Exploring Explainable AI Techniques for Text Classification in Healthcare: A Scoping Review

Ibrahim Alaa Eddine Madi, Akram Redjdal, Jacques Bouaud, Brigitte Seroussi

Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240544

Exploring Explainable AI Techniques for Text Classification in Healthcare: A Scoping Review

Ibrahim Alaa Eddine MADI[a,1], Akram REDJDAL[a], Jacques BOUAUD[a] and Brigitte SEROUSSI[a,b,c]
[a] *Sorbonne Université, Université Sorbonne Paris Nord, INSERM, Laboratoire d'Informatique Médicale et d'Ingénierie des connaissances en e-Santé, LIMICS, Paris, France*
[b] *AP-HP, Hôpital Tenon, Paris, France*
[c] *APREC, Paris, France*
ORCiD ID: Ibrahim Alaa eddine Madi https://orcid.org/0009-0000-4814-9726.

Abstract. Text classification plays an essential role in the medical domain by organizing and categorizing vast amounts of textual data through machine learning (ML) and deep learning (DL). The adoption of Artificial Intelligence (AI) technologies in healthcare has raised concerns about the interpretability of AI models, often perceived as "black boxes." Explainable AI (XAI) techniques aim to mitigate this issue by elucidating AI model decision-making process. In this paper, we present a scoping review exploring the application of different XAI techniques in medical text classification, identifying two main types: model-specific and model-agnostic methods. Despite some positive feedback from developers, formal evaluations with medical end users of these techniques remain limited. The review highlights the necessity for further research in XAI to enhance trust and transparency in AI-driven decision-making processes in healthcare.

Keywords. Artificial intelligence, Healthcare, Explainable AI (XAI), Text classification, Interpretability, Scoping review

1. Introduction

About 80% of hospital data is acquired in textual format [1]. Documents, such as discharge summaries and clinical notes, are a valuable source of information to optimize the clinical management of patients. However, the datafication of care allows to produce many documents, making access to the relevant information a complex task. Text classification methods allowing for the organization and categorization of medical information, helps information retrieval and analysis, thus improves the clinical decision-making process and advances medical research [2]. The evolution of data classification methods in healthcare, and especially text classification, has been notable, transitioning from rule-based systems relying on handcrafted rules to the adoption of Artificial

[1] Corresponding Author: Ibrahim Alaa eddine Madi; E-mail: madiibrahim01@gmail.com.

Intelligence (AI) technologies, including machine learning (ML) and deep learning (DL) algorithms, enabling the handling of large and complex datasets with heightened accuracy and efficiency [3]. However, AI systems are often perceived as "black boxes" that offer little insight into the reasoning behind their predictions, leading to hesitancy and low acceptance among healthcare practitioners. Therefore, current research works on explainable artificial intelligence (XAI) are conducted to provide interpretability of AI models to make them more useful, especially in the healthcare domain [4].

Two distinct approaches have been developed to interpret ML and DL models: model-specific methods tailored to the AI model structure and features, and model-agnostic methods explaining AI predictions regardless of how the AI model is working [5]. Two primary XAI methods have gained widespread adoption among researchers: LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME operates by generating local interpretable models, so-called White Box Models [4] (e.g., linear regression), to understand why the complex model made certain predictions for specific data instances. For instance, with a DL model trained to diagnose patient conditions from text reports, for a patient with an unusual condition, LIME allows to modify the report through slight changes. Based on these changes, LIME compute coefficients to represent the importance of each word or phrase in the DL model prediction allowing the model simplification to approximate the local decision boundary around the unusual patient condition in the report. On the other hand, SHAP provides insights to the importance of each feature of the DL prediction model, by calculating SHAPley values [4] representing the marginal contributions of the model features. For instance, in a model predicting patient mortality based on age and medical history, SHAP would determine the contribution of each feature as explanations of the model prediction.

In the perspective of explaining the identification of complex breast cancer clinical cases from tumor board reports with the system Oncolog-IA [6], we conducted a scoping review to explore and evaluate the predominant XAI techniques utilized in clinical text classification, focusing on whether model-agnostic or specific methods are favored. Additionally, we sought to verify if widely adopted techniques in other domains, such as SHAP and LIME, are similarly prevalent in clinical text classification. Our investigation extended to examining whether the effectiveness of XAI was evaluated by clinicians.

2. Methods

We performed a literature search, using PubMed to identify articles published between January 2015 and March 2024, focusing on XAI methods applied to medical text classification tasks. We used the following query*: (("XAI") OR (("Trustworthy" OR "Explainable" OR "Interpretable" OR "Transparent" OR "Explainability") AND ("AI" OR "Artificial intelligence" OR "deep learning" OR "neural networks" OR "Machine Learning" OR "Algorithm" OR "Model"))) AND ("Classification" OR "Diagnostic Prediction" OR "Risk Assessment" OR "Evaluation") AND ("Clinical Notes" OR "EHR" OR " Electronic Health Record" OR "Operative Notes" OR "Clinical Narratives" OR "Text") NOT ("Image" OR "signal" OR "ECG") NOT (Review OR Survey*).* The search results were refined by applying specific exclusion and inclusion criteria (mentioned in Figure 1), retaining articles explicitly mentioning XAI methods in their titles or abstracts and verified on their full text as relevant to text-based healthcare data analysis. Selected papers were analyzed to explore the different XAI methods used based on diverse types

of ML and DL models. The assessment of XAI techniques performance involving experts of the medical field was studied.

3. Results

The results of the literature search are displayed in Figure 1. From 148 articles initially retrieved, 40 articles were selected from title and abstract and 16 of them were selected based on the full text review. One article was included based on references cited within the selected literature leading to a total of 17 articles at the end.

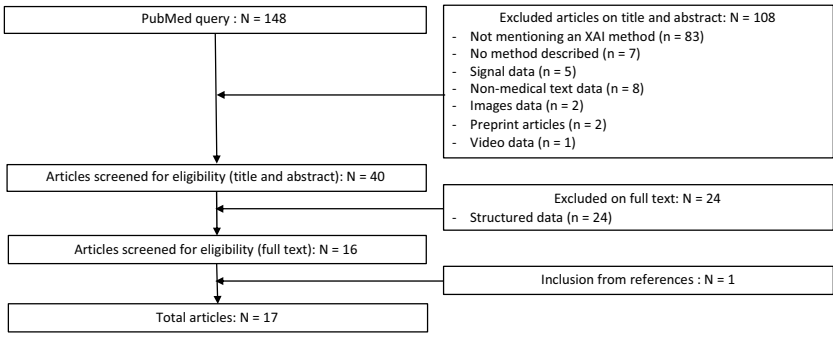

Figure 1. Flowchart outlining the process conducted for the selection of studies articles.

These 17 articles explore nine XAI methods of medical text classification, split between five model-agnostic and four model-specific techniques. LIME is the most prevalent, used in seven studies, while attention mechanisms and SHAP are behind with five and three articles, respectively. The classification tasks varied, ranging from predicting medical conditions and treatment outcomes to medication detection and International Classification of Diseases (ICD) codes classification from clinical notes. Interestingly ICD coding task was the most frequently represented. The predominant models paired with XAI techniques were transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT), indicating a trend towards advanced neural network architectures.Model-agnostic methods, especially LIME and SHAP, showed broad applicability across different medical applications, highlighting their versatility. Conversely, model-specific techniques such as attention mechanisms and Grad-CAM were primarily applied to DL models like transformers, Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs). An exception was TreeExplainer, a model-specific method for explaining tree-based models like Extreme Gradient Boosting (XGBoost).

Across the different studies, only five articles included formal assessments of XAI techniques' performance involving medical field experts (highlighted in **bold** in Table 1), primarily using LIME and SHAP, where the authors concluded that these techniques generally align well with medical domain knowledge, as indicated in studies #1 and #6.

In study #3, LIME showed significant overlaps with expert-identified text segments. In study #13, SHAP demonstrated a Jaccard similarity of 72% compared to explanations provided by medical professionals. Additionally, in study #8, the sparse attention mechanism's explanations were found to be relevant, with an accuracy of around 60%.

4. Discussion

In various studies about XAI outside the healthcare domain, SHAP and LIME have consistently emerged as the most used methods for enhancing AI model's explanation capacities [5]. However, according to the studied articles in this scoping review, concerning text classification in the medical domain, LIME appears to be preferred as compared to SHAP.

Table 1 : Comprehensive Overview of Explainable AI Techniques

ID	PMID	Classification Task	XAI Technique	XAI-Type	Model used
1	**37046469**	Classification of MRI scans reports for multiple sclerosis	LIME, SHAP, Integrated Gradients	MA	TMB
2	38271086	Classification of PSE reports	LIME	MA	SVM with Roberta base
3	**28506904**	Prediction of ordinal symptom severity scores	LIME	MA	CNN
4	36996118	Classification of depression from speech responses	LIME	MA	TBM
5	33534720	Disease-treatment classification	LIME, BioCIE	MA	TBM, SVM, LSTM
6	**35811026**	Prediction of assertion types of medical concepts in clinical notes	LIME	MA	TBM
7	35673093	Keyword-based text summarization of nursing entries from EHRs	LIME	MA	bidirectional LSTM-based neural network
8	**34741890**	ICD coding	Attention_mechanism	MS	CNN
9	35995108	ICD coding	Attention_mechanism	MS	TBM
10	33711543	ICD coding	Attention_mechanism	MS	Hierarchical Label-wise Attention Network
11	34789241	ICD coding	Attention_mechanism	MS	CNN
12	37030658	Medication detection and medication change event extraction	Attention_mechanism	MS	TBM
13	**38096637**	Identifying medical symptoms	SHAP	MA	TBM
14	37915208	Predicting the likelihood of advanced epithelial ovarian cancer	SHAP, TF-IDF	MA	TBM and XGBoost
15	37760173	Classification for patient condition diagnosis	Grad-CAM	MS	ResNet, CNN, and Bi-LSTM
16	36660449	Prediction of surgical misadventures from operative notes	Red-Flag/Blue-Flag	MS	Text - CNN
17	36787990	Identification of patients with hypertension	TreeExplainer	MS	XGBoost

TBM: Transformer-based models. **MS**: Model-specific. **MA**: Model-Agnostic. **MLP:** Multi-Layer Perceptron. **SVM:** Support Vector Machine. **CNN:** Convolutional Neural Network. **LSTM:** Long Short-Term Memory. **TF-IDF:** Term Frequency Inverse Document Frequency. **LR:** Logistic Regression. **RF:** Random Forest. **ICD:** International Classification of Diseases

Interestingly, attention mechanism is gaining adoption, particularly in tasks like ICD coding, outpacing SHAP in certain instances. This may be attributed to the rise of transformer-based models, such asBERT [7], which intrinsically utilize attention mechanisms to discern the relevance and context of each input element, enhancing model transparency and interpretability. The integration of attention mechanisms in these

advanced models allows for dynamic analysis, where terms like 'chest pain,' and 'shortness of breath,' are highlighted, aiding in understanding a model's prediction of ICD codes related to heart disease. This aligns with the increasing adoption of large language models in medical text analysis, underscoring a growing need for interpretability within such sophisticated AI systems. Despite the positive reception of XAI techniques allowing for the delivery of insights about AI predictions, there remains a gap in formal evaluation, particularly involving end-users from the medical sector. Only five studies (#1, #3, #6, #8 and #13 in table 1) included a formal evaluation of XAI methods, with four of them using SHAP or LIME. Conversely, many studies focusing on developing novel XAI techniques did not undergo evaluation with end-users from the healthcare domain. Studies could benefit from involving medical professionals in the evaluation process to ensure that XAI methods not only provide technical explanations but also align with clinical reasoning semantics. This approach could enhance application of XAI in healthcare, fostering trust and transparency of AI systems.

5. Conclusions

The exploration of XAI with text classification models reveals the prevalence of both model-specific and model-agnostic approaches. Attention mechanisms dominate as model-specific approaches, especially with studies using neural networks, and LIME as a model-agnostic method. Despite some positive feedbacks, formal evaluations involving medical workers are limited. Nonetheless, studies like Lee et al [8] (#13), demonstrate the potential of XAI to align with expert medical knowledge, underscoring the importance of further research in this area to enhance trust and transparency in AI-driven decision-making processes.

References

[1] Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Jt Summits Transl Sci Proc. 2014 Apr 7;2014:218-23. PMID: 25717416.

[2] Chaib R, Azizi N, Hammami NE, Gasmi I, Schwab D, and Chaib A, 'GL-LSTM Model For Multi-label Text Classification Of Cardiovascular Disease Reports', in 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Mar. 2022, pp. 1–6. doi: 10.1109/IRASET52964.2022.9738147.

[3] Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. 2018 Jul;1:53-68. doi: 10.1146/annurev-biodatasci-080917-013315. Epub 2018 May 23. PMID: 31218278

[4] Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine. 2022 Nov 1;226:107161.

[5] Dwivedi R, et al. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys. 2023 Jan 13;55(9):1-33. doi: 10.1145/3561048.

[6] Redjdal A, Bouaud J, Gligorov J, Séroussi B. Using Machine Learning and Deep Learning Methods to Predict the Complexity of Breast Cancer Cases. Stud Health Technol Inform. 2022 May 25;294:78-82. doi: 10.3233/SHTI220400. PMID: 35612020.

[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.

[8] Lee S, Lee J, Park J, Park J, Kim D, Lee J, Oh J. 'Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage', Am. J. Emerg. Med., vol. 77, pp. 29–38, Dec. 2023, doi: 10.1016/j.ajem.2023.11.063.