Towards Verifying Results from Biomedical NLP Machine Learning Models Using the UMLS: Cases of Classification and Named Entity Recognition

Joan Byamugisha[1], **Waheeda Saib**[1], **Theodore Gaelejwe**[1], **Asad Jeewa**[1], **Maletsabisa Molapo**[1]
[1] IBM Research Africa
joan.byamugisha@ibm.com, wsaib@za.ibm.com, theodore.gaelejwe@ibm.com, asad.jeewa@ibm.com, maletsabisa.molapo@ibm.com

Abstract

Machine learning (ML) for biomedical research is one of the fastest growing research areas in the world today. For NLP specifically, free-text healthcare reports are an important resource whose processing can contribute potentially to patient diagnosis, treatment, and management. However, the inability to explain the outputs of ML algorithms is currently a barrier to the use of these models in a clinical setting. We present a method that uses the ontologies and knowledge-bases in the Unified Medical Language System (UMLS) to verify and explain the output of biomedical ML models. Our verifier takes as input the results from an ML model, and uses the UMLS to correlate the results of the task with the confidence of the model for each result. We applied this architecture to two tasks using textual cancer pathology reports: ICD-O topography classification, and named entity recognition. For the former, we identified that the presence of certain entities in a report is inversely related to the model's confidence values; while, for the latter, we identified categories of errors related to lower confidence values. Our approach, therefore, not only verifies the accuracy of ML model results, but provides explanations that may be used to improve model design and performance.

Introduction

Recent advances in Machine Learning have unlocked data processing capabilities that were hitherto very limited or not possible. The benefits of applying these algorithms to medical research have led to ground-breaking results in medical imaging (Erickson et al. 2017; Ravì et al. 2016), digital health (Triantafyllidis and Tsanas 2019), and insights extraction from textual data (Townsend 2013), among others. For textual data specifically, current efforts include: text mining from biomedical literature (Lee et al. 2020; Deng et al. 2019; Sheikhalishahi et al. 2019), text mining from free-text reports (Huang, Altosaar, and Ranganath 2020; Yala et al. 2017), categorizing findings in pathology and radiology reports (Imler et al. 2013; Saib et al. 2020; Pons et al. 2016), and generating structured information from free-text reports (Kreimeyer et al. 2017).

Despite the impressive results of machine learning research in healthcare, there remains a barrier to the deployment of these models in clinical settings and much of these efforts do not go beyond research and archival purposes (He et al. 2019; Wiens et al. 2019). The black-box nature of machine learning models is the major impediment to trusting the results they output (Mittelstadt, Russell, and Wachter 2019; Tjoa and Guan 2020). This has spawned the field of explainable-AI (XAI), which aims to create models that are transparent and whose results can be explained and understood (He et al. 2019). Current approaches range from: creating visualizations in order to understand the inner workings of the model (Karpathy, Johnson, and Fei-Fei 2015); analyzing features by isolating the contribution of individual features (Cotton 2017); intrinsically interpreting and explaining models through reasoning (Cotton 2017); and retrospectively justifying and explaining the output of the models (Danilevsky et al. 2020). Our focus is on the retrospective evaluation of model results, and we investigated this using a stable, accurate, and trusted biomedical standard, the Unified Medical Language System - the UMLS (Bodenreider 2004a).

We developed a novel model-agnostic architecture that applies an industry standard knowledge repository, the UMLS, to perform the role of a verifier by retrospectively validating and/or explaining the output from a machine learning model, thus providing a level of interpretability to a model's outputs. Depending on the machine learning task at hand, the UMLS-based verifier applies different terminologies to provide further information on an output-by-output basis of the results of the model. A final report is then generated, and this report consists of information from the UMLS that supports or contradicts the results of the model. We evaluated this architecture on two tasks, ICD-O topography classification and named entity recognition (NER) on cancer pathology reports; with the National Cancer Institute metathesaurus (NCIm) used for the former, and sixteen other terminologies used for the latter. Our results show that for classification, the presence of certain entities is inversely proportional to the model's confidence values; while for NER, the model's accuracy can be determined on an entity-by-entity basis, and even categorized, and these categories then correlated to the model's confidence values. To the best of our knowledge, this is the first attempt to use the UMLS in a model-agnostic architecture, to not only identify inaccuracies in the results of a machine learning model, but to correlate observations from the UMLS to a model's confidence values.

Related Work

Due to the high level of accountability needed within the medical domain, the requirement to justify machine learning model reliability is mandatory (Tjoa and Guan 2020). However, the existing formal guidelines and generally accepted practices for explainability and transparency are limited (Mittelstadt, Russell, and Wachter 2019).

One class of solutions involves examining models for potential bias (Wiens et al. 2019) and evaluating the system rigorously before deployment (Wiens et al. 2019), since model accuracy alone does not ensure that a model will gain clinical acceptance (Stultz 2019). For example, the detection of osteoporosis by deep learning models, three years before actual diagnosis, from an analysis of MRI scans was supported further by the explanation of the specific patterns that the model detected, and their association to the resulting diagnosis (Kundu et al. 2020). Additionally, a study on the use of convolutional neural networks (CNNs) to classify patient phenotypes showed that the results could be interpreted by computing the saliency of the inputs, which is a method similarly used in rule-based approaches that rely on text analysis and knowledge extraction (Gehrmann et al. 2017).

Another class of solutions is the integration of knowledge-based tools with machine learning models, in order to combine the interpretability of the former with the high efficiency of the latter (Holzinger et al. 2017). Examples of such systems include: the combination of the rule-based Clinical Text Analysis and Knowledge Extraction System (cTAKES) and Logistic regression (Gehrmann et al. 2017); and the combination of rule-based feature engineering and domain knowledge-infused CNNs for clinical text classification (Yao, Mao, and Luo 2019). The knowledge source can be either custom ontologies and taxonomies for explainability (Arrieta et al. 2019; Arya et al. 2019), such as the healthcare-centric explanation ontology by (Chari et al. 2020), or industry standard ontologies and taxonomies such as the Unified Medical Language System (UMLS) (National Library of Medicine 2020).

The UMLS, a repository of millions of names for about a million medical concepts with tens of millions of relations among these concepts from 216 families of biomedical terminologies (National Library of Medicine 2020; Bodenreider 2004b), has been heavily used within the clinical NLP domain (Humphreys, Del Fiol, and Xu 2020), in the projection of word embeddings onto interpretable lower dimensional spaces (Rothe, Ebert, and Schtze 2016) and in the development of hybrid models (Faralli et al. 2016).

Despite the positive steps taken towards a solution by the above approaches, significant limitations still remain regarding adequately solving the problem of trust in machine learning systems for healthcare (Ching et al. 2018). Efforts to develop high-level interdisciplinary guidelines for responsible AI in healthcare have thus far resulted in the development of industry standards for reasoning and data storage (such as the terminologies in the UMLS), but not regarding the development of machine learning algorithms themselves. Also, examining models for potential bias and evaluating the system rigorously before deployment leave a lot of the machine learning trust problems to be handled intrinsically (either by model design or system architectures), which in some cases will result in a situation of the fox guarding the henhouse. Additionally, with the architectures of machine learning models becoming more and more complex, with deeper and deeper layers, we are decades away from an intrinsic solution, yet healthcare systems are needed now. Finally, combining the interpretability of knowledge-based tools with the high efficiency of machine learning models is a safe middle-ground, that can be used to retrospectively justify and explain the output of these models (Danilevsky et al. 2020). However, the choice of the knowledge source is key. While the use of custom knowledge sources (Arrieta et al. 2019; Arya et al. 2019) achieves the desired goal, they lack the trust inherent in industry standard knowledge sources, such as the UMLS that integrates and distributes the resources associated with key terminology, classification, and coding standards (National Library of Medicine 2020). On the other hand, due to the high calibre of the medical experts, healthcare documentation (industry standards, policies, guidelines, and laws), and institutions (public, private, academic, research, and medical) that are involved in the construction of the terminologies in the UMLS, it has become the healthcare industry's most trusted repository, and is, therefore, a well justified knowledge source with which to verify the results from machine learning models.

UMLS-based Verification of Machine Learning Model Outputs

We focused on the task of retrospectively justifying and explaining the output of machine learning models, where the UMLS is the expert knowledge-base with which to verify model outputs. The proposed architecture of such a system is shown in Figure 1.

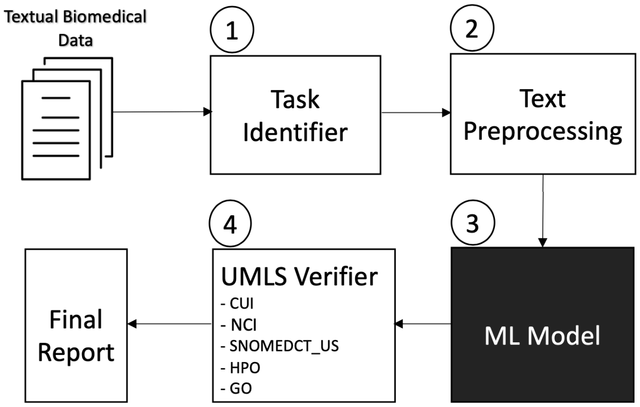

Figure 1: An architecture of the UMLS based verifier

In the architecture shown in Figure 1, the processing begins in *Task Identifier*, where the purpose of the system is determined. The task at hand can be classification, named entity recognition, information extraction, etc. The input into the system, *Textual Biomedical Data*, undergoes the text preprocessing procedures in *Text Preprocessing* that are appropriate for the task identified. The results from this are then fed into the *ML Model*, whose black-box nature is highlighted in the figure. Finally, the output from the *ML Model* passes on to the *UMLS Verifier* that generates a *Report* on the status of each of the model outputs, whether consistent or contradictory. The *UMLS Verifier* uses various terminologies (such as the *CUI*, *SNOMEDCT_US*, *NCI*, *GO*, and *HPO*) to reach its conclusions.

In practice, the UMLS-based verifier functions by taking raw text as input and, through medical concept extraction, obtains both the extracted medical entity and its Concept Unique Identifier (CUI). The CUI is then mapped to the terminologies associated with the task being performed, and if the mapping produces a concept, then further analysis is done on that concept. Conversely, if the mapping does not produce a concept, then it implies that the CUI is not associated with that terminology. The further implications of this are explained below for the classification and NER tasks.

In order to verify the consistency, or lack thereof, of the topographical ranges within a pathology report, the CUI for each extracted entity is mapped to the National Cancer Institute (NCI) terminology in order to obtain an NCIt code. The NCIt code is then used to obtain the entity's given label and synonyms. These, together with the NCIt code, are matched to the NCI's documented mappings of NCIt codes and labels to ICD-O topography codes, in order to obtain the most likely topographical range for that entity. This is repeated for every entity in the report, and results in a listing of all entities that have NCIt codes and their associated topographical range.

When verifying whether the correct tag has been assigned to an entity by the NER model, the entity's CUI is mapped to various terminologies to verify different tags. In cases where the terminology is an ontology instead of a knowledge-base, the CUI is first mapped to the ontology, and an ontology-specific code of the resulting concept, if available, is then used to obtain that concept's top-level concept. In this case, a single ontology can be used to verify different tags. For example, when using the SNOMED-CT ontology (which has nineteen top-level concepts) to verify an *Anatomy* tag, a code is obtained from the CUI-to-SNOMED-CT mapping, and if the top-level concept of this code is found to be *Body Structure*, then the *Anatomy* tag is considered to have an explainable basis. The current version uses sixteen terminologies from the UMLS metathesaurus for NER verification.

Materials And Methods for Evaluation of Architecture

We carried out an evaluation of the architecture shown in Figure 1 using two tasks: ICD-O topography classification and named entity recognition (NER).

For ICD-O topography classification, we were interested only in the topography (primary tumor site) in a pathology report. In the evaluation of the architecture for this task, we used 1964 unstructured and anonymized breast cancer pathology reports obtained from public and private health care laboratories across South Africa. The ground-truth coding for each report is based on manual annotations by expert human coders, and these codes formed the labels for the classification models (Saib et al. 2020). The deep learning model used was the Multi-Task Convolutional Neural Network (MT-CNN) with hard parameter sharing, which shares the hidden layers of the CNN across all tasks, while retaining task specific output layers for the different related tasks (Ruder 2017). MT-CNNs were preferred for this task because they have been shown to improve on the performance of single-task deep neural networks when applied to information retrieval and classification of primary tumor site and laterality (Yoon, Ramanathan, and Tourassi 2016). During classification, the most salient 1400 TF-IDF features were used to filter the corpus, thus allowing for the retention of the relevant words in each report (Saib et al. 2020). Finally, though the results from the MT-CNN included classification both by primary tumor site (topography) and by tumor cell origin (morphology), we only included the former in the evaluation.

For Named Entity Recognition (NER), we were interested in verifying the tags assigned to the entities in a pathology report. When evaluating the architecture for NER, we used unstructured and anonymized breast, small intestine, and large intestine cancer pathology reports. The deep learning model used was HunFlair (Weber et al. 2020), which uses a pre-trained BiLSTM-CRF model that incorporates character-level contextual embeddings of (Akbik, Blythe, and Vollgraf 2018), and FastText word embeddings (Bojanowski et al. 2017) which are based on the Skip-gram model as described in (Mikolov et al. 2013). This method was selected because the use of such embeddings enhances model performance over a variety of tasks including NER (Akbik, Blythe, and Vollgraf 2018), (Akbik, Bergmann, and Vollgraf 2019), (Peters et al. 2017), (Peters et al. 2018), (Wiedemann, Jindal, and Biemann 2018), (Zhai, Nguyen, and Verspoor 2018). Since the model is already pre-trained on 23 biomedical NER corpora, for five different entity types (Disease, Chemical, Gene, Cell line and Species) (Weber et al. 2020), it was directly applied to extract and label entities in the given reports.

The knowledge source used during the verification and explanation was the 2020 release of the UMLS metathesaurus, with the QuickUMLS tool (Soldaini and Goharian 2016) used to extract medical concepts (including multi-word medical terms) from text. For topography verification, a single terminology, the National Cancer Institute (NCI) was used, where a topographical range was identified for entities in the pathology report, and then consistency across topographical ranges was sought for the entire report. On the other hand, multiple terminologies that were associated with different tags were used to verify the NER results. On an entity-by-entity basis, a mapping was performed across the different terminologies in order to identify which terminology an entity is associated with. The label and properties of that entity within its associated terminology were then used to explain the agreement with, or contradiction to, the tag assigned by the NER model.

Results

Due to space limitations, we present only a summarized version of the results. For the ICD-O topography classification, as shown in Table 2 in the appendix, we found that the topographical range of breast cancer, *C50*, obtained through NCIt codes appears in at least one entity in every report. Some of the entities whose NCIt codes result in a topographical range of C50 (consistent entities) are: areola, axillary tail, breast, left breast, right breast, and nipple. However, we also observed the presence of other entities whose NCIt codes either do not result in a topographical range of C50 (inconsistent entities), or have no assigned topographical range (unclear entities). Some examples of the former include: axillary (C76), axillary nodes and axillary lymph nodes (C77), lymph (C42), muscle and vascular (C49), scar (C44), and trabecula (C41); while some examples of the latter include mucin, estrogen, tumour, and tumours. Further, as shown in Table 2 in the appendix, we found that the presence of inconsistent and unclear entities in a report relates to a model's performance in two ways: (1) when the model's prediction is correct, then the number of these entities is inversely proportional to the model's confidence value; and (2) when the model's prediction is incorrect, then the number of these entities is high.

For the NER model evaluation, due to space limitations, here we present results only on the *Disease* tagging of a single colon cancer report. As shown in Table 1 in the appendix, we identified that the accuracy of the NER model can be categorized into: (1) agreement, where the NER model and UMLS verifier produce the same output (sentences 2 and 5); (2) no annotation by the NER model, where entities identified as diseases by the verifier are missed by the NER model (Sentence 4); (3) partial annotation, where a multiword term is partially annotated by the NER model (sentences 1 and 6); (4) over annotation, where the NER model over-specifies a tag across multiple entity types (Sentence 3); and (5) incorrect annotation, where the NER model incorrectly tags non-disease entities as diseases (Sentence 7). Interestingly, the model confidence values are observed generally to be below 0.9 in the cases where the UMLS verifier identifies partial annotation. For example, in Table 1 in the appendix, *tumour* in Sentence 1 has a confidence value of 0.612; and *colorectal adenocarcinoma* in Sentence 6 has a confidence value of 0.592; though this is not always the case. On the other hand, the confidence values where the UMLS verifier agrees with the NER results are very high: *pneumoturia* in Sentence 1, and *haemorrhage* and *necrosis* in Sentence 5 with confidence values of 0.975, , 0.988, and 0.98, respectively.

In both these examples, the report output by the UMLS-verifier can include the use of these inconsistencies in analyses to explain a model's performance beyond its confidence in its predictions.

Discussion

Our architecture of a retrospective machine learning model output verifier based on a very large expert knowledge repository, the UMLS, is an important contribution towards trusting the results of AI systems in healthcare. The architecture presented in Figure 1 can be adopted to fit different machine learning models that are geared towards different tasks. This is different from some existing approaches whose main limitation lies in them being very model-specific, that is, carefully adapting a specific model to allow for both explainability and accuracy (Mullenbach et al. 2018). Our architecture provides a contribution towards a much-needed solution to the area of model-agnostic approaches that is currently very limited (Cotton 2017).

Additionally, the application of our architecture to verify the results of machine learning models brings to this problem area the full power and resource-richness of the UMLS. This can be used, not only to identify incorrect results from a machine learning model, but to go further and propose reasons for the incorrect results. This allows us to identify limitations in the model, an essential step in the healthcare domain (Myers et al. 2020; Stultz 2019). As an example, the evaluation of the classification model results suggest the entities that could be affecting model performance, and, most importantly, identifies characteristics in reports which can be used to explain incorrect model predictions, even when the model's confidence values are very high. For NER, the verification results point to causes of errors arising from partial entity extraction, no entity extraction, and over-extraction of entities. These reasons can provide a possible avenue to consider when developing the next class of NER models. This would be a directed investigation, as opposed to assuming that the errors can be corrected by, for example, tuning hyperparameters, using more training data, adding more hidden layers, applying different word-vector representations, etc..

Conclusion

In this paper, we presented an architecture for the retrospective verification of the results of machine learning models, as an avenue to creating trust in these models used in healthcare. The proposed architecture uses various terminologies from the UMLS to agree with or contradict the output from a machine learning model. We presented the results of the evaluation done on the architecture for two tasks, ICD-O topography classification and NER, showing how the UMLS verifier can be used to identify and explain incorrect results from a machine learning model. Future work will involve evaluating the verifier more deeply: by using biomedical text other than pathology reports, evaluating the fit-for-purpose of our architecture on more biomedical NLP tasks, and using the UMLS-based verifier in tandem with machine learning models to improve the accuracy of results on-the-fly.

Appendices

Samples of the results on Classification and NER.

Input Sentence	NER Model	UMLS Verifier
(1) A male with a rectosigmoid tumour and pneumoturia	tumour | pneumoturia	rectosigmoid tumour | pneumoturia
(2) Cystoscopy showed an infratrigonal fistula.	fistula	fistula
(3) Two previous biopsies showed high grade dysplasia of colonic mucosa.	high grade dysplasia of colonic mucosa	high grade dysplasia
(4) Sections show several fragments of tissue, showing predominantly ulceration with fibrinopurulent exudate on the surface.	-	ulceration | fibrinopurulent exudate
(5) There is extensive haemorrhage and necrosis associated with these fragments.	haemorrhage | necrosis	haemorrhage | necrosis
(6) There are free-lying cells as well as nests of stromal invasion within the fragments showing an invasive adenocarcinoma.	adenocarcinoma	invasive adenocarcinoma
(7) Necrotic debris is identified associated with the invasive component.	necrotic	-

Table 1: Comparison of disease entities identified by NER model and UMLS verifier in sample sentences from a single report

Predicted	Actual	Model Confidence	Consistent	Inconsistent	Unclear
Topography			Entities		
C50.0	C50.0	0.913	1	0	0
C50.1	C50.0	0.289	4	4	4
C50.2	C50.2	0.604	2	3	2
C50.4	C50.8	0.322	1	3	1
C50.0	C50.1	0.987	2	0	4
C50.9	C50.9	0.928	1	1	1
C50.9	C50.9	0.753	1	0	0
C50.5	C50.4	0.984	3	2	1
C50.5	C50.5	0.988	2	1	8
C50.5	C50.5	0.922	2	2	1

Table 2: Details on a sample of ten reports, showing the number of consistent, inconsistent, and unclear entities found in each report

References

Akbik, A.; Bergmann, T.; and Vollgraf, R. 2019. Pooled contextualized embeddings for named entity recognition. 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)*, 724–728.

Akbik, A.; Blythe, D.; and Vollgraf, R. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Arrieta, A. B.; Daz-Rodrguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garca, S.; Gil-Lpez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *arXiv:1910.10045 [cs]*. arXiv: 1910.10045.

Arya, V.; Bellamy, R. K. E.; Chen, P.-Y.; Dhurandhar, A.; Hind, M.; Hoffman, S. C.; Houde, S.; Liao, Q. V.; Luss, R.; Mojsilovi, A.; Mourad, S.; Pedemonte, P.; Raghavendra, R.; Richards, J.; Sattigeri, P.; Shanmugam, K.; Singh, M.; Varshney, K. R.; Wei, D.; and Zhang, Y. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv:1909.03012 [cs, stat]*. arXiv: 1909.03012.

Bodenreider, O. 2004a. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32(90001):267D–270.

Bodenreider, O. 2004b. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research* 32(1).

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

Chari, S.; Seneviratne, O.; Gruen, D. M.; Foreman, M. A.; Das, A. K.; and McGuinness, D. L. 2020. Explanation

Ontology: A Model of Explanations for User-Centered AI. *arXiv:2010.01479 [cs]*. arXiv: 2010.01479.

Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P.-M.; Zietz, M.; Hoffman, M. M.; et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* 15(141):20170387.

Cotton, C. 2017. Explanation and Justification in Machine Learning: A Survey. 6.

Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; and Sen, P. 2020. A Survey of the State of Explainable AI for Natural Language Processing. *arXiv:2010.00711 [cs]*. arXiv: 2010.00711.

Deng, Z.; Yin, K.; Bao, Y.; Armengol, V. D.; Wang, C.; Tiwari, A.; Barzilay, R.; Parmigiani, G.; Braun, D.; and Hughes, K. S. 2019. Validation of a semiautomated natural language processing–based procedure for meta-analysis of cancer susceptibility gene penetrance. *JCO Clinical Cancer Informatics* 3:1–9.

Erickson, B. J.; Korfiatis, P.; Akkus, Z.; and Kline, T. L. 2017. Machine learning for medical imaging. *Radiographics* 37(2):505–515.

Faralli, S.; Panchenko, A.; Biemann, C.; and Ponzetto, S. P. 2016. Linked Disambiguated Distributional Semantic Networks. In Groth, P.; Simperl, E.; Gray, A.; Sabou, M.; Krtzsch, M.; Lecue, F.; Flck, F.; and Gil, Y., eds., *The Semantic Web ISWC 2016*, volume 9982. Cham: Springer International Publishing. 56–64. Series Title: Lecture Notes in Computer Science.

Gehrmann, S.; Dernoncourt, F.; Li, Y.; Carlson, E. T.; Wu, J. T.; Welt, J.; Foote Jr, J.; Moseley, E. T.; Grant, D. W.; Tyler, P. D.; et al. 2017. Comparing rule-based and deep learning models for patient phenotyping. *arXiv preprint arXiv:1703.08705*.

He, J.; Baxter, S. L.; Xu, J.; Xu, J.; Zhou, X.; and Zhang, K. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine* 25(1):30–36.

Holzinger, A.; Biemann, C.; Pattichis, C. S.; and Kell, D. B. 2017. What do we need to build explainable AI systems for the medical domain? *arXiv:1712.09923 [cs, stat]*. arXiv: 1712.09923.

Huang, K.; Altosaar, J.; and Ranganath, R. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. In *ACM Conference on Health, Inference, and learning (CHIL) 2020 Workshop*.

Humphreys, B. L.; Del Fiol, G.; and Xu, H. 2020. The UMLS knowledge sources at 30: indispensable to current research and applications in biomedical informatics. *Journal of the American Medical Informatics Association* 27(10):1499–1501.

Imler, T. D.; Morea, J.; Kahi, C.; and Imperiale, T. F. 2013. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clinical Gastroenterology and Hepatology* 11(6):689–694.

Karpathy, A.; Johnson, J.; and Fei-Fei, L. 2015. Visualizing and Understanding Recurrent Networks. *arXiv:1506.02078 [cs]*. arXiv: 1506.02078.

Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S. F.; Forshee, R.; Walderhaug, M.; and Botsis, T. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics* 73:14–29.

Kundu, S.; Ashinsky, B. G.; Bouhrara, M.; Dam, E. B.; Demehri, S.; Shifat-E-Rabbi, M.; Spencer, R. G.; Urish, K. L.; and Rohde, G. K. 2020. Enabling early detection of osteoarthritis from presymptomatic cartilage texture maps via transport-based learning. *Proceedings of the National Academy of Sciences* 117(40):24709–24719. Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/117/40/24709.full.pdf.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4):1234–1240.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mittelstadt, B.; Russell, C.; and Wachter, S. 2019. Explaining Explanations in AI. _Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19_ 279–288. arXiv: 1811.01439.

Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; and Eisenstein, J. 2018. Explainable Prediction of Medical Codes from Clinical Text. *arXiv:1802.05695 [cs, stat]*. arXiv: 1802.05695.

Myers, P. D.; Ng, K.; Severson, K.; Kartoun, U.; Dai, W.; Huang, W.; Anderson, F. A.; and Stultz, C. M. 2020. Identifying unreliable predictions in clinical risk models. *NPJ digital medicine* 3(1):1–8.

National Library of Medicine. 2020. Unified medical language system (umls).

Peters, M. E.; Ammar, W.; Bhagavatula, C.; and Power, R. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pons, E.; Braun, L. M.; Hunink, M. M.; and Kors, J. A. 2016. Natural language processing in radiology: A systematic review. *Radiology* 279(2):329–343.

Ravì, D.; Wong, C.; Deligianni, F.; Berthelot, M.; Andreu-Perez, J.; Lo, B.; and Yang, G.-Z. 2016. Deep learning for health informatics. *IEEE journal of biomedical and health informatics* 21(1):4–21.

Rothe, S.; Ebert, S.; and Schtze, H. 2016. Ultradense Word Embeddings by Orthogonal Transformation. *arXiv:1602.07572 [cs]*. arXiv: 1602.07572.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Saib, W.; Sengeh, D.; Dlamini, G.; and Singh, E. 2020. Hierarchical deep learning ensemble to automate the classification of breast cancer pathology reports by icd-o topography. *arXiv preprint arXiv:2008.12571*.

Sheikhalishahi, S.; Miotto, R.; Dudley, J. T.; Lavelli, A.; Rinaldi, F.; and Osmani, V. 2019. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics* 7(2):e12239.

Soldaini, L., and Goharian, N. 2016. Quickumls: A fast, unsupervised approach for medical concept extraction. In *Medical Information Retrieval (MedIR) Workshop, SIGIR*, 1–4.

Stultz, C. M. 2019. The advent of clinically useful deep learning.

Tjoa, E., and Guan, C. 2020. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* 1–21. arXiv: 1907.07374.

Townsend, H. 2013. Natural language processing and clinical outcomes: the promise and progress of nlp for improved care. *Journal of AHIMA* 84(2):44–45.

Triantafyllidis, A. K., and Tsanas, A. 2019. Applications of machine learning in real-life digital health interventions: review of the literature. *Journal of medical Internet research* 21(4):e12286.

Weber, L.; Sänger, M.; Münchmeyer, J.; Habibi, M.; Leser, U.; and Akbik, A. 2020. Hunflair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. *arXiv preprint arXiv:2008.07347*.

Wiedemann, G.; Jindal, R.; and Biemann, C. 2018. microner: A micro-service for german named entity recognition based on bilstm-crf. *arXiv preprint arXiv:1811.02902*.

Wiens, J.; Saria, S.; Sendak, M.; Ghassemi, M.; Liu, V. X.; Doshi-Velez, F.; Jung, K.; Heller, K.; Kale, D.; Saeed, M.; Ossorio, P. N.; Thadaney-Israni, S.; and Goldenberg, A. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine* 25(9):1337–1340.

Yala, A.; Barzilay, R.; Salama, L.; Griffin, M.; Sollender, G.; Bardia, A.; Lehman, C.; Buckley, J. M.; Coopey, S. B.; Polubriaginof, F.; et al. 2017. Using machine learning to parse breast pathology reports. *Breast Cancer Research and Treatment* 161(2):203–211.

Yao, L.; Mao, C.; and Luo, Y. 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC medical informatics and decision making* 19(3):71.

Yoon, H.-J.; Ramanathan, A.; and Tourassi, G. 2016. Multitask deep neural networks for automated extraction of primary site and laterality information from cancer pathology reports. In *INNS Conference on Big Data*, 195–204. Springer.

Zhai, Z.; Nguyen, D. Q.; and Verspoor, K. 2018. Comparing cnn and lstm character-level embeddings in bilstm-crf models for chemical and disease named entity recognition. *arXiv preprint arXiv:1808.08450*.