biases or errors before they can be trusted by clinicians and patients [16]. In order for these models to be used in practice they need to provide valid explanations for how the decisions are made.

Deep learning machines can perform this protocol classification. Previous work has been done using machine learning techniques such as SVM, Random Forests, and Gradient Boosted Machine [17]. A deep neural network approach demonstrated a slight boost over k-Nearest Neighbors (KNN) and random forests methods [18]. However, these models are limited by the size of the model and the use of classical word embeddings which do not provide deep contextual word embeddings [19]. Newer models, such as the bidirectional recurrent neural networks (RNN) and Transformers can improve text representation to be sensitive to its local context in a sentence and optimized for specific tasks by using a self-attention mechanism to help embed the context of each word [20]. Large language

models such as BERT (Bidirectional Encoder Representations from Transformers) [21] and ELMo (Embeddings from Language Model) [22] have been shown to provide substantial performance improvements for language modeling and text classification.

In this study, we designed machines to perform a protocol classification task using Transformer-based models. We adapted several large pre-trained BERT-based language models to classify neuroradiologic orders. The ML models will learn the medical language used to indicate a neuroradiologic order and assign the best protocol accordingly. This is a complex task because the models need to understand language in the context of human anatomy and pathology from a short vignette. We hypothesize that the use of context-dependent token embeddings will substantially improve medical text classification and model interpretation compared to conventional ML models. A BERT model that was pre-trained on biomedical literature [23] and another pre-trained on clinical text [24] will be included for comparison. This will provide the best contextual token embeddings for the model to understand the physicians' notes. The performances of these BERT-based Transformer models were compared to several machine learning models.

In addition, we evaluate the model's ability to provide explanations of its decision based on word importance. A trustworthy algorithm should be able to demonstrate it is making complex decisions using similar rational to a human. For this application, explanation is increasingly complex because the model will need to understand language in the context of human anatomy and physiology. Figure 1 illustrates a proposed system in which physician notes are fed as input to a model, which then outputs an imaging protocol along with an explanation for its decision-making process.

The main contributions of this study are as follows:

- We fine-tune different pre-trained BERT model using a medical dataset of medical imaging protocol text, and demonstrate that it achieves state-of-the-art performance compared to previous studies.
- We employ a gradient-based method called integrated gradients to quantify the contribution that each word in the input text makes to the model's decision.
- We validate the model's word importance claims using a technique called erasure.
- We analyze the model's mistakes using word importance and identify systematic errors that may pose potential safety risks and need to be addressed before the model can be safely deployed in a clinical setting.

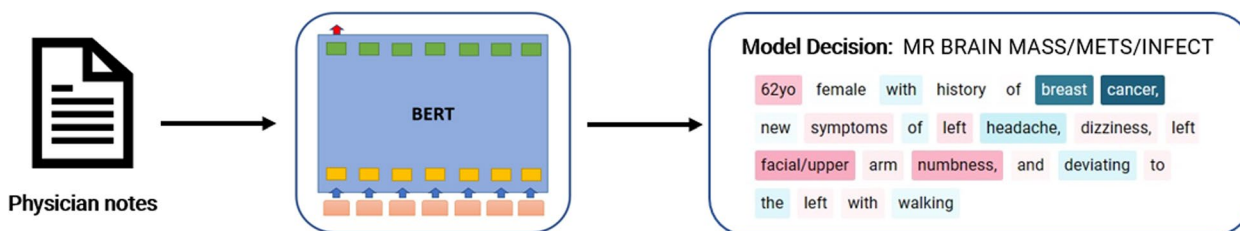


Fig 1 A proposed system in which physician notes are used as input to a model. The output of the model is an imaging protocol, with color coding to denote the significance of terms in influencing the model’s decision: red signifies words that negatively influence the prediction, blue denotes the most important words that positively influence the prediction, and white indicates a neutral influence. This system aims to provide a more efficient and accurate method for determining appropriate imaging protocols, while also offering insight into the decision-making process of the model. By incorporating an explainability component, the proposed system has the potential to enhance trust and understanding in the use of machine learning for medical image protocol assignment

Table 1 The 10 most commonly assigned protocols and their frequencies

Protocol name	Number of entries
MR brain demyelinating	4,076
MR brain mass/metastases/infection	32,587
MR moya-moya with Diamox	1,765
MR nasopharynx oropharynx	3,945
MR orbit sinus face	4,289
MR seizure	3,476
MR sella	5,297
MR skull base	4,390
MR stroke	23,704
MR vascular malformation/hemorrhage/trauma	4,523
Total	88,052

Data

In order to train a specialized model for medical text classification, we have compiled a new large-scale dataset for image protocol review. This dataset consists of order entries and assigned protocols for magnetic resonance (MR) neuroradiology studies that were conducted at our institution between June 2018 and July 2021. Each row in the dataset represents a single radiology order and includes the ‘reason for exam’, patient age and gender, and the protocol assigned by the radiologist.

We have excluded orders for spine imaging from this study, as the assigned protocol typically reflects the specific segment of the spine indicated in the order. From the original dataset of 119,093 rows, we removed the most common protocol, ‘routine brain’, as it can be used for a wide range of indications and serves as the default protocol at our institution. The remaining dataset was narrowed down to the 10 most common protocols (Table 1).

To ensure the accuracy and quality of the data, we performed a thorough review by an experienced radiologist (ET) with 10 years of experience. We also applied standard text preprocessing techniques such as handling of

missing outputs, and expansion of acronyms, to further clean and organize the data. Furthermore, in medical documentation, the use of standardized terminology and phrases is common practice, which could lead to similar entries in our dataset. To address this, we have meticulously removed any duplicate entries prior to the randomization process to prevent overfitting. The remaining similar entries accurately reflect real-world medical reporting practices where standardized language is prevalent. Our methodology ensures that the model’s training and validation are as close to the clinical reality as possible. The final dataset includes 88,000 recorded notes with expert-annotated imaging protocols.

Methods

This retrospective study was conducted with the approval of the Stanford Institutional Review Board (IRB) and under a waiver of informed consent. The study was approved for collaboration between Stanford University and the University of California, Berkeley.

BERT fine tuning

We approach the problem of text classification as predicting the class that corresponds to a given input text. In our dataset, we have 10 possible classes that can be predicted. To achieve this, we fine-tuned four pre-trained language models using the HuggingFace Transformers library: BERT, RoBERTa, BioBERT, and ClinicalBERT [25].

BERT (Bidirectional Encoder Representations from Transformers) is a machine learning framework for natural language processing (NLP) that serves as the foundation of these models. BERT was pre-trained on a dataset of English text consisting of books, articles, and websites, including Wikipedia. The model was trained using an unsupervised pre-training method, where the model is trained to predict missing words in a sentence or a sequence of text (also known as “masked language modeling”).

Before being processed by the encoder, the input data is transformed by passing it through three embedding layers: a token embedding layer, a segment embedding layer, and a position embedding layer. In the token embedding layer, the input sentences are tokenized. Each token is then transformed into a fixed-dimensional vector representation (e.g., a 768-dimensional vector). Special classification [CLS] and separator [SEP] tokens are also inserted at the beginning and end of the tokenized sentence to serve as input representations and sentence separators for the classification task. The [CLS] token in the last hidden state of BERT contains the embedding of the entire input and is used for classification (Fig. 2)

While all models share this core architecture, they differ in their pre-training data, which tunes them for specific domains: RoBERTa was optimized on extended data for improved performance; BioBERT was further pre-trained on biomedical literature; and ClinicalBERT was further fine-tuned on clinical text. Our contribution involves adapting and integrating these models for accurately predicting neuroradiology protocol assignment from physician notes.

Resemblant to the clinical setting, the number in each protocol is not evenly distributed (Table 1). More than half of the imaging protocol entries belong to two of the classes. To mitigate this imbalance we up sample the remaining 8 imaging protocols so that the dataset is approximately balanced between all 10 classes of imaging protocols. Before performing the up sampling, the data is randomly split into a train, validation and test sets. We have 70% of the protocols make up the train set, 20% make up the validation set, and 10% make up the

test set. The validation set was used to perform a hyper-parameter grid search. The learning rate was tuned from the range of 1×10^{-4} to 1×10^{-6} , using a step size of 2×10^{-5} . During our experiments we found the model would converge after 10 epochs and training for any longer would degrade performance. The model is trained using a single A6000 GPU.

Model baseline

In order to establish a baseline and compare the performance of our fine-tuned BERT model against traditional machine learning methods, we conducted experiments using several well-known algorithms, namely Random Forest (RF), XGBoost, and Deep Neural Networks (DNN). These algorithms have been used in previous studies for medical imaging protocol assignment and provide a benchmark to evaluate the effectiveness of our approach.

For the RF, XGBoost, and KNN models, we employed TF-IDF vectorization to transform the input text into numerical feature vectors. For the DNN model we integrated pre-trained GloVe embeddings [26]. To implement and evaluate the traditional machine learning methods, we used popular and widely adopted Python libraries for each of the algorithms. For RF, and XGBoost, we utilized the scikit-learn library. For the DNN, we employed Keras for building a 1D Convolutional Neural Network (CNN) classifier.

Word importance

For the purposes of this study, we use the concept of "word importance" as a means of interpreting the model.

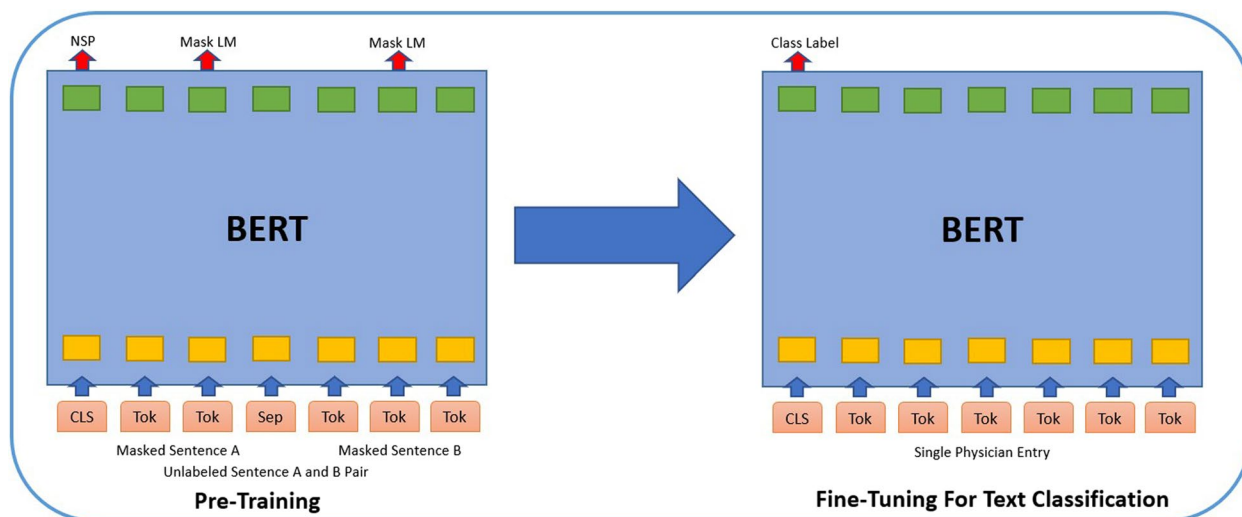


Fig 2 (Left) Original pre-trained BERT that is trained to perform 'next sentence prediction (NSP)' and 'masked-language modeling (MLM)'. Special classification [CLS] and separator [SEP] tokens are inserted into the input to facilitate learning. (Right) BERT is fine-tuned for this classification task using labeled data from physician entries. The output is a class label corresponding to the assigned protocol

Table 2 A comparison of imaging protocol F1 scores

Protocol Name	BERT	RoBERTa	ClinicalBert	BioBERT	DNN	XGBoost	RF	KNN
MR BRAIN DEMYELINATING	0.92	0.92	0.95	0.94	0.91	0.92	0.90	0.75
MR BRAIN MASS/METS/INFECT	0.85	0.85	0.86	0.87	0.77	0.71	0.66	0.59
MR BRAIN MOYA-MOYA DIAMOX	0.96	0.96	0.96	0.98	0.96	0.98	0.97	0.90
MR NASOPHARYNX OROPHARYNX	0.89	0.94	0.93	0.97	0.92	0.93	0.91	0.75
MR ORBIT SINUS FACE	0.85	0.84	0.89	0.88	0.83	0.81	0.75	0.68
MR BRAIN SEIZURE	0.95	0.95	0.96	0.96	0.77	0.78	0.68	0.66
MR SELLA	0.96	0.96	0.97	0.97	0.94	0.94	0.89	0.74
MR SKULL BASE	0.82	0.82	0.89	0.96	0.79	0.74	0.64	0.61
MR STROKE	0.84	0.84	0.88	0.88	0.83	0.79	0.73	0.72
MR VASCULAR MALFORMATION	0.87	0.84	0.89	0.88	0.84	0.83	0.75	0.65
Weighted Average	0.89	0.89	0.91	0.92	0.85	0.84	0.77	0.70

Word importance quantifies the contribution that each word in the input text makes to the model's prediction. To calculate word importance, we utilize a gradient-based method called integrated gradients [27, 28].

Validating word importance

The assumption to use heat-maps of attribution values over the inputs as explanations is particularly popular for natural language processing. To test the validity of these explanations, "stress tests" can be designed using a method called erasure, where the most or least important parts of the input, as indicated by the explanation, are removed and the model's prediction is observed for changes [29]. Specifically, we erase the most (or least) important word from the input sentence and measure the resulting model accuracy.

Aggregating word attribution

We aggregate the word attributions across multiple texts for each imaging protocol. Integrated gradient assigns attribution scores to each prediction made on a text segment that is a maximum of 512 sub-words long. We calculate the top 5 words for each imaging protocol by taking the average attribution value for each word across all text for a given imaging protocol, and select the top words as those with the highest average attribution value. We further filter out words that appear in less than 3 texts. A trained radiologist assigned a measure of word importance across all text for a given imaging protocol. This measure was based on a numerical score, with a value of 1 indicating a strong influence on the radiologist's decision, 0.5 indicating a slight influence, and 0 indicating a neutral influence. For each word, the human word importance score was determined as the average of all word scores across a single image protocol class. These methods were employed to generate lists of the most

influential words for each imaging protocol, utilizing both the BERT model and the judgments of the trained radiologist.

Results

The results of our fine-tuning experiment on the BERT model are shown in Table 2. The model's performance was evaluated using three metrics: precision, recall, and F1 score. The F1 score is a measure of the model's accuracy, taking into account both the precision and recall of the model. We found that the BERT, RoBERTa, ClinicalBERT, and BioBERT models had an F1 score of 0.89, 0.89, 0.91, and 0.92 respectively. This represents a significant improvement over the results of previous studies using other machine learning methods. One such study using deep neural network, random forest algorithm, and k-nearest neighbors (kNN) achieved a F1 scores of only 0.83, 0.81 and 0.76 respectively [18].

For our dataset, we also measured the weighted average F1 scores of the traditional machine learning models: XGBoost achieved an F1 score of 0.84, RF scored 0.77, KNN obtained 0.70, and the DNN yielded an F1 score of 0.85. These results are comparable to the performance of existing studies. Overall, the results of our experiment demonstrate the superior performance of the pre-trained BERT models compared to non-Transformer based approaches. The BERT models were able to achieve a higher level of accuracy, as indicated by the higher F1 score, and outperformed other methods in this task.

The performance of the BioBERT model was compared with a human readers. The number of errors and the accuracy in each category are tabulated in Table 3. The accuracy is tied in the 'MR moya-moya with Diamox' and 'MR seizure' categories. Otherwise, the neuroradiologist achieved higher accuracy than the BioBERT model in the remaining 8 categories

Table 3 Performance results of the biobert model compared with neuroradiologists. Human outperform the model in all but 2 categories

	Number of entries	Number of BioBERT errors	Accuracy %	Number of Human errors	Accuracy %
MR brain demyelinating	395	19	0.95	13	0.97
MR brain mass/mets/infection	457	46	0.90	9	0.98
MR moya-moya with Diamox	184	4	0.98	4	0.98
MR nasopharynx oropharynx	384	11	0.97	5	0.99
MR orbit sinus face	426	60	0.86	15	0.96
MR seizure	355	4	0.99	4	0.99
MR sella	521	13	0.98	1	0.99
MR skull base	443	44	0.90	20	0.95
MR stroke	497	55	0.89	17	0.97
MR vascular malformation/hemorrhage/trauma	433	47	0.88	25	0.94
Weighted Average			0.93		0.97

Word importance

The attribution scores assigned to individual words by the integrated gradients are intended to reflect the influence of those words on the model’s decisions.

To verify the validity of these attribution scores, we conducted a “stress test” using a technique called erasure. This involved systematically removing the most and least important words from the input text and measuring the resulting impact on the performance of the BERT models. The results of this stress test are shown in Fig. 3. We can see that the removal of the least important words had a relatively small effect on the model’s performance, causing a decline in the F1 score from 0.89 to 0.86. In contrast, the removal of the most important words had a much more significant impact, with the F1

score dropping sharply from 0.89 to 0.62 when the top-most important words were removed. Each subsequent removal of the most important words also resulted in a decremental drop in the F1 score. The stress test was also performed on RoBERTa, ClinicalBERT and BioBERT yielding similar results.

These results provide strong evidence that the attribution scores generated by the integrated gradients method are valid, as they accurately reflect the influence of each word on the model’s performance. The stress test demonstrates that the most important words have a substantial impact on the model’s ability to make accurate predictions, and that the words with the highest attribution scores are particularly influential in the model’s decision making process. We aggregate word attribution scores

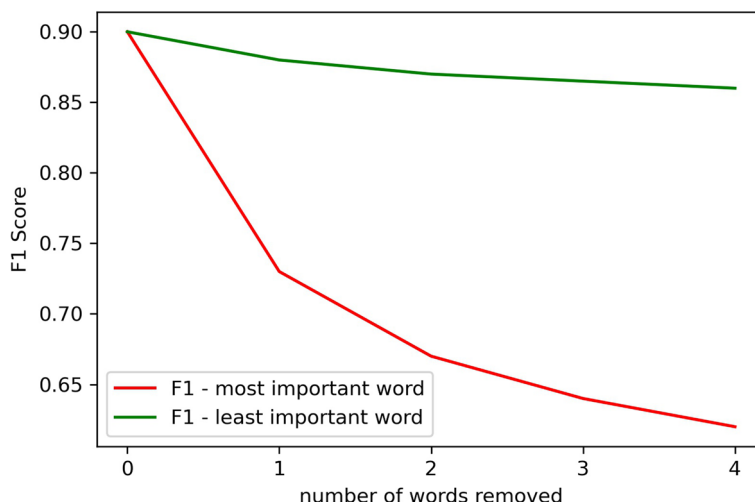














Fig 3 Model performance after step-wise removal of the 4 most important words and the 4 least important words from the text prompt. The results show that the least important words are less likely to degrade model performance while the most important words substantially degrade the performance

a) BERT

MR Brain MASS/METS			MR Brain Seizure			MR Stroke		
 & 			 & 			 & 		
mets	intracranial	post	seizure	hippocampus	winter	stroke	mental	pit
cancer	cyberknife	stereo	epilepsy	temporal	onset	transient	mra	headache
tumor	brain	treatment	visualase	cortical	disorder	mca	cva	history
lung	lesions	rule	lobe	dysplasia	protocol	vertigo	facial	mri
meningioma	lymphoma	date	confusion	coronal	axial	defuse	weakness	memory

b) BioBERT













MR Brain MASS/METS			MR Brain Seizure			MR Stroke		
 & 			 & 			 & 		
mets	cyberknife	post	seizure	hippocampus	winter	stroke	mental	pit
cancer	brain	mass	epilepsy	temporal	focus	transient	vertigo	ischemic
tumor	lesions	treatment	visualase	cortical	onset	weakness	cva	headache
lung	lymphoma	rule	lobe	dysplasia	disorder	mra	facial	arm
intracranial	astrocytoma	staging	confusion	coronal	brain	mca	weakness	neck

Fig 4 Top 5 words where human (trained radiologist) and (a) BERT or (b) BioBERT agree or disagree for 3 selected protocols. Human & robot are words both human and model agree are important. Human only are words with high human importance but low model importance. Robot only are words with high model importance but low human importance

for each image protocol and investigate the difference in the word importance ranks of BERT, and those of a radiologist (Fig. 4a). Both human (trained radiologist) and the BERT models picked the words most frequently mentioned in the indications for brain mass workup. There was no difference in the top 5 aggregate words between BERT and RoBERTa, and only minimal differences were observed between BERT and ClinicalBERT, as well as between BERT and BioBERT (Fig. 4b).

Meningioma is the most common type of brain tumor and lung cancer is the most common cause of brain metastases. Mets is a very commonly used shorthand for metastases. Both human and BERT picked up words suggesting a history of treatment for brain tumors, human picked ‘cyberknife’, while BERT picked ‘post, stereo, treatment’. ‘Rule’ and ‘date’ favored by BERT are most likely due to bias.

Seizure and epilepsy (a condition with prolonged or repetitive seizures) are obviously important for the seizure protocol, both human and BERT agreed. They also consider ‘visualase’, which is an ablation technique for treating seizures, important. BERT did not recognize the specific anatomic structures (hippocampus, temporal lobe) and specialized medical term that are considered important for humans. Instead BERT was biased by some non-specific words.

The top 5 words in agreement for stroke protocol are indeed critical, specific, and frequently used. Again BERT was biased by a few generic words, and failed to recognize words that describe the symptoms of stroke or the medical acronym for stroke (‘cva’).

Furthermore we examine individual texts and their word attribution values to assess the model’s understanding of language in the context of human anatomy and pathology. Figure 5 presents a physician’s text alongside the model’s corresponding word attribution values. In the first example, the model places emphasis on the patient’s history of breast cancer and a headache. In older patients, headaches can often indicate the presence of a brain tumor, and cancer can spread from the breast to the brain, leading to brain metastasis. Despite the presence of symptoms such as dizziness, facial, and numbness, which suggest the possibility of a stroke, the model de-emphasizes these words and correctly determines that brain metastasis is the most likely cause, given the patient’s history of breast cancer and a headache. In the second example, we see a case where the model makes an incorrect decision. The mention of possible edema on a computerized tomography scan suggests the possibility of a brain tumor. Additionally, the model ignores the age of the patient, which is relevant because for patients over the age of 50, seizures are often caused by brain tumors. While an MRI to diagnose brain seizure is plausible, the reasons described indicate that an MRI to diagnose brain metastasis is generally more likely in this case.

Error analysis

In order to understand the errors made by our fine-tuned BERT model on the test set, we conducted an analysis of the model’s explanations and looked for any systematic patterns in the mistakes. Our analysis identified four broad categories of errors: (1) multifarious choices, (2)

Predicted Label	True Label	Indication For Exam
MR Brain METS	MR Brain METS	62yo female with history of breast cancer, new symptoms of left headache, dizziness, left facial/upper arm numbness, and deviating to the left with walking
MR Brain Seizure	MR Brain METS	59 yo w left posterior headache possible seizure, concern for edema on computer tomography. Brain tumor at age 18. epilepsy w seizure and possible edema on computer tomography. gender male
MR NASOPHARYNX OROPHARYNX	MR NASOPHARYNX OROPHARYNX	70 Year-old male with a 50 pack-year smoking history (quit 11/2016) and a T3 N2 squamous cell carcinoma of the left upper lobe treated with chemoradiation therapy to 66 Gy in 30 fractions with concurrent cisplatin and etoposide completed on 6/19/17

Fig 5 Selected samples from the dataset with color coded word importance. Red signifies words that negatively influence the prediction, blue denotes the most important words that positively influence the prediction, and white indicates a neutral influence. The indication for the exam is provided by the ordering physician, which briefly summarizes the symptoms, relevant medical history, and the medical questions. The 'true label' is the protocol, assigned manually by a trained radiologist, that is most suitable for the indication. The 'predicted label' is the protocol predicted by the AI model

age-related results, (3) ambiguous entries, and (4) flagrant errors.

The most common type of mistake occurred when the clinical question was too complex or broad, with multiple clinical questions, regions of interest, or complex medical histories. In these cases, there may be multiple valid imaging protocols, and the model struggled to select the most appropriate one. This accounted for 52% of the errors in the test set.

Errors in the second category, age-related results, occurred when the model failed to consider the age of the patient in its prediction. For example, the best protocol for a patient with intracranial hemorrhage may vary depending on their age group. This category accounted for 15% of the errors in the test set. Errors in the third category, ambiguous entries, occurred when the model was unable to make a prediction due to ambiguous or esoteric language in the input text. This could include stems that were too rare or cryptic, or protocols that could not be designated to ambiguous stems. This category accounted for 5% of the errors in the test set.

Finally, flagrant errors, the fourth category, occurred when the model made a wrong prediction or the order of word importance did not make sense for the prediction. This category accounted for 28% of the errors in the test set. A visual breakdown of these mistakes is provided in Fig. 6.

Overall, the largest issue for the model was its difficulty in understanding the hierarchical ordering of protocols. This accounted for 52% of the errors in the test set, and will require further work to address before the model can be used in a clinical setting. Another issue was the

model's partial capture of important regions of the input text, which accounted for 15% of the errors. This may be due to biases or limitations in the training data, and will also require further work to address. By understanding the patterns of errors made by the model, we can begin to identify areas for improvement and fine-tune the model to achieve even better performance.

Discussion

Protocols are a crucial task for radiologists to ensure that the appropriate sequences are acquired in response to clinical questions. However, manual protocols can be time-consuming, disruptive, and prone to errors. In recent years, the volume of radiologic orders has increased, making protocols an increasingly costly burden. To address these challenges, we utilized a large pre-trained language model that was fine-tuned by training it with a large dataset of radiologic orders. This allowed the model to learn medical terminology and accurately process orders, which frequently contain typos, acronyms, and grammatical errors, and are often written in shorthand using specialized medical terminology.

In response to the increasing demand for 'explainable AI', we investigated the decision-making process of our model. We evaluated the model's ability to provide explanations of its decision based on 'word importance'. Model explanation techniques were applied to estimate the importance of each word within the text of each radiologic order. This allowed us to delve into the model's decision-making process and determine whether it was making correct predictions for the right reasons, as

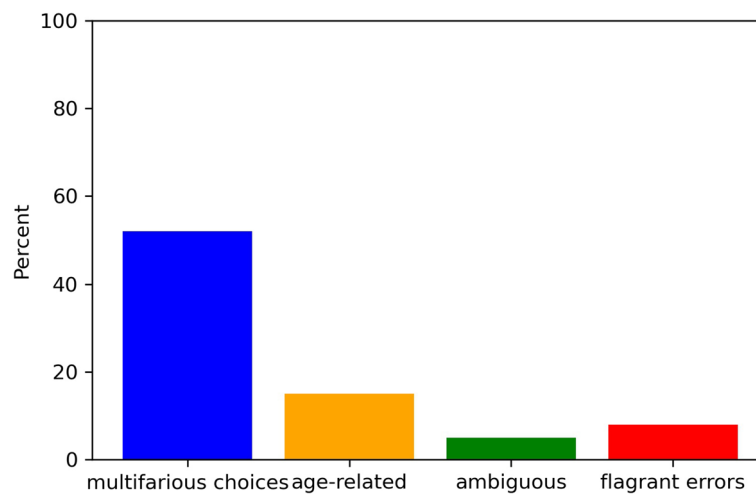


Fig 6 The bar plot decomposes the mistakes into four categories: multifarious choices, age-related, ambiguous text, and flagrant errors

well as to identify the root causes of any mistakes. Our results indicate that the BERT based models are able to identify relevant words that are highly indicative of the target protocol.

Our error analysis revealed that the model struggled most with understanding complex indications involving multiple clinical questions, leading to incorrect protocol selection in some cases. For example, the model may have difficulty distinguishing between protocols for a patient with acute neurologic deficits after brain tumor resection, as it may not fully comprehend the hierarchical ordering of protocols. Furthermore, we identified that approximately 15% of the model's mistakes were due to insufficient capture of important regions of the input text. This could be due to various factors such as bias in the training data or limited examples of certain edge cases.

We recognize the limitations of deep learning explainability tools like integrated gradients, which, although effective for text-based models such as BERT, may not universally apply to other data types or models. Other methods, such as LIME [30] or attention weights offer alternative ways to interpret model decisions.

It's essential to recognize that current explainability techniques are, at best, approximations. Recent studies have shown that these methods provide our best guess at explaining model decisions, but they are not always entirely accurate or intuitive [31]. The field has yet to discover a universally intuitive and completely reliable explainability technique. This uncertainty underscores the importance of ongoing research and critical evaluation of explainable AI models, especially in high-stakes environments like healthcare.

Furthermore, the emergence of advanced generative models like GPT-4 introduces new challenges and

opportunities for large language models [32]. These models excel in text generation with their nuanced language understanding and production. Yet, their internal complexity raises significant explainability challenges. Some recent studies have begun to clarify their functionality [33], but there is still much to uncover. Future studies will extend this research to compare the interpretability and explainability of pretrained BERT models with such generative AI models.

Lastly, in future studies, we aim to delve deeper into the granularity of medical imaging by extending our model's capabilities beyond protocol-level classification to the labelling of individual imaging acquisitions. Given that certain acquisitions may be common across multiple protocols, this refined approach could unveil how well the model discerns the nuanced differences and similarities between them. Such an investigation would not only enhance the model's precision in predicting appropriate acquisitions but also provide a more detailed understanding of its alignment with the intricate requirements of each protocol.

Limitation

There are several limitations to consider in the context of this study. First, our dataset comprised of neuroradiologic orders from a single center, and thus may be limited in its representation of the racial, social, and ethnic diversity of other regions. Validation with datasets from different institutions is necessary to more accurately compare the model's performance. Additionally, we limited the number of protocols to the ten most commonly used protocols in this study, which may not fully capture the breadth of protocols used in clinical practice. The data was collected from routine clinical work, which means that protocols were assigned by multiple radiologists

with varying levels of experience, potentially leading to inter-operator variability. While the dataset is relatively large at over 80,000 entries, it is possible that additional data could further improve model performance.

Additionally, it is important to note that there may be significant variations in the importance of certain words when considering the perspectives of different radiologists. In this study, we were constrained to a single radiologist when evaluating word-level agreement with BERT. However, in future studies, it would be beneficial to evaluate word importance from the perspectives of a diverse group of radiologists to achieve more robust results.

Related work

Previous work has been done using classification models to predict imaging protocol from a physician's notes using machine learning techniques such as SVM, Random Forests, and Gradient Boosted Machine [34]. More recently, a deep neural network approach was used to automate radiological protocols which showed a slight boost over kNN and random forests. However, these models are limited by the size of the model and the use of classical word embeddings which don't provide deep contextual word embeddings. To date, there has been no research on explainable medical text for image protocol classification tasks or on the decision-making process of these models to identify potential systematic errors that may need to be addressed.

Recently bidirectional RNN's and transformers have improved text representation to be sensitive to its local context in a sentence and optimized for specific tasks by using a self-attention mechanism to help embed the context of each word. Large language models such as BERT and ELMo have been shown to provide substantial performance improvements for language modeling and text classification. We hypothesize that the use of context-dependent token embeddings will provide a substantial improvement for medical text classification and model interpretation. While there has been recent work evaluating large pretrained models for specialized tasks such as legal contract review [35], to the best of our knowledge, this paper is the first to evaluate how these models will perform on this specialized medical text which poses different challenges.

Furthermore, in the case of high stake applications, both accuracy and trust are necessary for the adoption of the model's decisions. Recent studies have focused on incorporating model explanations to improve trust [36, 37]. Explainable models have been developed to visualize word importance and attention layers. This has provided researchers with insight into understanding the model's decisions [38]. However, to the best of our knowledge, no other group

has attempted to evaluate if machine learning models can provide valid explanations for specialized medical texts.

Conclusion

In this study, we demonstrate state-of-the-art performance for the radiologic protocol classification task and provide a better understanding of how natural language processing (NLP) models make decisions in the medical domain. Using a large dataset of over 80,000 entries annotated by medical experts, we evaluated different pretrained BERT models and found that they significantly outperformed existing machine learning methods. We showed that BERT is able to identify relevant words that are highly indicative of the target protocol. The differences in BERT and human word importance were driven by BERT not recognizing specific anatomic structures and specialized medical terms that are important for humans. Furthermore, our analysis of the errors revealed that the largest source of errors was due to the model's difficulty in understanding the hierarchy of protocol assignments, while the third largest contributor was potential limitations or biases in the dataset.

Overall, our findings demonstrate that BERT can provide valuable insight into its decision making process for specialized medical tasks. This insight is valuable in understanding the error profile of the model. Understanding BERT's decision making process is a necessary step to deploying it in a real-life clinical environment.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12911-024-02444-z>.

Additional file 1: Supplemental. Word Importance

Acknowledgements

Not applicable.

Authors' contributions

ET and ST conceived of the research study with input from MM. ST contributed toward the design, implementation and evaluation of machine learning models. ET curated the dataset and evaluated the model's errors. AL and GY performed the human reader study. GZ provided project feedback. ET, MM ST managed the project vision and implementation along with writing of the manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets utilized during this study are not publicly available due to reasonable privacy and security concerns. The data is not easily redistributable to researchers other than those engaged in the Institutional Review Board-approved research collaborations with Stanford University. The dataset remains proprietary to the institution, and formal proceedings to make it publicly available has not been undertaken. Requests for data access can be made by Stanford affiliated users.

Declarations

Ethics approval and consent to participate

This retrospective study (and all experimental protocols) was conducted with the approval of the Stanford Institutional Review Board (IRB) and under a waiver of informed consent. The study was approved for collaboration between Stanford University and the University of California, Berkeley. All methods were carried out in accordance with the relevant guidelines and regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 August 2023 Accepted: 28 January 2024

Published online: 07 February 2024

References

- Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annual review of biomedical engineering. 2017;19:221.
- Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for health-care: review, opportunities and challenges. Briefings in bioinformatics. 2018;19(6):1236–46.
- Madani A, Ong JR, Tibrewal A, Mofrad MR. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. NPJ digital medicine. 2018;1(1):1–11.
- Yoojoong Kim, et al. "Predicting medical specialty from text based on a domain-specific pre-trained BERT." *Int J Med Inform.* 2023;170:104956.
- Turchin Alexander, Masharsky Stanislav, Zitnik Marinka. Comparison of BERT implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked.* 2023;36: 101139.
- Wang A, Pruksachatkun Y, Nangia N, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver; 2019. p. 3261–3275.
- Pandey B, Kumar Pandey D, Pratap Mishra B, Rhmann W. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. *J King Saud Univ Comput Inf Sci.* 2021:1–17.
- F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint [arXiv:1702.08608](https://arxiv.org/abs/1702.08608), 2017.
- Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri OS Alamoodi AH, Bai J, Salhi A, et al. A systematic review of trustworthy and explainable artificial Intelligence in healthcare: assessment of quality, bias risk, and data fusion. *Inf Fusion.* 2023;96:156–91.
- (2019) Explainable ai: the basics policy brief. [Online]. Available: <https://royalsociety.org/-/media/policy/projects/explainable-ai/985-AI-and-interpretability-policy-briefing.pdf>
- G. Cina, T. Ro"ber, R. Goedhart, and I. Birbil, "Why we do need explainable ai for healthcare," arXiv preprint [arXiv:2206.15363](https://arxiv.org/abs/2206.15363), 2022.
- van Beek EJ, Kuhl C, Anzai Y, Desmond P, Ehman RL, Gong Q, Gold G, Gulani V, Hall-Craggs M, Leiner T, et al. Value of mri in medicine: more than just another test? *Journal of Magnetic Resonance Imaging.* 2019;49(7):e14–25.
- Blackmore CC, Mecklenburg RS, Kaplan GS. Effectiveness of clinical decision support in controlling inappropriate imaging. *Journal of the American College of Radiology.* 2011;8(1):19–25.
- Boland GW, Duszak R, Kalra M. Protocol design and optimization. *Journal of the American College of Radiology.* 2014;11(5):440–1.
- Schemmel A, Lee M, Hanley T, Pooler BD, Kennedy T, Field A, Wiegmann D, John-Paul JY. Radiology workflow disruptors: a detailed analysis. *Journal of the American College of Radiology.* 2016;13(10):1210–4.
- Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine.* 2018;378(11):981.
- Brown AD, Marotta TR. Using machine learning for sequence-level automated MRI protocol selection in neuroradiology. *Journal of the American Medical Informatics Association.* 2018;25(5):568–71. <https://doi.org/10.1093/jamia/ocx125>.
- Kalra A, Chakraborty A, Fine B, Reicher J. Machine Learning for Automation of Radiology Protocols for Quality and Efficiency Improvement. *Journal of the American College of Radiology.* 2020;17(9):1149–58. <https://doi.org/10.1016/j.jacr.2020.03.012>.
- Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics.* 2018;87:12–20. <https://doi.org/10.1016/j.jbi.2018.09.008>.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics; 2019:4171–4186. doi: <https://doi.org/10.18653/v1/N19-1423>
- Peters ME, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics; 2018:2227–2237. doi: <https://doi.org/10.18653/v1/N18-1202>
- Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2020;36(4):1234–40. <https://doi.org/10.1093/bioinformatics/btz682>.
- Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission." arXiv preprint [arXiv:1904.05342](https://arxiv.org/abs/1904.05342) (2019).
- T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz et al., "Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.
- Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
- Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In Proc. 34th International Conference on Machine Learning. 2017;70:3319–28.
- N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan et al., "Captum: A unified and generic model interpretability library for pytorch," arXiv preprint [arXiv:2009.07896](https://arxiv.org/abs/2009.07896), 2020.
- D. Alvarez-Melis and T. S. Jaakkola, "A causal framework for explaining the predictions of black-box sequence-to-sequence models," arXiv preprint [arXiv:1707.01943](https://arxiv.org/abs/1707.01943), 2017.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.
- Jain SWallace BC. Attention is not explanation. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019. p. 3543–56.
- Achiam, OpenAI Josh et al. "GPT-4 Technical Report" (2023).
- Bills S, Cammarata N, Mossing D, Tillman H, Gao L, Goh G, Sutskever I, Leike J, Wu J, Saunders W. Language models can explain neurons in language models. 2023. URL <https://openai-public.blob.core.windows.net/neuron-explainer/paper/index.html>. Accessed 14 May 2023.
- D Brown Andrew, R Marotta Thomas. A natural language processing-based model to automate mri brain protocol selection and prioritization. *Acad Radiol.* 2017;24(2):160–6.
- D. Hendrycks, C. Burns, A. Chen, and S. Ball, "Cuad: An expert-annotated nlp dataset for legal contract review," arXiv preprint [arXiv:2103.06268](https://arxiv.org/abs/2103.06268), 2021.

36. Lai V, Tan C. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In Proceedings of the conference on fairness, accountability, and transparency. 2019. pp. 29–38.
37. Hao Y, Dong L, Wei F, Xu K. Self-attention attribution: Interpreting information interactions inside transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 14. 2021. pp. 12 963–12 971.
38. Hayati SA, Kang D, Ungar L. Does bert learn as humans perceive? understanding linguistic styles through lexica. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2021. URL <https://arxiv.org/abs/2109.02738>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.