

Exploring the Performance and Explainability of BERT for Medical Image Protocol Assignment

Salmon Talebi*
stalebi@berkeley.edu
University of California, Berkeley
Berkeley, CA, USA

Elizabeth Tong*
etong@stanford.edu
Stanford University
Stanford, CA, USA

Mohammad R. K. Mofrad
mofrad@berkeley.edu
University of California, Berkeley
Berkeley, CA, USA

1 Abstract

2 Although deep learning has become state of the art for nu-
3 merous tasks, it remains untouched for many specialized
4 domains. High stake environments such as medical settings
5 pose more challenges due to trust and safety issues for deep
6 learning algorithms. In this work, we propose to address
7 these issues by evaluating the performance and explainability
8 of a Bidirectional Encoder Representations from Transform-
9 ers (BERT) model for the task of medical image protocol
10 assignment. Specifically, we evaluate the performance and
11 explainability on this medical image protocol classification
12 task by fine tuning a pre-trained BERT model and measur-
13 ing the word importance by attributing the classification
14 output to every word through a gradient based method. We
15 then have a trained radiologist review the resulting word
16 importance scores and assess the validity of the model's
17 decision-making process in comparison to that of a human.
18 Our results indicate that the BERT model is able to identify
19 relevant words that are highly indicative of the target proto-
20 col. Furthermore, through the analysis of important words in
21 misclassifications, we are able to reveal potential systematic
22 errors in the model that may be addressed to improve its
23 safety and suitability for use in a clinical setting.

24 1 Introduction

25 Machine learning systems are being rapidly adopted for
26 many applications including high-stakes settings such as
27 medical applications [18, 19, 22]. Recent progress with self-
28 attention techniques, and specifically Transformers, have
29 dominated the field of text processing and classification tasks.
30 Large pretrained Transformers have outperformed humans
31 on language understanding tasks such as SuperGLUE [26].
32 However, many specialized text analysis tasks do not make
33 use of modern machine learning methods. It remains ques-
34 tionable how well existing pretrained models will transfer
35 to large, specialized texts.

36 In many high-stake applications, such as medicine, law,
37 or security where the main workers are humans trained in
38 specialized tasks, the direct application of these machine
39 learning algorithms, without human oversight, is currently
40 inappropriate. This reason is not only due to accuracy con-
41 cerns, but also arise from the lack of explainability and trust

42 humans have for the machine learning algorithm. Therefore,
43 in order to implement a machine learning algorithm to help
44 with specialized medical tasks it must not only have human
45 level performance, but also provide trustworthy explanations
46 to the user [10]. Furthermore, model explainability is being
47 driven by laws and regulations which state that decisions
48 from machine learning algorithms must provide information
49 about the logic behind those decisions [1]. In fact, the lack of
50 explainability of ML models often plagues medical artificial
51 intelligence (AI) [8]. For these reasons, in high-stake settings,
52 explainability should be a priority for researchers.

53 In this study, we focus on the specialized task of identi-
54 fying medical imaging protocols within text descriptions.
55 Medical imaging plays a crucial role in modern healthcare,
56 allowing physicians to visualize the inside of the human
57 body in order to diagnose and manage various conditions.
58 Clinicians often order radiologic studies, such as magnetic
59 resonance imaging (MRI) or computed tomography (CT), to
60 help answer clinical questions and guide treatment decisions
61 [24].

62 Typically, when a physician orders an imaging study, he/she
63 will provide a brief description of the indication for the exam
64 outlining the patient's signs and symptoms, medical history,
65 and any relevant clinical findings. These requests are then
66 sent to the radiologists, who are responsible for reviewing
67 the orders and recommending a radiologic protocol that best
68 addresses the clinical question. A radiologic protocol is a
69 specific set of instructions that defines the type of radiologic
70 exam to be performed on a particular body part, taking into
71 account the patient's presentation and the expected imag-
72 ing findings. The protocol may involve different imaging
73 sequences contrast agents, imaging planes, field of view, etc
74 in an MRI exam.

75 Assigning the appropriate protocol requires a thorough
76 understanding of the radiological appearance of different
77 pathologies, as well as a detailed knowledge of the patients'
78 clinical presentation and medical history. It also requires
79 familiarity with the types of protocols offered by the institu-
80 tion, as different facilities may have different capabilities and
81 resources. In MRI, accurate protocol assignment is particu-
82 larly crucial to patient care, as the chosen protocol dictates
83 which sequences are obtained and can impact the quality
84 and diagnostic accuracy of the exam [3, 4].

*Both authors contributed equally to this research.

85 Traditionally, protocol assignment to each radiologic order
86 is done manually by the radiologists or radiology technol-
87 ogists. This can incur substantial costs to the healthcare
88 system. This tedious task may take up to at least 6% of the
89 radiologists' time [21]. With increasing radiology orders, an
90 automated process with high throughput and accuracy is
91 desirable to ensure patient care and to avoid radiologists'
92 burnout. However, given the high stakes of medical tasks,
93 machine learning models must be evaluated for any system-
94 atic biases or errors before they can be trusted by clinicians
95 and patients [7]. In order for these models to be used in
96 practice they need to provide valid explanations for how the
97 decisions are made.

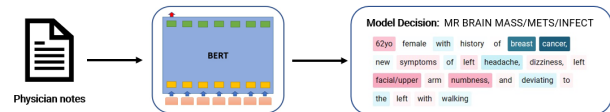
98 To address these problems, we fine-tuned a BERT model
99 using thousands of archived physician orders to learn the
100 medical language used to describe a given radiological exam.
101 Physicians' orders are generally written poorly, with many
102 typos and grammatical errors. In many cases they are written
103 with a few keywords to try and convey their point. Further-
104 more, they use terminology that only make sense in the
105 context of human anatomy or physiology. This can pose
106 challenges for existing pre-trained models as there is a distri-
107 bution shift between the physician's text and what existing
108 models have been trained on [15]. In addition, we evaluate
109 the model's ability to provide explanations of its decision
110 based on word importance. A trustworthy algorithm should
111 be able to demonstrate it is making complex decisions using
112 similar rationale to a human. For this application, explana-
113 tion is increasingly complex because the model will need to
114 understand language in the context of human anatomy and
115 physiology.

116 The main contributions of this study are as follows:

- 117 • We fine-tune a pre-trained BERT model using a medi- 151
118 cal dataset of medical imaging protocol text, and demon- 152
119 strate that it achieves state-of-the-art performance. 153
- 120 • We employ a gradient-based method called integrated 154
121 gradients to quantify the contribution that each word 155
122 in the input text makes to the model's decision. 156
- 123 • We validate the model's word importance claims using 157
124 a technique called erasure.
- 125 • We demonstrate that the model is capable of complex 158
126 decisions in a manner similar to that of a trained radi- 159
127 ologist.
- 128 • We analyze the model's mistakes using word impor- 160
129 tance and identify systematic errors that may pose 161
130 potential safety risks and need to be addressed before 162
131 the model can be safely deployed in a clinical setting. 163

132 2 Data

133 In order to train a specialized model for medical text clas- 164
134 sification, we have compiled a new large-scale dataset for 165
135 image protocol review. This dataset consists of deidentified 166
136 order entries and assigned protocols for magnetic resonance 167
168
169



137 **Figure 1.** A proposed system in which physician notes are
138 used as input to a model. The output of the model is an
139 imaging protocol, as well as an explanation of the process
140 by which the protocol was determined. This system aims
141 to provide a more efficient and accurate method for deter-
142 mining appropriate imaging protocols, while also offering
143 insight into the decision-making process of the model. By
144 incorporating an explainability component, the proposed
145 system has the potential to enhance trust and understanding
146 in the use of machine learning for medical image protocol
147 assignment.

148 (MR) neuroradiology studies that were conducted at our in-
149 stitution between June 2018 and July 2021. Each row in the
150 dataset represents a single radiology order and includes the
151 'reason for exam', patient age and gender, and the protocol
152 assigned by the radiologist.

153 We have excluded orders for spine imaging from this study,
154 as the assigned protocol typically reflects the specific seg-
155 ment of the spine indicated in the order. From the original
156 dataset of 119,093 rows, we removed the most common pro-
157 tocol, 'routine brain', as it can be used for a wide range of
158 indications and serves as the default protocol at our institu-
159 tion. The remaining dataset was narrowed down to the 10
160 most common protocols (Table 1).

161 To ensure the accuracy and quality of the data, we per-
162 formed a thorough review by an experienced radiologist
163 (ET) with 10 years of experience. We also applied standard
164 text preprocessing techniques, such as the removal of redun-
165 dant fields, handling of missing outputs, and expansion of
166 acronyms, to further clean and organize the data. The final
167 dataset includes 88,000 recorded notes with expert-annotated
168 imaging protocols.

169 3 Methods

170 This retrospective study was conducted with the approval of
171 the Stanford Institutional Review Board (IRB) and under a
172 waiver of informed consent. The study was approved for col-
173 laboration between Stanford University and the University
174 of California, Berkeley.

175 3.1 BERT Fine Tuning

176 We approach the problem of text classification as predict-
177 ing the class that corresponds to a given input text. In our
178 dataset, we have 10 possible classes that can be predicted. To
179 achieve this, we fine-tune a pre-trained BERT model using
180 the HuggingFace Transformers library [28].

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

Table 1. A comparison of imaging protocol F1 scores.

170 Before being processed by the BERT encoder, the input
171 data is transformed by passing it through three embedding
172 layers: a token embedding layer, a segment embedding layer,
173 and a position embedding layer. In the token embedding
174 layer, the input sentences are tokenized and each token is
175 transformed into a fixed-dimensional vector representation
176 (e.g., a 768-dimensional vector). Special classification [CLS]
177 and separator [SEP] tokens are also inserted at the beginning
178 and end of the tokenized sentence to serve as input represen-
179 tations and sentence separators for the classification task.

180 The segment embedding layer is useful for classifying
181 a text when provided with a pair of input texts. The po-
182 sitional embedding layer encodes the relative position of
183 tokens within a sentence using a sinusoidal function. The
184 final input embedding is the sum of these three individual
185 embeddings, which is then passed to the transformer for
186 further processing.

187 Resemblant to the clinical setting, the number in each pro-
188 tocol is not evenly distributed. More than half of the imaging
189 protocol entries belong to two of the classes. To mitigate this
190 imbalance we up sample the remaining 8 imaging protocols
191 so that the dataset is approximately balanced between all 10
192 classes of imaging protocols. The data is randomly split into
193 a train, validation and test sets. We have 70% of the protocols
194 make up the train set, 20% make up the validation set, and
195 10% make up the test set. The validation set was used to per-
196 form a hyperparameter grid search. The learning rate was
197 tuned from the range of 1×10^{-4} to 1×10^{-6} . During our exper-
198 iments we found the model would converge after 20 epochs
199 and training for any longer would degrade performance. The
200 model is trained using a single A6000 GPU.

201 3.2 Model Baseline

202 In order to establish a baseline and compare the performance
203 of our fine-tuned BERT model against traditional machine
204 learning methods, we conducted experiments using several
205 well-known algorithms, namely K-Nearest Neighbors (KNN),
206 Random Forest (RF), XGBoost, and Deep Neural Networks

(DNN). These algorithms have been used in previous stud-
ies for medical imaging protocol assignment and provide a
benchmark to evaluate the effectiveness of our approach.

To implement and evaluate the traditional machine learn-
ing methods, we used popular and widely adopted Python
libraries for each of the algorithms. For KNN, RF, and XG-
Boost, we utilized the scikit-learn library. For the DNN, we
employed Keras for building a 1D Convolutional Neural Net-
work (CNN) classifier.

216 3.3 Word Importance

217 For the purposes of this study, we use the concept of "word
218 importance" as a means of interpreting the model. Word
219 importance quantifies the contribution that each word in
220 the input text makes to the model's prediction. To calculate
221 word importance, we utilize a gradient-based method called
222 integrated gradients [16, 23].

223 Integrated gradients exploit the gradient information of
224 the model by integrating first-order derivatives. This method
225 does not require the model to be differentiable or smooth,
226 making it particularly suitable for large and complex models
227 such as Transformers. We use integrated gradients to accu-
228 rately estimate the importance of individual words within
229 an input sentence.

230 The integrated gradients method can be formally defined
231 as follows: let x be the input sentence, represented as a
232 (x_1, \dots, x_m) , and let x' be a "blank" baseline input. We have
233 a trained model F , and $F(x)_n$ is the output of the model at
234 time step n . The contribution of the m th word in x to the
235 prediction of $F(x)_n$ can be calculated by taking the integral
236 of gradients along the straight line path from x' to the input
237 x . In other words, we are measuring how much the predic-
238 tion at time step n changes as we move from the baseline
239 input x' to the actual input x , and specifically how much the
240 m th word in x contributes to this change.

241 The word importance value of each word in the input is cal-
242 culated by summing the scalar attributions across the dimen-
243 sions of the input embeddings. A positive attribution value

244 indicates that the word contributed to the prediction made
245 by the model, while a negative attribution value indicates
246 that it opposed the prediction. In cases of the BERT model,
247 which uses sub-word tokenization to divide rare words into
248 smaller pieces, we can obtain word-level attributions that
249 are more understandable to humans by taking the sub-word
250 with the highest absolute attribution value as the attribution
251 for the entire word.

252 3.4 Validating Word Importance

253 The assumption to use heat-maps of attribution values over
254 the inputs as explanations is particularly popular for natural
255 language processing. To test the validity of these explana-
256 tions, "stress tests" can be designed using a method called
257 erasure, where the most or least important parts of the input,
258 as indicated by the explanation, are removed and the model's
259 prediction is observed for changes [2]. Specifically, we erase
260 the most (or least) important word from the input sentence
261 and measure the resulting model accuracy.

262 3.5 Aggregating word attribution

263 We aggregate the word attributions across multiple texts for
264 each imaging protocol. Integrated gradient assigns attribu-
265 tion scores to each prediction made on a text segment that
266 is a maximum of 512 sub-words long. We calculate the top 5
267 words for each imaging protocol by taking the average attri-
268 bution value for each word across all text for a given imaging
269 protocol, and select the top words as those with the highest
270 average attribution value. We further filter out words that
271 appear in less than 3 texts. A trained radiologist assigned
272 a measure of word importance across all text for a given
273 imaging protocol. This measure was based on a numerical
274 score, with a value of 1 indicating a strong influence on the
275 radiologist's decision, 0.5 indicating a slight influence, and
276 0 indicating a neutral influence. For each word, the human
277 word importance score was determined as the average of
278 all word scores across a single image protocol class. These
279 methods were employed to generate lists of the most influ-
280 ential words for each imaging protocol, utilizing both the
281 BERT model and the judgments of the trained radiologist.

282 4 Results and Analysis:

283 The results of our fine-tuning experiment on the BERT model
284 are shown in Table 1. The model's performance was evalu-
285 ated using three metrics: precision, recall, and F1 score. The
286 F1 score is a measure of the model's accuracy, taking into
287 account both the precision and recall of the model. We found
288 that the BERT model had an F1 score of 0.89, which repre-
289 sents a significant improvement over the results of previous
290 studies using other machine learning methods. One such
291 study using deep neural network, random forest algorithm,
292 and k-nearest neighbors (kNN) achieved a F1 scores of only
293 0.83, 0.81 and 0.76 respectively [14].

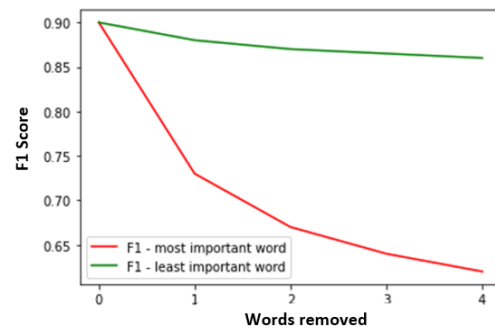


Figure 2. Model performance after step-wise removal of the 4 most important words and the 4 least important words from the text prompt. The results show that the least important words are less likely to degrade model performance while the most important words substantially degrade the performance.

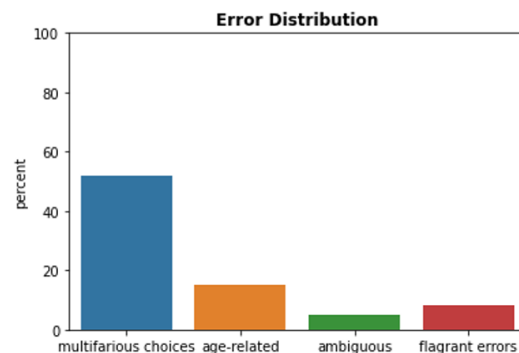


Figure 3. The bar plot decomposes the mistakes into four categories: multifarious choices, age-related, ambiguous text, and flagrant errors.

294 For our dataset, we also measured the weighted average F1
295 scores of the traditional machine learning models: XGBoost
296 achieved an F1 score of 0.84, RF scored 0.77, KNN obtained
297 0.70, and the DNN yielded an F1 score of 0.85. These results
298 are comparable to the performance of existing studies. Over-
299 all, the results of our experiment demonstrate the superior
300 performance of the pre-trained BERT model compared to
301 non-Transformer based approaches. The BERT model was
302 able to achieve a higher level of accuracy, as indicated by
303 the higher F1 score, and outperformed other methods in this
304 task.

305 4.1 Word Importance

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

309 To verify the validity of these attribution scores, we con-
310 ducted a "stress test" using a technique called erasure. This













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

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

involved systematically removing the most and least important words from the input text, and measuring the resulting impact on the performance of the BERT model. The results of this stress test are shown in Figure 2. 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 topmost important words was removed. Each subsequent removal of the most important words also resulted in a decremental drop in the F1 score.

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 for each image protocol and investigate the difference in the word importance ranks of BERT and those of a radiologist (figure 4). Both human (trained radiologist) and BERT picked the words most frequently mentioned in the indications for brain mass workup. 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.

4.2 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) 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

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

Figure 5. Selected samples from the dataset. 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.

390 protocols, and the model struggled to select the most appropriate one. This accounted for 52% of the errors in the test set.

391
392
393 Errors in the second category, age-related results, occurred
394 when the model failed to consider the age of the patient in
395 its prediction. For example, the best protocol for a patient
396 with intracranial hemorrhage may vary depending on their
397 age group. This category accounted for 15% of the errors in
398 the test set. Errors in the third category, ambiguous entries,
399 occurred when the model was unable to make a prediction
400 due to ambiguous or esoteric language in the input text.
401 This could include stems that were too rare or cryptic, or
402 protocols that could not be designated to ambiguous stems.
403 This category accounted for 5% of the errors in the test set.

404 Finally, flagrant errors, the fourth category, occurred when
405 the model made a wrong prediction or the order of word im-
406 portance did not make sense for the prediction. This category
407 accounted for 28% of the errors in the test set.

408 Overall, the largest issue for the model was its difficulty
409 in understanding the hierarchical ordering of protocols. This
410 accounted for 52% of the errors in the test set, and will require
411 further work to address before the model can be used in a
412 clinical setting. Another issue was the model’s partial capture
413 of important regions of the input text, which accounted for
414 15% of the errors. This may be due to biases or limitations
415 in the training data, and will also require further work to
416 address. By understanding the patterns of errors made by
417 the model, we can begin to identify areas for improvement
418 and fine-tune the model to achieve even better performance.

419 5 Discussion

420 Protocoling is a crucial task for radiologists to ensure that
421 the appropriate sequences are acquired in response to clinical
422 questions. However, manual protocolling can be time-
423 consuming, disruptive, and prone to errors. In recent years,
424 the volume of radiologic orders has increased, making pro-
425 tocoling an increasingly costly burden. To address these
426 challenges, we utilized a large pre-trained language model
427 that was fine-tuned by training it with a large dataset of
428 radiologic orders. This allowed the model to learn medical
429 terminology and accurately process orders, which frequently
430 contain typos, acronyms, and grammatical errors, and are
431 often written in shorthand using specialized medical termi-
432 nology.

433 Furthermore, in response to the increasing demand for ‘ex-
434 plainable AI’, we investigated the decision-making process
435 of our model. We evaluated the model’s ability to provide ex-
436 planations of its decision based on ‘word importance’. Model
437 explanation techniques were applied to estimate the impor-
438 tance of each word within the text of each radiologic order.
439 This allowed us to delve into the model’s decision-making
440 process and determine whether it was making correct pre-
441 dictions for the right reasons, as well as to identify the root
442 causes of any mistakes. Our results indicate that the BERT
443 model is able to identify relevant words that are highly in-
444 dicative of the target protocol.

445 Our error analysis revealed that the model struggled most
446 with understanding complex indications involving multiple
447 clinical questions, leading to incorrect protocol selection
448 in some cases. For example, the model may have difficulty
449 distinguishing between protocols for a patient with acute

450 neurologic deficits after brain tumor resection, as it may 500
451 not fully comprehend the hierarchical ordering of protocols. 501
452 Furthermore, we identified that approximately 15% of the 502
453 model's mistakes were due to insufficient capture of impor- 503
454 tant regions of the input text. This could be due to various 504
455 factors such as bias in the training data or limited examples 505
456 of certain edge cases. 506

457 Overall, the utilization of integrated gradients in our anal- 507
458 ysis has provided valuable insights into the model's decision- 508
459 making process compared to that of a trained radiologist. 509
460 This information was used to evaluate the strengths and 510
461 weaknesses of the model, and will be instrumental in making 511
462 the model more robust and trustworthy before its application 512
463 in clinical settings. 513

464 6 Limitation 514

465 There are several limitations to consider in the context of 517
466 this study. First, our dataset comprised of neuroradiologic 518
467 orders from a single center, and thus may be limited in its 519
468 representation of the racial, social, and ethnic diversity of 520
469 other regions. Validation with datasets from different insti- 521
470 tutions is necessary to more accurately compare the model's 522
471 performance. Additionally, we limited the number of proto- 523
472 cols to the ten most commonly used protocols in this study, 524
473 which may not fully capture the breadth of protocols used in 525
474 clinical practice. The data was collected from routine clinical 526
475 work, which means that protocols were assigned by multiple
476 radiologists with varying levels of experience, potentially
477 leading to inter-operator variability. While the dataset is rela-
478 tively large at over 80,000 entries, it is possible that additional
479 data could further improve model performance. 527

480 Additionally, it is important to note that there may be sig- 529
481 nificant variations in the importance of certain words when 530
482 considering the perspectives of different radiologists. In this 531
483 study, we were constrained to a single radiologist when eval- 532
484 uating word-level agreement with BERT. However, in future 533
485 studies, it would be beneficial to evaluate word importance 534
486 from the perspectives of a diverse group of radiologists to 535
487 achieve more robust results. 536

488 7 Related Work 537

489 Previous work has been done using classification models 540
490 to predict imaging protocol from a physician's notes using 541
491 machine learning techniques such as SVM, Random Forests, 542
492 and Gradient Boosted Machine [5, 6]. More recently, a deep 543
493 neural network approach was used to automate radiolog- 544
494 ical protocols which showed a slight boost over kNN and 545
495 random forests [14]. However, these models are limited by 546
496 the size of the model and the use of classical word embed- 547
497 dings which don't provide deep contextual word embeddings 548
498 [27]. To date, there has been no research on explainable med- 549
499 ical text for image protocol classification tasks or on the 550

decision-making process of these models to identify poten-
tial systematic errors that may need to be addressed.

514 Recently bidirectional RNN's and transformers have im-
515 proved text representation to be sensitive to its local context
516 in a sentence and optimized for specific tasks by using a self-
517 attention mechanism to help embed the context of each word
518 [25]. Large language models such as BERT [9] and ELMO
519 [20] have been shown to provide substantial performance
520 improvements for language modeling and text classification.
521 We hypothesize that the use of context-dependent token
522 embeddings will provide a substantial improvement for med-
523 ical text classification and model interpretation. While there
524 has been recent work evaluating large pretrained models for
525 specialized tasks such as legal contract review [13], to the
526 best of our knowledge, this paper is the first to evaluate how
527 these models will perform on this specialized medical text
528 which poses different challenges.

529 Furthermore, in the case of high stake applications, both
530 accuracy and trust are necessary for the adoption of the
531 model's decisions. Recent studies have focused on incorpo-
532 rating model explanations to improve trust [17]. Explainable
533 models have been developed to visualize word importance
534 and attention layers [11]. This has provided researchers with
535 insight into understanding the model's decisions [12]. How-
536 ever, to the best of our knowledge, no other group has at-
537 tempted to evaluate if machine learning models can provide
538 valid explanations for specialized medical texts.

539 8 Conclusion 540

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

557 Overall, our findings demonstrate that BERT can provide
558 valuable insight into its decision making process for special-
559 ized medical tasks. This insight is valuable in understanding
560 the error profile of the model. Understanding BERT's deci-
561 sion making process is a necessary stop to deploying it in a
562 real life clinical environment.

9 Competing Interests

The authors declare that there are no competing interests.

10 AUTHOR CONTRIBUTIONS

ET and ST conceived of the research study. ST contributed toward the design, implementation and evaluation of machine learning models. ET curated the dataset and evaluated the model's errors. ET, MM ST managed the project vision and implementation along with writing of the manuscript.

11 DATA AVAILABILITY

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.

References

- [1] 2019. *Explainable AI: the basics policy brief*. <https://royalsociety.org/-/media/policy/projects/explainable-ai/985AI-and-interpretability-policy-briefing.pdf>
- [2] David Alvarez-Melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint arXiv:1707.01943* (2017).
- [3] C Craig Blackmore, Robert S Mecklenburg, and Gary S Kaplan. 2011. Effectiveness of clinical decision support in controlling inappropriate imaging. *Journal of the American College of Radiology* 8, 1 (2011), 19–25.
- [4] Giles W Boland, Richard Duszak, and Mannudeep Kalra. 2014. Protocol design and optimization. *Journal of the American College of Radiology* 11, 5 (2014), 440–441.
- [5] Andrew D Brown and Thomas R Marotta. 2017. A natural language processing-based model to automate MRI brain protocol selection and prioritization. *Academic Radiology* 24, 2 (2017), 160–166.
- [6] Andrew D Brown and Thomas R Marotta. 2018. Using machine learning for sequence-level automated MRI protocol selection in neuroradiology. *Journal of the American Medical Informatics Association* 25, 5 (2018), 568–571.
- [7] Danton S Char, Nigam H Shah, and David Magnus. 2018. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine* 378, 11 (2018), 981.
- [8] Giovanni Cinà, Tabea Röber, Rob Goedhart, and Ilker Birbil. 2022. Why we do need explainable ai for healthcare. *arXiv preprint arXiv:2206.15363* (2022).
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [10] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [11] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 12963–12971.
- [12] Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does bert learn as humans perceive? understanding linguistic styles through lexica. *arXiv preprint arXiv:2109.02738* (2021).
- [13] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268* (2021).
- [14] Angad Kalra, Amit Chakraborty, Benjamin Fine, and Joshua Reicher. 2020. Machine learning for automation of radiology protocols for quality and efficiency improvement. *Journal of the American College of Radiology* 17, 9 (2020), 1149–1158.
- [15] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, 5637–5664.
- [16] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896* (2020).
- [17] Vivian Lai and Chenhao Tan. 2019. 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*. 29–38.
- [18] Ali Madani, Jia Rui Ong, Anshul Tibrewal, and Mohammad RK Mofrad. 2018. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *NPJ digital medicine* 1, 1 (2018), 1–11.
- [19] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics* 19, 6 (2018), 1236–1246.
- [20] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*. 2227–2237.
- [21] Andrew Schemmel, Matthew Lee, Taylor Hanley, B Dustin Pooler, Tabassum Kennedy, Aaron Field, Douglas Wiegmann, and J Yu John-Paul. 2016. Radiology workflow disruptors: a detailed analysis. *Journal of the American College of Radiology* 13, 10 (2016), 1210–1214.
- [22] Dinggang Shen, Guorong Wu, and Heung-Il Suk. 2017. Deep learning in medical image analysis. *Annual review of biomedical engineering* 19 (2017), 221.
- [23] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.
- [24] Edwin JR van Beek, Christiane Kuhl, Yoshimi Anzai, Patricia Desmond, Richard L Ehman, Qiyong Gong, Garry Gold, Vikas Gulani, Margaret Hall-Craggs, Tim Leiner, et al. 2019. Value of MRI in medicine: more than just another test? *Journal of Magnetic Resonance Imaging* 49, 7 (2019), e14–e25.
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [26] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems* 32 (2019).
- [27] Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics* 87 (2018), 12–20.
- [28] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.