SIIM20
VIRTUAL MEETING
Powered by SIIMU Online Learning
June 24-26
Reimagining the Future.
CELEBRATING 40 YEARS

Are Commercial Deep Learning Builders Superior to Traditionally Built Machine Learning Models for Natural Language Processing? A Head-to-head Comparison using Abdominal CT Protocol Classification

Brian A. Xavier, MD, Cleveland Clinic; Po-Hao Chen, MD, MBA

Introduction

Radiologists and radiology technologists constantly decipher provider submitted text data in the form of study indications to select the correct protocol for each imaging study requested. Protocol assignment is amenable to natural language processing (NLP). However, developing and maintaining state-of-the-art NLP models can be challenging. While some data scientists deploy a combination of programming languages and GUI workbenches to build models, commercial automatic machine learning (ML) tools attempt to simplify that process with simple and wizard-based tools to train advanced machine learning models. We compared machine learning models for abdominal CT protocoling built with an automatic commercial tool with those built using more manual workflows.

Hypothesis

We predict that automatic machine learning tools will perform on par with traditionally designed machine learning algorithms for predicting abdominal CT protocols in radiology.

Methods

94,510 abdominal CT studies performed between 12/30/2015 and 09/15/2019 were evaluated. Only the free text components (ordering provider's study indication and diagnosis text) and the final examination protocol were used. 11 categories of abdominal CT protocols were included: abdomen & pelvis, abdomen, pelvis, renal stone, kidneys, adrenal, pancreas, liver, urograms, cystograms, and enterographies (figure 1). The user-provided indication text data was processed using Knime (Knime AG, Zurich, Switzerland) including removal of stop words, stemming using a Snowball stemmer, and conversion to unigrams and bigrams before being vectorized using inverse document frequency (IDF). Four machine learning algorithms and one deep learning algorithm were constructed including a Random Forest (RF), Tree Ensemble (TE), Gradient Boosted Tree (GBT), and multi-layer perceptron (MLP). The Universal Language Model based deep learning algorithm (ULMFiT) was built using the FastAI Python library and trained on Wikipedia articles followed by fine-tuning on radiology text for the protocol classification task. 90% of the data was used for model training and 10% for validation. Finally, the same data was processed using the automatically generated machine learning platform Google AutoML (Alphabet, Inc., Mountain View, CA) for comparison.

Results

Metrics of precision, recall, and F1 score were obtained for all models. The machine learning algorithms all performed similarly with F1 scores of 0.834 for RF, 0.833 for TE, 0.835 for GBT, 0.816 for MLP, and 0.844 for ULMFiT. The Universal Language Model performed the best with a recall of 85.1% and precision of 84.2%. The automatically generated model with identical input data performed better with a recall of 83.9%, precision of 87.0%, and F1 score of 0.854. The models performed best on the abdomen and pelvis protocols category (F1 scores: 0.880-0.898) followed by the adrenal protocols (F1 scores: 0.743-0.889).

Conclusion

The commercial, automatic deep learning platform can quickly create models with good classification results for abdominal CT protocol classification on par with or better than deep learning and non-deep learning models created through less automated methods.

Statement of Impact

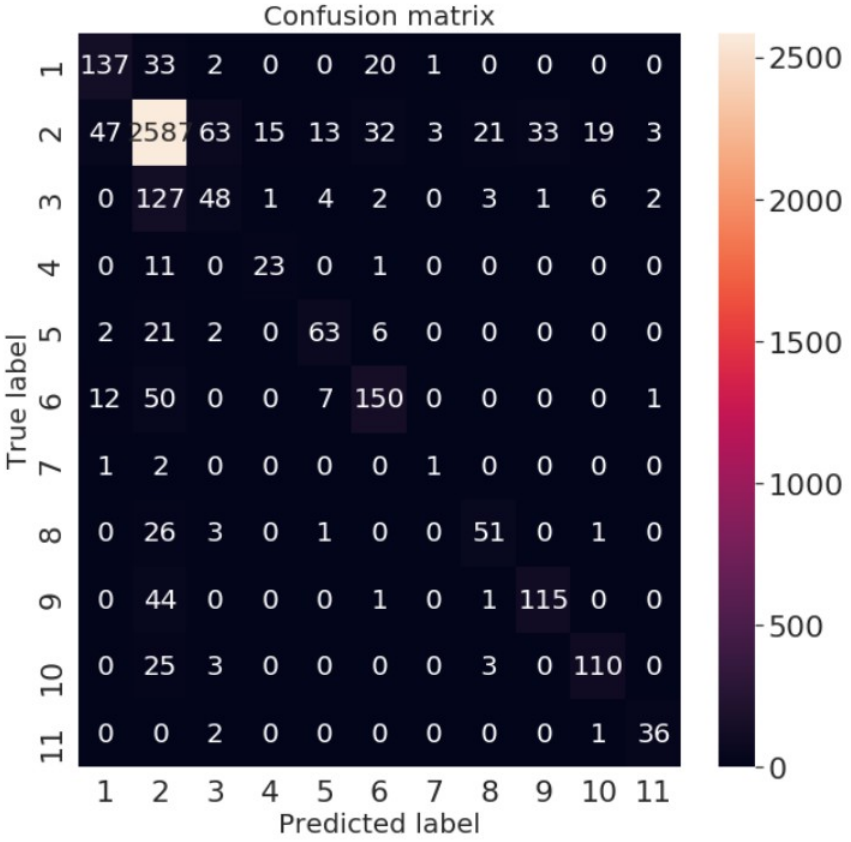
NLP using automatically generated deep learning models for classification offers low barrier-of-entry and high performance, warranting further assessment for the development of clinically useful models.

Keywords
natural language processing, machine learning, deep learning, artificial intelligence, radiology, abdominal Imaging

Figure 1
Abdominal CT protocol categories

1	Renal Stone	Acute flank pain/Renal stone protocol
2	Abdomen and Pelvis	With IV and oral contrast, with IV, oral and rectal contrast, with IV without oral contrast, without IV with oral contrast, without iv without oral
3	Abdomen	With IV and oral contrast, with IV without oral contrast, without IV with oral contrast, without iv without oral
4	Pelvis	With IV with oral, with IV, oral, and rectal, without IV with oral and rectal
5	Kidney	Triple phase kidneys with pelvis, Triple phase kidneys without pelvis
6	Urogram	Triple phase urogram, Urogram with split bolus
7	Cystogram	Cystogram with IV contrast, Cystogram without IV contrast
8	Pancreas	Pancreas protocol with pelvis, Pancreas protocol without pelvis
9	Enterography	1-Phase Enterography, 2-Phase Enterography
10	Liver	2-Phase Liver protocol, 2-Phase Liver protocol with pelvis, 3-Phase Liver protocol, 3- Phase Liver protocol with pelvis, 4-Phase Liver protocol, 4-Phase Liver protocol with pelvis
11	Adrenal	Adrenal protocol with contrast, Adrenal protocol without contrast, Adrenalectomy protocol

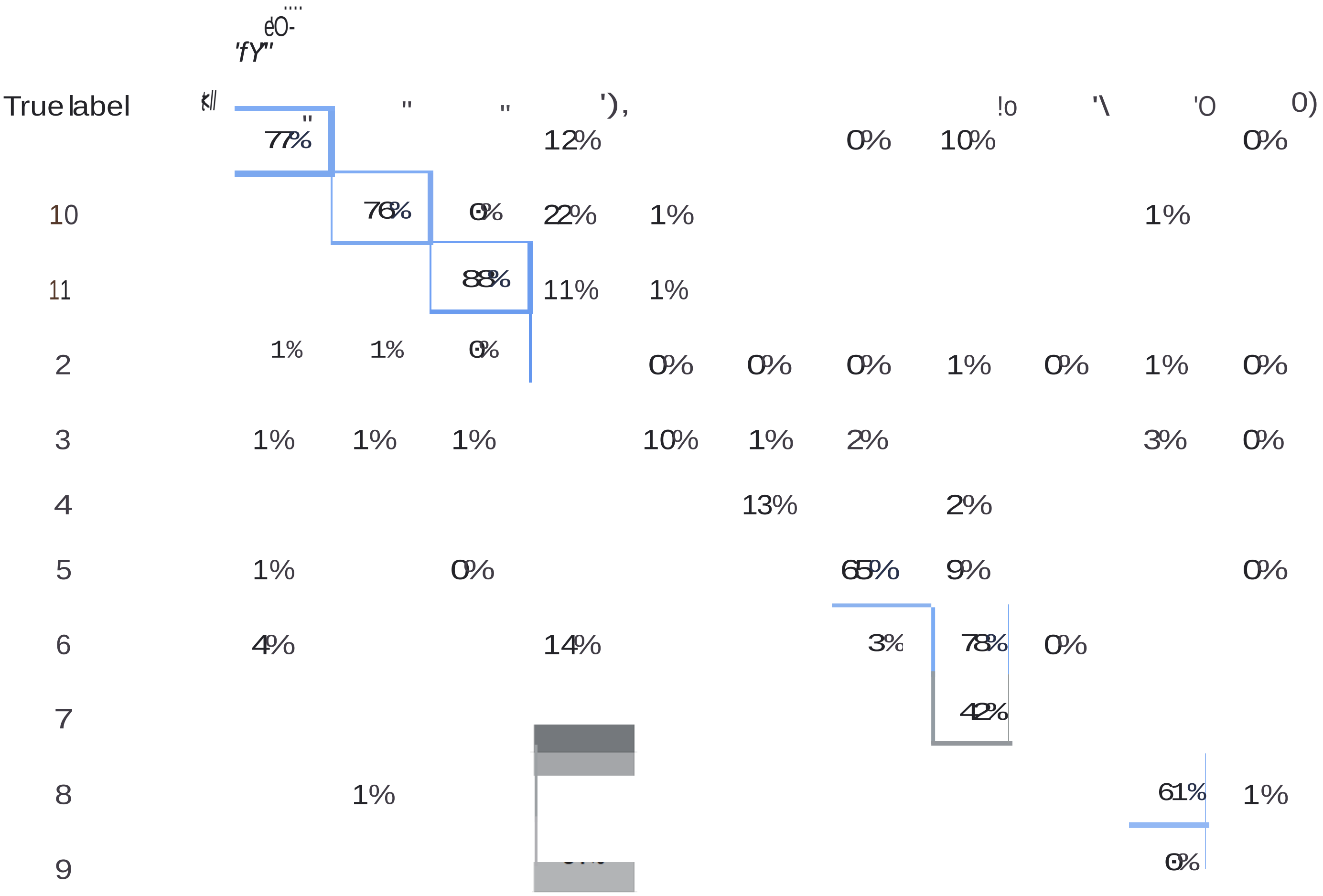
Figure 2
ULMFiT Model Confusion Matrix



	1	2	3	4	5	6	7	8	9	10	11
1	137	33	2	0	0	20	1	0	0	0	0
2	47	2587	63	15	13	32	3	21	33	19	3
3	0	127	48	1	4	2	0	3	1	6	2
4	0	11	0	23	0	1	0	0	0	0	0
5	2	21	2	0	63	6	0	0	0	0	0
6	12	50	0	0	7	150	0	0	0	0	1
7	1	2	0	0	0	0	1	0	0	0	0
8	0	26	3	0	1	0	0	51	0	1	0
9	0	44	0	0	0	1	0	1	115	0	0
10	0	25	3	0	0	0	0	3	0	110	0
11	0	0	2	0	0	0	0	0	0	1	36

Figure 3

Automatically Generated ML Model Confusion Matrix



True label											
	77%			12%			0%	10%			0%
10		76%	0%	22%	1%					1%	
11			88%	11%	1%						
2	1%	1%	0%		0%	0%	0%	1%	0%	1%	0%
3	1%	1%	1%		10%	1%	2%			3%	0%
4						13%		2%			
5	1%		0%				65%	9%			0%
6	4%			14%			3%	78%	0%		
7								42%			
8		1%								61%	1%
9										0%	