




# Attention-Based Explainability Approaches in Healthcare Natural Language Processing

Haadia Amjad<sup>1</sup>, Mohammad Shehroz Ashraf<sup>2</sup>, Syed Zoraiz Ali Sherazi<sup>2</sup>, Saad Khan<sup>1</sup>,  
Muhammad Moazam Fraz<sup>1</sup><sup>a</sup>, Tahir Hameed<sup>3</sup><sup>b</sup> and Syed Ahmad Chan Bukhari<sup>4</sup><sup>c</sup>

<sup>1</sup>Department of Computing, National University of Sciences and Technology, Islamabad, Pakistan

<sup>2</sup>Department of Electrical Engineering, National University of Sciences and Technology, Islamabad, Pakistan

<sup>3</sup>Girard School of Business, Merrimack College, North Andover, MA 01845, U.S.A.

<sup>4</sup>Division of Computer Science, Mathematics and Science, St. John's University, Queens, NY 11439, U.S.A.


**Keywords:** Attention Mechanism, Explainability, Natural Language Processing, AI, Healthcare, Clinical Decision Support Systems.


**Abstract:** Artificial intelligence (AI) systems are becoming common for decision support. However, the prevalence of the black-box approach in developing AI systems has been raised as a significant concern. It becomes very crucial to understand how an AI system makes decisions, especially in healthcare, since it directly impacts human life. Clinical decision support systems (CDSS) frequently use Natural Language Processing (NLP) techniques to extract information from textual data including Electronic Health Records (EHRs). In contrast to the prevalent black box approaches, emerging 'Explainability' research has improved our comprehension of the decision-making processes in CDSS using EHR data. Many researches use 'attention' mechanisms and 'graph' techniques to explain the 'causability' of machine learning models for solving text-related problems. In this paper, we conduct a survey of the latest research on explainability and its application in CDSS and healthcare AI systems using NLP. For our work, we searched through medical databases to find explainability components used for NLP tasks in healthcare. We extracted 26 papers that we found relevant for this review based on their main approach to develop explainable NLP models. We excluded some papers since they did not possess components for inherent explainability in architectures or they included explanations directly from the medical experts for the explainability of their work, leaving us with 16 studies in this review. We found attention mechanisms are the most dominant approach for explainability in healthcare AI and CDSS systems. There is an emerging trend using graphing and hybrid techniques for explainability, but most of the projects we studied employed attention mechanisms in different ways. The paper discusses the inner working, merits and issues in the underlying architectures. To the best of our knowledge, this is among the few papers summing up latest explainability research in the healthcare domain mainly to support future work on NLP-based AI models in healthcare.


## 1 INTRODUCTION

The development of AI systems has helped many fields achieve benchmarks of efficiency. AI technologies aim to provide faster and more accurate results significantly improving the decision-making process in many industries. These systems aid professionals in making domain-related decisions in

nearly every field, from home lifestyle to banking, to critical applications such as security and healthcare. Their use in critical domains requires higher accuracy as they directly impact human life. One common issue in AI systems is they lack transparency about inner working of the system. They typically follow a black-box approach which makes it hard for the user to comprehend their reasoning.

<sup>a</sup> <https://orcid.org/0000-0003-0495-463X>

<sup>b</sup> <https://orcid.org/0000-0002-6824-6803>

<sup>c</sup> <https://orcid.org/0000-0002-6517-5261>

AI systems make decisions based on criteria that they can learn from the data and requirements of the system. How these systems arrive at a decision is not inherently interpretable, commonly referred as the black-box approach of the AI (Castelvecchi, 2016). This creates a lack of trust in the system for the user. By not understanding the formation of intelligence of a machine learning model, the user tends not to believe in the correctness of the system (Nourani, King, & Ragan, 2020). While a base understanding of the working of a model could help this issue, the role of the data segments that aided the learning process remains unclear. Explainability is a concept that helps in explaining the AI system to the user in such a manner that they can understand how a model makes a decision (IBM, n.d.). The use of Explainable Artificial Intelligence (xAI) methods aids domain experts in trusting these systems and making better decisions, especially in healthcare AI and CDSS (DeGrave et al., 2021; Savage, 2022).

In the medical domain, all decisions directly impact human life. Any intelligent decision-support system in this field should have high performance and accuracy to be able to help a professional. While the accuracy of the system is gravely important, the understanding of their working directs the confidence of the professional in using these systems (Cabitza et al., 2021). While AI systems are involved in many healthcare decisions and tasks, when it comes down to the types of data being used, the two most typical are medical images and clinical texts. Text data is available in large amounts and formats. It is an interesting task for a system to learn from these documents and notes to make a decision.

Manually handling text data and using it to make decisions in the healthcare domain is a lengthy procedure, requiring massive attention to detail (Wang et al., 2018). For an AI system to achieve this with high accuracy, the model must have sufficient data to train from and enough experimentation to validate its use. The concept of manual attention translates to AI systems in the form of attention mechanism components (Hu, 2020). The attention mechanisms, for text data, focus on the building blocks of these texts, also called tokens. The tokens could be words, phrases or even sentences. Putting such attention on the data helps figure out the importance of those segments that largely influence the system's decision. Talking of manual decision-making, this is how a domain expert would also do it, by extracting the segments that point towards a labelled decision. This attention mechanism aids causable explanations for the AI system.

In our work, we review attention mechanisms more closely for text data that is used in the healthcare domain and summarise how different researchers have made use of this component in AI architectures to introduce explainability into their work. The paper contributes mainly by reviewing latest literature on emerging field of explainability approaches as applied in the healthcare domain and by identifying critical needs of improvement especially a need for novel objective and comparable evaluation metrics and criterion for explainability. In the rest of the paper, we cover our methodology, a thorough literature review, and our systematic observations and findings of our review. We hope this helps guide future researchers who aim to work in the area of explainable AI and trust in AI systems.

## 2 METHODOLOGY

Attention mechanisms have been applied in a variety of domains, including healthcare, resulting in significant research content availability. Thus, it was important to identify the scope of the papers being presented in this review. To start with, the authors jotted down the keywords that closely related to the work being aimed for: 'NLP', 'explainability' and 'clinical text'. These keywords were then used to carry out searches on literary databases, namely PubMed (J) and ScienceDirect (Hunter, 1998). PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics, with ScienceDirect being a website which provides access to a large bibliographic database of scientific and medical publications of the Dutch publisher Elsevier. Furthermore, to ensure the relevancy of the work we opted to include the research that has been carried out between 2018 and 2022 inclusive. Any surveys and papers that included research on medical images were excluded as they did not fit the NLP explainability theme. It resulted in 26 papers that matched the above criteria.

The subset of papers was then passed through a second layer of filtering. Papers in which the concepts of explainability were vague, making it difficult for users to comprehend the work, were excluded. The finalized papers were then peer-reviewed by four members to ensure that the selected papers apply explainable NLP concepts as well as to identify a common methodological trend being followed. In the end, 16 out of 26 papers used attention mechanisms in their architecture to emphasize explainable results. This was concluded by looking at the base

architectures and the accompanying changes that were carried out to incorporate attention. A focus was also placed on the datasets being used, and the goal of the base model on which attention is applied, mainly because base model results determine the type of attention to be used for explainability. Thus, this directed our work to focus on the role of attention in NLP for healthcare systems. The remaining 10 papers did not fall into inherent explainable architectures. They included descriptive explanations from medical professionals and researchers themselves. Hence, we decided not to include those in our review.

### 3 ATTENTION MECHANISM IN HEALTHCARE INFORMATION PROCESSING

In this section, we deeply review the work done in healthcare to attain explainability in NLP problems. To make AI architectures explainable, researchers make additions to existing architectures or use Explainable AI (xAI) models to attain interpretable results.

Naylor et al (2021) used multiple classifiers with varying interpretability and complexity on the MIMIC-III dataset, a dataset including 40,000 records published by Beth Israel Deaconess Medical Center. These classifiers include Logistic Regression (LR), Explainable Boosting Machine (EBM), Random Forest (RF), DL8.5, Boosted Rule Sets, Bayesian Rule Lists, Optimal Classification Trees, Certifiably Optimal Rule ListS (CORELS) and BigBird. For LR, RF, and EBM, they use Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) for explainability. For other models, they considered techniques such as saliency and integrated gradients. To conclude their work, they used the Local Lipschitz metric for evaluation. They concluded that a trade-off may exist between model quality and interpretability for models not typically used for text classification. Xu et al. (2018) predict ICD-10 diagnostic codes using a multimodal architecture making an ensemble of Text-CNN, Decision Tree and a combination of Char-CNN and Bidirectional LSTM. For their interpretability section, they extracted training parts and calculated an influence score for text features. They employed LIME for tabular features. Their interpretability results using Jaccard Similarity Coefficient (JSC) are 0.1806 on text data and 0.3105 on tabular data. For most work in this area, researchers use attention mechanism and knowledge

graphs that have been covered in the following paragraphs.

#### 3.1 Transformer-Based Architectures

Attention mechanisms are used for selective concentration on feature extractions in deep learning networks. The following studies induce novelty in transformer-based architectures using attention layers. Liu et al. (2022) introduced a novel architecture in their work known as Hierarchical Label-wise Attention Transformer Model (HiLAT). They made ICD-9 code predictions on clinical documents using HiLAT with an explainability component. HiLAT uses a two-level label-wise attention mechanism in a hierarchical manner that produces label-specific representations of the document. They also employed XLNet-Base on clinical notes to develop ClinicalplusXLNet. HiLAT + ClinicalplusXLNet attained an F1 score of 73.5.

Mayya et al (2021) developed LATA, a neural model built upon Label Attention Transformer Architectures. The model comprises of multiple encoder layers based on BERT model. The target automatic assignment on ICD-10 codes using CodiEsp dataset. Their model outperformed BERT by 33-49% in precision, recall and F1 scores. The label attention mechanism in LATA introduced the explainability components in prediction. Feng et al. (2020) proposed a novel architecture pipeline to predict ICD-9 codes with an explainability component. They used MIMIC-III discharge summaries for testing their model. Their architecture extracted sentences from documents and tokenized them before passing them through BERT feature extractor. This process generated word embeddings that were input to a base Text-CNN resulting in sentence features that were fed into a transformer. The encoded features from the transformer were passed on to a label-wise attention layer. The attention representations were ultimately given to a classifier for multi-label code predictions. Biswas et al. (2021) also predicted explainable ICD-9 codes with a transformer-based architecture combined with a code-wise attention mechanism. They put deeper attention to tokens of the clinical documents and extracted code-specific representations. They achieve a micro-AUC score of 0.923. Nguyen et al. (2021) experimented with biomedical text classification and pre-trained language models with label attention. They developed a transformer-based architecture with label attention for text classification. The label attention mechanism was used in the fine-tuning

process of pre-trained language models. The base architecture used in this work is pre-trained BERT.

### 3.2 Convolutional Neural Networks

Some studies made use of traditional CNN models and placed attention layers in the architecture to achieve interpretability in multi-label classifiers. Feucht et al. (2021) predicted explainable ICD-9 codes using description-based label attention. Their model learnt code embeddings by integrating code descriptions and then later assigning those embeddings to text representation which in turn resulted in a label-wise attention mechanism. Trigueros et al. (2022) developed a multi-label classifier for electronic health records in Spanish with convolutional attention. They used the OSA Spanish dataset and the MIMIC-III dataset. They compared the results of a baseline CNN, attentive CNN and regularised CNN. Hu et al. (2020) proposed Shallow and Wide Attention convolutional Mechanism (SWAM) for explainable model code prediction. They used convolution layers to extract informative snippets and important features, and the model then selected the snippets of each code using an attention mechanism and ultimately made code predictions.

The base architecture used in many of the above-noted architectures is the Convolutional Attention Mechanism for Multi-Label classification (CAML) presented by Mullenbach et al. (2018). They used a label-wise attention mechanism on max pooling layers in a convolutional neural network for predicting ICD-9 codes using the MIMIC-III dataset. Mayya et al. (2021) designed a multi-channel convolutional neural network with an attention mechanism for medical code prediction on unstructured discharge summaries, known as Enhanced Convolutional Attention Network for Multi-Label classification (EnCAML). They experimented using MIMIC-III and CodiEsp datasets. The text embeddings pass through convolutional layers of different kernel sizes followed by individual attention mechanisms concatenated and followed by dense layers ultimately predicting ICD-9 code prediction groups. The explainability metric is evaluated using the Jaccard score.

### 3.3 Long Short-Term Memory (LSTM) and Gated Residual Network (GRN)

The use of attention layers can be said to achieve different results based on their type and placement in

the architecture, as seen from the work mentioned thus far. Researchers have also used multiple attention mechanisms for the same purpose of interpretability with variants of Recurrent Neural Networks (RNNs) the works we mention going forward. Wang et al. (2021) created a novel architecture in their work using multi-head attention with the gated residual network to predict ICD-10 codes. They used Chinese electronic health records as their dataset. In their architecture, the data embeddings passed through a gated residual network for supervised predictions and employed dot and additive attention mechanisms on dilated convolutional layers for explainability. They achieved an F-1 score of 92.11.

In the work of Dong et al. (2021), Hierarchical Label-wise Attention (HLAN) is introduced with the concept of quantifying the importance of words using attention weights. Their architecture includes two bi-directional gated residual units, one with word-wise attention and the second with label-wise attention. They tested it on the MIMIC-III dataset and achieved an F1 score of 64.1. Li et al. (2021) introduced JLAN, a joint learning architecture of attention and denoising mechanisms for ICD-9 codes using the MIMIC-III dataset. Their approach tended to reduce the noise at code allocation along with attention-based architectures. Specifically, they employed self-attention and label-attention mechanisms.

### 3.4 Multimodal Architectures including Knowledge Graphs

Knowledge graphs (KGs) organise data from multiple sources, capture information about entities of interest in a given domain or task and forge connections between them (The Alan Turing Institute, n.d.).

Knowledge Graphs are considered to be inherently explainable. For this reason, they have been adapted to many explainability tasks specifically in the healthcare domain. Teng et al. (2020) predicted ICD-9 codes using knowledge graphs, adversarial learning and an attention mechanism. They introduced G-Coder, which takes two inputs, a knowledge graph and medical documents. The medical documents are processed with adversarial samples resulting in sentence encodings and the knowledge graph results in graph embeddings. Both of these embeddings are passed through multiplied attention mechanism which then passes down a classifier to predict ICD codes. The architecture achieves an F1 score of 69.92.

## 4 FINDINGS AND OBSERVATIONS

We have observed that research work done so far on text-related medical tasks using attention mechanisms follows a standard pipeline (Figure 1). The text-based data is converted into embeddings. Then a combination of convolutional layers is used along with an attention layer. This attention layer may be token-wise, word-wise, label-wise or sentence-wise. Putting attention on the embeddings or the results of convolutional layers assists in interpreting the parts of the documents that influence the prediction or classification of labels.

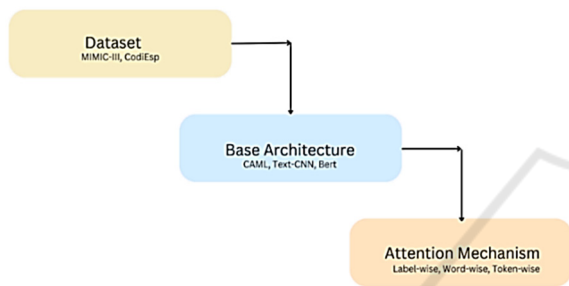


Figure 1: A typical pipeline for NLP in healthcare using an attention mechanism; key differences in the base architecture.

The models discussed above primarily use two benchmark datasets, MIMIC-III and CodiEsp and are compared based on their F1 score. The base architecture for these research works is primarily Text-CNN, BERT or CAML. From these base models, CAML comes with an inherent attention layer while Text-CNN and BERT are used due to their performance and increased usage across various tasks. These models also vary based on their layer assembly and learning techniques, such as joint, adversarial or others. (See Table 1).

While some of these works include an evaluation for the explainable components, such as by using Jaccard scores, these papers follow a pattern of depicting their results using graphs containing either highlighted text visualisation over documents, heat maps or other graphs. The idea is to depict the local features identified by different attention mechanisms to emphasise the explainability of their models. (Table 2).

Table 2: Evaluation techniques used for explainability components in different research work.

Reference	Explainability Technique	Explainability Metric/ Evaluation Criteria
Xu et al. (2019)	LIME, Influence scores	Jaccard Similarity Coefficient (0.1806 on text and 0.3105 on tabular data)
LATA	Layer-wise attention	Jaccard Similarity Coefficient
HiLAT + Clinicalplu sXLNet	Label-wise attention	Visualization of attention weights
DLAC	Description-based label attention	Visualization of attention weights
Feng et al. (2020)	Label-wise attention	Medical Professional + Novel approach finding attention weights

Specifically, for the use of attention mechanisms for explainability, researchers have experimented with types of attention, such as dot attention or concatenation, the focus of attention, such as label or word, their placement in the pipeline and their occurrence. These innovations combined with base architectures, techniques of embeddings and architectural variations have resulted in achievements in explainability for text-related tasks in the healthcare domain.

We must highlight that the evaluation of explainability components in all the work we have reviewed is not directly comparable. There has been no benchmark against which to measure the correctness of the explainable results. Hence, we cannot rate one above another. For aspiring researchers in this direction, it is important to include a detail of their explainability evaluation either with a comparison of the work included in this paper or using metrics such as the Jaccard Similarity Coefficient or document similarity index. We can, however, assume that a model with the highest F1 score, or another performance metric, would have picked the optimal features from the text data. Hence, we believe the attention scores that are highlighted in a visualization using the attention mechanism of the model explain the results of the model better.

Table 1: Comparison of the attention mechanisms, base architectures, datasets and their performance.

Research Work	Model	Dataset	Base Architecture	Attention Mechanism	Results (F1-Score)
Liu et al (2022)	HiLAT + ClinicalplusXLNet	MIMIC-III [28]			
XLNet, BERT	Label-wise attention	73.5			
Feucht et al (2021)	DLAC	MIMIC-III	BERT, CNN	Description-based label attention	62.2
Mayya et al (2021)	LATA	CodiEsp	BERT	Label-wise attention	64.3
Wang et al (2021)	MA-GRN	Chinese electronic health records	GRN	Self-attention (dot and additive), Label-wise attention	92.11
Li et al (2021)	JLAN	MIMIC-III	LSTM	Self-attention, Label-wise attention	66.5
Mayya et al (2021)	EnCAML	MIMIC-III, CodiEsp	CAML	Layer-wise attention	85.59
Dong et al (2021)	HLAN	MIMIC-III	GRN	Label-wise attention, Word-wise attention	64.1
Mullenbach et al (2018)	CAML	MIMIC-III	CNN	Label-wise attention	88.0
Feng et al (2020)	-	MIMIC-III	BERT, Text-CNN	Label-wise attention	78.9
Biswas et al (2021)	TransICD	MIMIC-III	Transformer	Code-wise attention	56.2
Nguyen et al (2021)	LAME	Hallmarks of Cancers (HoC), Medical abstracts of diseases (Disease5)	BERT	Label-wise attention	83.3
Trigueros et al (2022)	-	Osa, MIMIC-III	CNN	Layer-wise attention	64.4
Hu et al (2021)	SWAM	MIMIC-III	CAML	Label-wise attention	59.3

\*macro: computed using the arithmetic mean of all the per-class F1 scores

## 5 CONCLUSIONS

In our work, we have reviewed the cutting-edge applications of the explainability components in NLP tasks in the healthcare domain. Based on our review and analysis of the work done from 2018 to 2022, we have found that around 60% of the work uses attention mechanisms. There was some research on graphs and hybrid approaches but attention remains the dominant one.

All the studies and projects included in this review involved NLP tasks focused on CDSS and healthcare applications and they shared some commonalities. All of them used a base architecture for text label prediction. This architecture was then modified with attention layers, of various types and in different

numbers, to add explainability component. Then, they created visualization of the attention results using most suitable techniques that could effectively demonstrate the results of these attention mechanisms. It is important to note the researches included in this study review a similar overall three-set pipeline (See Fig. 1). However, the difference and novelty lies in the base architectures and these projects have carefully designed architectural pipelines that help them strive for better performance and results while addressing specific issues. Hence, these base architectures are the source of performance differences and which architectures may be suitable for specific use cases.

We have summarized the contributions of researchers and highlighted the base architectures

they used. We have also detailed how these architectures vary in the use of attention mechanisms. We have found that CAML, Text-CNN and BERT are the most commonly used base architectures, while label-wise and self-attention are the most common mechanisms to develop such novel architectures. Many of these studies build on architectures from previous similar studies to introduce a pipeline with better results in the area. We can also see the popular use of attention mechanisms with transformer models, CNNs and even RNNs. However, all of these researches are focused on evaluation of label predictions and visualization of attention scores. In order to develop plausible interpretability, a standard set of explainability evaluations should be performed. We recognize an immediate need to not only consider explainability that is much more relatable to end-users' cognition and learning, but is also objective in terms of being measurable and comparable. Such evaluation measures combined with overall performance of the NLP and AI models in the healthcare domain would aid the readers as well as the end-users in understanding the outputs better, hence working with these AI models and systems with higher confidence and trust.

## REFERENCES

- Biswas, B., Pham, T. H., & Zhang, P. (2021, June). *Transicd: Transformer based code-wise attention model for explainable icd coding*. In International Conference on Artificial Intelligence in Medicine (pp. 469-478). Springer, Cham.
- Cabitza, F., Campagner, A., & Datteri, E. (2021). To err is (only) human. Reflections on how to move from accuracy to trust for medical AI. In *Exploring Innovation in a Digital World* (pp. 36-49). Springer, Cham.
- Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature News*, 538(7623), 20.
- DeGrave, A. J., Janizek, J. D., & Lee, S. I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7), 610-619
- Dong, H., Suárez-Paniagua, V., Whiteley, W., & Wu, H. (2021). Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116, 103728.
- Feng, J., Shaib, C., & Rudzicz, F. (2020, November). *Explainable clinical decision support from text*. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 1478-1489).
- Feucht, M., Wu, Z., Althammer, S., & Tresp, V. (2021). *Description-based Label Attention Classifier for Explainable ICD-9 Classification*. arXiv preprint arXiv:2109.12026.
- IBM. Explainable AI (XAI). <https://www.ibm.com/watson/explainable-ai>. Accessed [Nov, 2022]
- Hu, D. (2019, September). *An introductory survey on attention mechanisms in NLP problems*. In Proceedings of SAI Intelligent Systems Conference (pp. 432-448). Springer, Cham.
- Hu, S., Teng, F., Huang, L., Yan, J., & Zhang, H. (2021). An explainable CNN approach for medical codes prediction from clinical text. *BMC Medical Informatics and Decision Making*, 21(9), 1-12.
- Hunter, K. (1998). ScienceDirect™. *The Serials Librarian*, 33(3-4), 287-297.
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1), 1-9.
- Li, X., Zhang, Y., Dong, D., Wei, H., & Lu, M. (2021). JLAN: medical code prediction via joint learning attention networks and denoising mechanism. *BMC bioinformatics*, 22(1), 1-21.
- Liu, L., Perez-Concha, O., Nguyen, A., Bennett, V., & Jorm, L. (2022). *Hierarchical Label-wise Attention Transformer Model for Explainable ICD Coding*. arXiv preprint arXiv:2204.10716.
- Mayya, V., Kamath, S. S., & Sugumaran, V. (2021, October). *Label Attention Transformer Architectures for ICD-10 Coding of Unstructured Clinical Notes*. In 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (pp. 1-7). IEEE.
- Mayya, V., Kamath, S., Krishnan, G. S., & Gangavarapu, T. (2021). Multi-channel, convolutional attention based neural model for automated diagnostic coding of unstructured patient discharge summaries. *Future Generation Computer Systems*, 118, 374-391.
- Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., & Eisenstein, J. (2018). *Explainable prediction of medical codes from clinical text*. arXiv preprint arXiv:1802.05695.
- Naylor, M., French, C., Terker, S., & Kamath, U. (2021). *Quantifying explainability in nlp and analyzing algorithms for performance-explainability tradeoff*. arXiv preprint arXiv:2107.05693.
- Nguyen, B., & Ji, S. (2021). *Fine-tuning Pretrained Language Models with Label Attention for Explainable Biomedical Text Classification*. arXiv preprint arXiv:2108.11809.
- Nourani, M., King, J., & Ragan, E. (2020, October). *The role of domain expertise in user trust and the impact of first impressions with intelligent systems*. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (Vol. 8, pp. 112-121).
- Savage, N. (2022). Breaking into the black box of artificial intelligence. *Nature*, March 2022 <https://doi.org/10.1038/d41586-022-00858-1>

- The Alan Turing Institute. Knowledge Graphs, How do we encode knowledge to use at scale in open evolving systems, decentralised systems? <https://www.turing.ac.uk/research/interest-groups/knowledge-graphs>. Accessed [Nov, 2022]
- Teng, F., Yang, W., Chen, L., Huang, L., & Xu, Q. (2020). Explainable prediction of medical codes with knowledge graphs. *Frontiers in Bioengineering and Biotechnology*, 8, 867.
- Trigueros, O., Blanco, A., Lebeña, N., Casillas, A., & Pérez, A. (2022). Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention. *International Journal of Medical Informatics*, 157, 104615.
- Wang, X., Han, J., Li, B., Pan, X., & Xu, H. (2021, December). *Automatic ICD-10 Coding Based on Multi-Head Attention Mechanism and Gated Residual Network*. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 536-543). IEEE.
- Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Change*, 126, 3-13.
- White, J. (2020). PubMed 2.0. *Medical Reference Services Quarterly*, 39(4), 382-387.
- Xu, K., Lam, M., Pang, J., Gao, X., Band, C., Mathur, P., & Xing, E. P. (2019, October). *Multimodal machine learning for automated ICD coding*. In Machine learning for healthcare conference (pp. 197-215). PMLR.

