Semantic Enrichment of Explanations of AI Models for Healthcare

Luca Corbucci[1](✉), Anna Monreale[1], Cecilia Panigutti[1], Michela Natilli[2], Simona Smiraglio[1], and Dino Pedreschi[1]

[1] Department of Computer Science, University of Pisa, Pisa, Italy
luca.corbucci@phd.unipi.it
[2] ISTI-CNR Pisa, Pisa, Italy

Abstract. Explaining AI-based clinical decision support systems is crucial to enhancing clinician trust in those powerful systems. Unfortunately, current explanations provided by eXplainable Artificial Intelligence techniques are not easily understandable by experts outside of AI. As a consequence, the enrichment of explanations with relevant clinical information concerning the health status of a patient is fundamental to increasing human experts' ability to assess the reliability of AI decisions. Therefore, in this paper, we propose a methodology to enable clinical reasoning by semantically enriching AI explanations. Starting with a medical AI explanation based only on the input features provided to the algorithm, our methodology leverages medical ontologies and NLP embedding techniques to link relevant information present in the patient's clinical notes to the original explanation. Our experiments, involving a human expert, highlight promising performance in correctly identifying relevant information about the diseases of the patients.

1 Introduction

Recent efforts in Artificial Intelligence (AI) have shown great potential in helping physicians in several of their daily clinical practices, for example, the interpretation of medical scans [30] and the accurate assessment of prognosis [9] and treatment recommendation [5]. While some worries have been raised about AI systems replacing the role of doctors, human reasoning and oversight remain indispensable for the proper functioning of such systems [10]. Indeed, current AI applications focus on narrow tasks and have been shown to be sensitive to adversarial attacks [23] and biased datasets and algorithms [26]. These shortcomings raised several concerns about the trustworthiness of such systems, especially because most state-of-the-art AI-based solutions are hardly interpretable by humans. The transparency of AI systems in high-stakes domains such as healthcare has been subject to many recent European regulatory efforts like the GDPR and the recent proposal to regulate AI (AI Act).

For example, the European General Data Protection Regulation (GDPR), which came into full effect in May of 2018, prescribes providing the data subject of any automated decision-making process with "meaningful information about

A. Bifet et al. (Eds.): DS 2023, LNAI 14276, pp. 216–229, 2023.
https://doi.org/10.1007/978-3-031-45275-8_15

the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" [14]. Furthermore, the recent proposal for a European regulation of AI (AI Act) prescribes high-risk AI systems to be developed in such a way that they enable users to interpret their output correctly and use them appropriately [15]. In response to these ethical and legal issues, in the past years, the research community has been very active in developing several techniques to explain the reasoning of *black box* AI models, i.e., models whose internal decision-making process is obscure. The research field that studies the interpretability of AI systems is that of eXplainable Artificial Intelligence (XAI) [4]. Most XAI techniques offer interpretations of the black box behaviour by providing *explanations*, i.e., interfaces between humans and algorithms that allow the user to understand the AI decision-making process. Developing AI systems able to support medical decision-making requires creating appropriate human-computer interfaces to enable clinical reasoning. However, most XAI explanations are designed to provide insights on model behaviour to AI developers [3].

In this paper, we present a novel methodology that exploits access to the patient's clinical notes and the domain knowledge encoded in medical ontologies to semantically enrich the explanations provided by a state-of-the-art XAI technique for clinical decision support systems (DSS). While the original explanation considers only patient features that the AI algorithm received as input, our methodology exploits medical ontologies to link such features to an external source of knowledge on the patient. The result is an augmented explanation that allows the physician to reason over the clinical context. Our experiments, involving a human expert, show promising performance in correctly identifying relevant information about the diseases of the patients.

The paper is structured as follows. In Sect. 2 we briefly present the field of XAI, its applications in the healthcare context and the related uses of ontologies. In Sect. 3 we formalize the problem we address in the paper, while in Sect. 4 we describe the details of our methodology. Section 5 presents the experiments used to validate our methodology. Finally, in Sect. 6 we discuss our results and we present our ideas for future developments of our methodology.

2 Related Work

In this section, we overview some research work linked to our methodology.

XAI in Healthcare. XAI research studies how to provide explanations for AI systems behaviour in *human-understandable* terms [4]. The need for XAI techniques stems from the fact that many AI systems have an opaque internal reasoning process, i.e., they are considered black boxes. In the literature, the transparency of AI systems is achieved mainly in two ways: by building *transparent-by-design* models and by extracting explanations from black box models [16]. Some examples of transparent-by-design models employed in healthcare are models that allow the visualization of the relationships between input features and model output [6] and case-based reasoning models where the decision-making process is entirely interpretable [2]. However, it is not always possible to build transparent-by-design models for the task at hand. Therefore it might be necessary to extract explanations from black box models. The two most known examples of such XAI techniques are LIME [27] and SHAP [22]. While LIME trains a local linear model on a feature space neighbourhood of the data point to be explained and uses its weights as a local explanation for the model classification, SHAP assigns to each feature an importance value using a game theory approach. In the healthcare field, an example of an explainer is MARLENA [25], a model-agnostic solution to explain classifiers that perform multi-label tasks such as multi-morbidity classification or unknown genes functional expressions. Another example is Doctor XAI [24], the XAI algorithm employed in our experiments which we detail in the next paragraph. However, none of these works really takes into consideration end-user needs and domain expertise in the design of their explanations.

Some examples of transparent-by-design models employed in healthcare are the ones presented in [6] and [2]. In [6], the authors use Generalized Additive Models (GAM) with pairwise interactions to predict the probability of 30-days readmission to the hospital and the probability of death from pneumonia. GAM allows the visualization of the relationships between single and pairs of input features with the output, enabling the user to inspect what the model has learned. In [2], the authors develop a case-based interpretable Deep Learning model to classify mass lesions in mammographies. The case-based reasoning, highlighting the classification-relevant parts of the image used to make the decision, makes the model interpretable. However, it is not always possible to build transparent-by-design models for the task at hand. Therefore it might be necessary to extract explanations from black box models. The type of XAI technique that we employ in this paper is post-hoc and model-agnostic. Post-hoc XAI techniques extract explanations from trained models, and model-agnostic ones can extract such explanations from any type of black box model because they do not use any of its internal parameters in the explanation extraction process.

This kind of XAI techniques are agnostic w.r.t. the black box model, however, they are not agnostic w.r.t. the type of input and output data processed by the model. Therefore, they are considered specific for healthcare when they are able to deal with the peculiarities of healthcare data.

Doctor AI and Doctor XAI. In this paper we semantically enrich the explanations of Doctor XAI, which is a post-hoc model-agnostic XAI technique able to deal with multi-label classification tasks and ontology-linked sequential data. Doctor XAI exploits medical ontologies in its explanation extraction process. Once the user selects one data point whose outcome needs an explanation, Doctor XAI first finds a set of semantically close neighbours of that data point from a set of available instances by employing an ontological distance metric. Then, it augments such neighbourhood by *ontologically perturbing* the neighbour's data points, i.e. it masks ontologically similar features and queries the black box on such perturbed data points. Finally, it learns a multi-label decision tree on such an augmented neighbourhood and extracts an explanation from it in the form of a decision rule matching the decision path for that data point on the tree.

Doctor XAI is used to explain the outcomes of Doctor AI [12], a Recurrent Neural Network (RNN) trained on the sequential representation of patients' clinical histories encoded using International Classification of Diseases (ICD) codes. Doctor AI predicts patients' next clinical events, i.e. the set of diseases (represented as ICD codes) that each patient will have in future visits to the hospital.

Therefore, we use it in our experiments as clinical DSS and study how to improve Doctor XAI explanations of Doctor AI predictions to enable clinical reasoning.

Ontologies Use in XAI. Some XAI works already explored how to use ontologies (or knowledge graphs) to improve the explanation process or to tailor explanations to specific user needs or characteristics. Besides Doctor XAI, also the authors of *Trepan Reloaded* [13] use the ontology in the explanation extraction process. In particular, they use ontology to constrain the training of the decision tree acting as a local interpretable model. Closer to our research, other works use ontologies to tailor the explanation to user-specific needs [7,21]. The authors of [8] use an ontology that encodes all types of explanations to find the most appropriate one for user questions.

In [28] the authors use an ontology to customize the explanation to user needs. However, to the best of our knowledge, ours is the first attempt to enrich explanations of clinical DSS to enable clinical reasoning and the first method that extracts sentences from clinical notes guided by ontology and ICD-9 codes (the ninth revision of ICD). We are aware of the existence of semantic annotation tools like [1], and [19]. However, our method is different, and it does not tag each sentence in the clinical note with a corresponding entity. Our method highlights only the relevant sentences of the note based on the associated ICD-9 codes and the relations extracted from the ontology. This difference did not allow us to compare our method with the already existing tools.

3 Problem Statement

Our aim is to use medical ontologies and external sources of medical knowledge to semantically enrich the explanation provided by state-of-the-art explainability techniques for clinical DSS. In particular, we are interested in augmenting explanations, that consider only the features given as input to the model, with external sources of knowledge in order to present to the end-user the complete clinical picture relevant for a particular algorithmic decision and enabling clinical reasoning. We focus our effort on the post-hoc explanations provided by Doctor XAI [24] and use clinical notes representing the patient's discharge summary as an external source of knowledge. We have already presented Doctor XAI in Sect. 2 and now we provide more details on its explanations. In Fig. 1, we

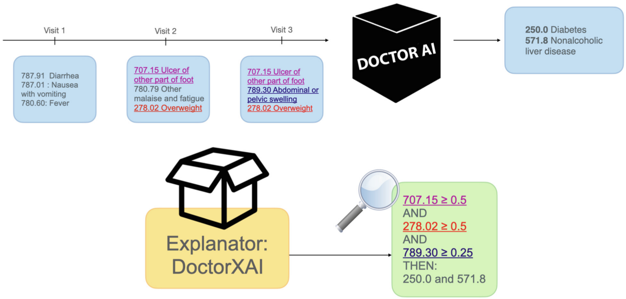

Visit 1
Visit 2
Visit 3
787.91 Diarrhea
787.01 : Nausea with vomiting
780.60: Fever
707.15 Ulcer of other part of foot
780.79 Other malaise and fatigue
278.02 Overweight
707.15 Ulcer of other part of foot
789.30 Abdominal or pelvic swelling
278.02 Overweight
DOCTOR AI
250.0 Diabetes
571.8 Nonalcoholic liver disease
Explanator: DoctorXAI
707.15 ≥ 0.5
AND
278.02 ≥ 0.5
AND
789.30 ≥ 0.25
THEN:
250.0 and 571.8

Fig. 1. An example of Doctor XAI Explanation.

show an example of an explanation of a Doctor AI outcome for a patient having three visits. Each visit is represented by a set of ICD-9 codes and the explanation for the multi-label classification provided by Doctor AI is the decision rule depicted in the bottom right. Each conjunction of the rule premise follows the following pattern: ICD_code $\gtrless$ *threshold_value*. Here, the *threshold_value* is a split value assigned by a decision tree to that ICD code. The internal encoding of Doctor XAI allows giving a temporal interpretation of such value, e.g., $threshold\_value = 0.5$ means that the ICD code was present in the last visit. At the top of the image, we have a more readable representation of the explanation. The ICD-9 codes of the patient's clinical history identified as meaningful by Doctor XAI have been coloured to enhance the readability. However, the final user who wants to exploit this explanation has to analyse the description associated with each highlighted code and derive the possible relationships and their meaning. Furthermore, the explanation does not provide any information on the clinical context of the patient. The method we are proposing aims to enrich the Doctor XAI explanation with information derivable from clinical notes associated with each visit and written by nurses and physicians.

Our methodology enriches such an explanation by highlighting the parts of the patient's clinical notes mostly correlated with the ICD-9 codes and uses medical ontologies to identify if, in that clinical note, there are references to clinically relevant information such as the ICD-9 description, the parts of the body affected by the disease, its causes and its effects.

4 Methodology

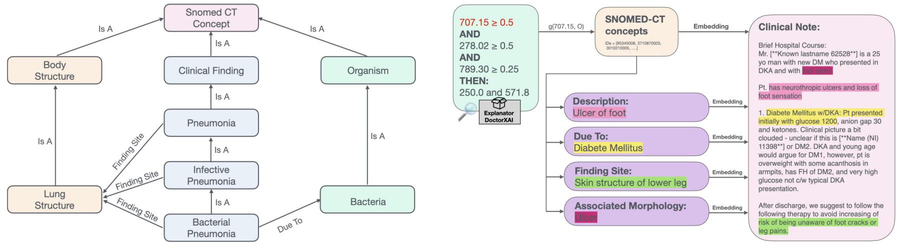
Our methodology exploits the SNOMED-CT medical ontology [29] to semantically enrich XAI explanations. The SNOMED-CT ontology contains a comprehensive representation of clinical healthcare terminology including diseases, symptoms, signs, diagnoses, medications and procedures. Our methodology first finds all the SNOMED-CT concepts related to each ICD-9 code in the explanation, then it selects some clinically relevant ontological relationships associated with these concepts (more details in Sect. 4.1), and finally uses clinical embeddings to find the parts of the patient's clinical note most related to these relationships and highlights them on the clinical note itself (more details in Sect. 4.2). A bird-view of our methodology is provided in Fig. 2b. In Fig. 2a we show an example of some of the concepts and relations contained in the SNOMED-CT Ontology. In particular, for the concept "Bacterial Pneumonia" we have two different relations, the *Finding Site*, which is "Lung Structure", and the *Due to* which is "Bacteria". Note that all the diseases are also involved in a Parent-Child relation where the parent node represents a more general disease than the child e.g. "Pneumonia" is more general than "Infective Pneumonia".

4.1 SNOMED-CT Relationships Extraction

Each ICD-9 of the explanation has a one-to-many mapping to the concepts in the SNOMED-CT ontology. For example, consider the ICD-9 code 707.15, which stands for "Ulcer of other part of foot". This code is mapped to a set of SNOMED-CT concepts such as "Ulcer of foot", "Ulcer of big toe" and "Diabetic foot". For providing the clinician with the most accurate clinical context related to the decision, we first consider all of these possibilities and for each of them, we extract all the relevant clinical information. We focus on three SNOMED-CT ontological relationships: (a) *Finding Site*, i.e., the body site affected by a condition; (b) *Associated Morphology*, i.e., the morphological changes seen at the tissue or cellular level that are characteristics of a disease; and (c) *Due to*, i.e., the cause of the clinical finding, might be another clinical finding or a procedure.

More formally, we define a function g that given an ICD-9 code cd and the SNOMED-CT ontology O returns the corresponding set of the SNOMED-CT concepts SC, i.e., $g(cd, O) = SC$. Then, starting from the set of concepts SC, our method navigates the SNOMED-CT ontology and derives:

- A set of descriptions $D = \cup_{s \in SC} d_s$, where each d_s is the description associated with the SNOMED-CT concept s;
- A set of finding sites $F = \cup_{s \in SC} F_s$, where $F_s = f_s^1, f_s^2, \ldots, f_s^n$ is the set of finding sites associated with the SNOMED-CT concept s;


Snomed CT Concept
Is A
Is A
Is A
Body Structure
Clinical Finding
Organism
Is A
Is A
Is A
Pneumonia
Finding Site
Is A
Infective Pneumonia
Finding Site
Lung Structure
Finding Site
Is A
Bacteria
Bacterial Pneumonia
Due To
707.15 ≥ 0.5
AND
278.02 ≥ 0.5
AND
789.30 ≥ 0.25
THEN:
250.0 and 571.8
Explanator
DoctorXAI
g(707.15, O)
SNOMED-CT concepts
IDs = [95345008, 3710870003, 3010210005, ...]
Embedding
Description:
Ulcer of foot
Due To:
Diabete Mellitus
Finding Site:
Skin structure of lower leg
Associated Morphology:
Ulcer
Embedding
Embedding
Embedding
Embedding
Clinical Note:
Brief Hospital Course:
Mr. [**Known lastname 62528**] is a 25 yo man with new DM who presented in DKA and with foot ulcer.
Pt. has neurothropic ulcers and loss of foot sensation
1. Diabete Mellitus w/DKA: Pt presented initially with glucose 1200, anion gap 30 and ketones. Clinical picture a bit clouded - unclear if this is [**Name (NI) 11398**] or DM2. DKA and young age would argue for DM1, however, pt is overweight with some acanthosis in armpits, has FH of DM2, and very high glucose not c/w typical DKA presentation.
After discharge, we suggest to follow the following therapy to avoid increasing of risk of being unaware of foot cracks or leg pains.

(a) (b)

Fig. 2. (a) SNOMED-CT Ontology relationships and (b) Bird-view of our methodology.

- A set of associated morphology caused by the disease $M = \cup_{s \in SC} M_s$, where $M_s = m_s^1, m_s^2, \ldots, m_s^k$ is the set of associated morphology associated with the SNOMED-CT concept s;
- A set of causes of the disease $C = \cup_{s \in SC} C_s$, where $C_s = c_s^1, c_s^2, \ldots, c_s^h$ is the set of causes associated to the SNOMED-CT concept s.

We denote by $f \in F$, $m \in M$ and $c \in C$ any of the finding sites, associated morphology and causes extracted from the ontology.

4.2 Information Extraction from Clinical Notes

We exploit biomedical word embeddings to encode the description of each clinically relevant piece of information found in the previous step and find the most similar piece of text in the clinical note associated with the patient. Given an ICD-9 code and a clinical note N, our methodology, by using the function g (defined above), first extracts from the ontology O the set of descriptions D related to the concepts SC in SNOMED-CT, or the corresponding sets of finding sites F, associated morphology M, causes C. Then, for each $d_s \in D$, $f \in F$, $m \in M$ or $c \in C$ To this end, we use a sliding window of length r that generates a set of word sequences W composed of r contiguous words that can be used to represent the note N. We then embed each element of W obtaining the corresponding set of pairs $\langle embedding, sentence \rangle$ denoted by E. We also compute the embedding for each d_s, $f \in F$, $m \in M$, or $c \in C$ and for each of them we identify the most similar embedded sentence E_w corresponding to the pair $\langle E\_w, w \rangle \in E$.

We use the cosine similarity metric to compute a similarity score between these embeddings and those generated using the sliding window. Given two embeddings A and B, the similarity is computed as follows: $Similarity = \frac{A \cdot B}{||A|| ||B||}$. Thus, we obtain that each element of D, F, M and C is associated with the most similar sentence of the note and a similarity score i.e., we have four score vectors D_{score}, F_{score}, M_{score} and C_{score}.

To identify the descriptions in D, the finding sites in F, the associated morphology in M and the causes in C referred to in the note N, we select from these sets only the elements with a similarity score higher than a threshold τ. In our experiments, the threshold τ is computed as the 90th percentile of the score vectors. We compared several thresholds. In the end, we chose the one that allows us to have the highest number of correctly highlighted sentences. By highlighting all the sentences of the discharge summary having $T\_scores \geq \tau$, we present to the end-user only the information relevant to the patient under study.

The length r of the sliding window has a clear impact on the embedding-based representation of each note and on the resulting parts of the text that are associated with specific concept descriptions, finding sites, etc. We propose to select for each type of relationship the more appropriate r value by using a data-driven approach. In particular, it finds the suitable r value for a given relationship type by testing several sliding window length values on a separate set of clinical notes and by selecting the value leading on average to the highest similarity score.

5 Experiments and Results

In this section, we experimentally show the ability of our approach to identify the correct sentences of the patient's clinical notes for explanation enrichment.[1]

We carried out two types of experiments with the help of a human expert. In the first experiment, the human expert manually annotated a set of clinical notes with the ICD-9 descriptions and their relevant ontological relationships. This allowed us to build a ground truth for the automatic extraction of our methodology. In the second experiment, we used our methodology to extract the sentences from another set of clinical notes and then, we asked the human expert to validate whether the identified sentences were correct.

5.1 Dataset

We tested our methodology on the Medical Information Mart for Intensive Care database (Mimic-III) [18]. This dataset contains de-identified data of approximately 40.000 patients collected between 2001 and 2012 in the Beth Israel Deaconess Medical Center data in Boston. Data is stored in 26 different tables; in particular, we used the NoteEvents table which contains all the clinical notes written by nursing and clinicians during a patient's stay in the hospital.

Note Cleaning. We applied a pre-processing to the clinical notes to clean them and reduce noise: we have lower-cased the text; we have removed numbers; we have substituted odd characters with space; we have removed stopwords; we have removed the punctuation; and we have replaced the contractions in the text with an extended form using a dictionary of possible contractions.

5.2 Implementation Details

We trained Doctor AI for 50 epochs, splitting MIMIC-III using 70% of its patients as a training set, 15% as a validation set, and 15% as a test set. We then used Doctor XAI as detailed in the original paper. To navigate the ontology, we used a Python Library called PyMedTermino [20]. For the embedding of the clinical notes' sentences, we used three different methods:

- *BioWordVec* [31], a pre-trained word embedding for biomedical natural language processing trained on PubMed and Mimic-III;
- *ClinicalBert* [17], a Bert based embedding trained on Mimic-III;
- and *BioSentVec* [11], a biomedical sentence embedding with sent2vec trained on Mimic-III and PubMed.

[1] Code available at: https://github.com/lucacorbucci/Semantic-Enrichment. Hardware used: NVIDIA Quadro RTX 6000 GPU, Intel(R) Core(TM) i9-10980XE CPU @ 3.00 GHz, 256 gb of RAM.

BioSentVec and BioWordVec are based on Word2Vec, the embeddings are context-independent and we can use them without the model that generated them because we just have <$key, value$> pairs where the keys are the words and the values are the embeddings. ClinicalBert is based on Bert, the embeddings are generated considering the context of a word, this means that we have to give a sentence as input to the model and it will return the embedding. This is computationally more expensive than the Word2Vec model.

Before applying the embedding we identified the suitable length value r of the sliding window for each type of relationship: $r = 7$ for the *Finding Site* and *Associates Morphology* relationships, $r = 9$ for *Due to* relationship and $r = 10$ for the *Description.* To this end, we tested values in the range between 3 and 30 using a subset of 500 clinical notes contained in the original dataset.

5.3 Human Validated Experiment

Clinical Notes Manual Annotation. The domain expert took into account each ICD-9 code associated with the clinical notes and highlighted by Doctor XAI. The notes were manually annotated highlighting the most similar sentences to the following information: *i)* Code description; *ii)* Cause of the disease associated with the code; *iii)* Finding site of the disease associated with the code; and *iv)* Associated morphology of the disease associated with the code. In particular, we considered the clinical histories of nine different patients, involving a total of 32 clinical notes. These patients have been diagnosed with ICD-9 code 250.00 i.e. diabetes, 584.9 i.e. Acute kidney failure, 428.0 i.e. Congestive heart failure and 401.9 i.e. Unspecified essential hypertension, among other diseases. Once the domain expert annotated the notes, we tested our method to compare the extracted sentences with the manually annotated ones. In Table 1, we report *Accuracy*, *F1-Score*, *Precision* and *Recall* for each ontological relationship and the corresponding confidence interval at $1 - \alpha = 0.95$ confidence level. The results are divided according to the type of relationship and in bold we highlight the best performance. The confidence intervals for all the metric values are tight meaning that the performances of the methods are reliable. To evaluate these metrics, we have defined:

- *True Positive* as the number of sentences that were manually annotated in the clinical notes and are correctly annotated by our method;
- *False Negative* as the number of sentences in the clinical notes that our method does not annotate because they have a similarity score lower than the input threshold and that were manually annotated by the domain expert.
- *False Positive* as the number of sentences in the clinical notes that our method annotates and that do not have a corresponding manual annotation.
- *True Negative* as the number of sentences in the clinical notes that our method does not annotate because they have a similarity score lower than the input threshold and that do not have a corresponding manual annotation.

Table 1 shows that BioWordVec presents the best performance across all relationships, for this reason, we chose to employ it in our methodology. Furthermore,

Table 1. Validation on 32 manually annotated clinical notes of 9 patients. Confidence of Accuracy, Precision, Recall and F1-score at $1 - \alpha = 0.95$ of confidence level.

Relationship	Embedding		Accuracy	F1-score	Precision	Recall
Description	BioWordVec	Value Confidence	**0.718** 0.715–0.719	**0.707** 0.704–0.708	**0.819** 0.815–0.819	**0.622** 0.619–0.624
Description	BioSentVec	Value Confidence	0.662 0.659–0.663	0.664 0.661–0.665	0.804 0.800–0.804	0.566 0.563–0.567
Description	ClinicalBert	Value Confidence	0.640 0.637–0.641	0.602 0.599–0.603	0.654 0.651–0.655	0.557 0.555–0.559
Finding site	BioWordVec	Value Confidence	**0.743** 0.740–0.744	0.274 0.273–0.277	0.170 0.169–0.173	**0.708** 0.705–0.709
Finding site	BioSentVec	Value Confidence	0.726 0.723–0.727	**0.294** 0.293–0.297	**0.200** 0.199–0.203	0.555 0.553–0.557
Finding site	ClinicalBert	Value Confidence	0.686 0.683–0.687	0.214 0.213–0.217	0.150 0.149–0.153	0.375 0.373–0.377
Due to	BioWordVec	Value Confidence	**0.666** 0.647–0.673	**0.451** 0.440–0.466	**0.350** 0.342–0.368	**0.636** 0.618–0.644
Due to	BioSentVec	Value Confidence	0.600 0.582–0.609	0.091 0.091–0.119	0.050 0.050–0.080	0.500 0.486–0.513
Due to	ClinicalBert	Value Confidence	0.568 0.552–0.579	0.214 0.211–0.238	0.150 0.149–0.176	0.375 0.366–0.392
Associated morphology	BioWordVec	Value Confidence	**0.856** 0.845–0.856	**0.577** 0.571–0.581	**0.464** 0.459–0.470	**0.764** 0.755–0.766
Associated morphology	BioSentVec	Value Confidence	0.803 0.793–0.803	0.409 0.405–0.415	0.321 0.318–0.329	0.562 0.556–0.566
Associated morphology	ClinicalBert	Value Confidence	0.734 0.726–0.736	0.339 0.336–0.347	0.321 0.318–0.329	0.360 0.356–0.367

BioWordVec computational runtime was an order of minutes shorter if compared with ClinicalBert, as previously observed in [19]. Our experiment pointed out that the *Description* is the easiest relation to search for and in most of the cases, our methodology is able to extract the same sentence highlighted during the manual annotation phase. On the contrary, it is not easy to deal with *Finding Site* and *Due to*. As explained by our domain expert usually this information is often underlined by the clinicians and is not explicitly written in the notes. This sometimes led to the extraction of wrong sentences.

A Kruskal-Wallis test was used to determine whether or not there are statistically significant differences between the medians of accuracy, precision, recall and F1-score of the different embedding methods reported in Table 1 (BioWordVec, BioSentVec, and ClinicalBert). The Kruskal-Wallis test is the non-parametric test considered equivalent to the One-Way ANOVA and, given the low number of observations that we are comparing, it is the best fitting for our setting. The Kruskal-Wallis test uses the following null and alternative hypotheses: H_0: "The median is equal across all embedding methods", H_1: "The median is not equal across all embedding methods". For the *accuracy* we obtained that the Kruskal-Wallis Statistics is 2.192 with a p-value of 0.334 (> 0.05) meaning no statistically significant difference among the accuracy medians, so the H_0 hypothesis cannot be rejected. For the *recall* we obtained that the Kruskal-Wallis statistic is 8.800

Table 2. Kruskal-Wallis test for the recall: results.

	BioWordVec	BioSentVec	ClinicalBert
BioWordVec	stat: 1.000 p: 0.000	stat: 3.036 p: 0.0814	stat: 5.398 p: **0.0202**
BioSentVec		stat: 1.000 p: 0.000	stat: 5.333 p: **0.0209**
ClinicalBert			stat: 1.000 p: 0.000

Table 3. Human validation of the extracted sentences using a 90th percentile threshold.

Relationship	Description		Finding site		Associated Morphology		Due to	
Embedding	*valid*	*non-valid*	*valid*	*non-valid*	*valid*	*non-valid*	*valid*	*non-valid*
BioWordVec	75 (75%)	25 (25%)	**65 (65%)**	35 (35%)	**21 (75%)**	7 (25%)	**13 (65%)**	7 (35%)
BioSentVec	**77 (77%)**	23 (23%)	56 (56%)	44 (44%)	**21 (75%)**	7 (25%)	**13 (65%)**	7 (35%)
ClinicalBert	67 (67%)	33 (33%)	32 (32%)	68 (68%)	18 (64%)	10 (36%)	9 (45%)	11 (55%)

with a p-value of 0.012 meaning a statistically significant difference among the recall medians, so the H_0 hypothesis has been rejected. We performed both a Kruskal-Wallis test and a Mann Whitney test on the pairs to verify which are the pairs with a significant difference. In Table 2 we report the results of the pairwise comparisons of the Kruskal-Wallis test (equal results were obtained with the Mann Whitney test).

Looking at Table 2, it is interesting to note how the pairwise comparisons between BioWordVec *vs* ClinicalBert and BioSentVec *vs* ClinicalBert give statistically significant differences between the pairs (always lower for ClinicalBert). For the *F1-score* (Kruskal-Wallis Statistics 1.505 with a p-value of 0.471) and the *precision* (Kruskal-Wallis Statistics 1.462 with a p-value of 0.481) we found no significant difference among at least one of the medians (both for F1-score and precision), so the H_0 hypotheses have to be accepted in both cases. Thus, to summarize the three embedding methods present statistically significant differences only for the recall: BioWordVec and BioSentVec perform statistically better than ClinicalBert, reinforcing the choice of one of these two embeddings.

Classification of Extracted Sentences. We made a second experiment exploiting the knowledge of the domain expert. We selected almost 100 notes classified with the ICD-9 250.00 (diabetes), 584.9 (Acute kidney failure), 428.0 (Congestive heart failure) and 401.9 (Unspecified essential hypertension).

Then, we ran our method on each note to highlight the most similar sentences to the *Description*, *Finding Site*, *Associated Morphology* and *Due To* relation associated with all the ICD-9 with which the note is associated. We used the previously mentioned method and the three different embeddings to compute the similarity. After extracting the sentences with our method, the domain expert analysed each sentence evaluating the correlation with the relation with which similarity was calculated and if the highlighted sentence provided helpful information about the patient's clinical history. Each of the extracted sentences was classified as "*Valid Sentence*" or "*Non-Valid Sentence*". In Table 3 we report the result of this experiment with a similarity threshold of the 90th percentile. The results show that the performances using embeddings BioWordVec and BioSentVec are very similar while those using ClinicalBert are slightly worse.

6 Conclusion and Future Work

We presented a methodology to semantically enrich the explanation of an XAI technique in the healthcare context by exploiting SNOMED-CT ontology and clinical notes. In particular, it highlights the relevant clinical information related to one algorithmic decision directly on the patient's clinical note. Thanks to the domain expert, we were able to annotate a small part of the dataset and to have a preliminary "human validation" of our methodology. The presence of a "human validator" was crucial in our methodology. Unfortunately, we have not found any pre-annotated dataset that could fit our needs and that could be used as a ground truth. The "human-validated" experiment showed promising results concerning the identification of sentences related to the description of the disease and the associated morphology while selecting the correct finding site and cause of the disease is more challenging. We studied many different approaches to extract the information, and we compared different embeddings to have a better representation of our notes. In terms of embeddings, we compared the performances achieved with BioWordVec, BioSentVec and ClinicalBert, and we concluded that, for the same performance, BioWordVec performs slightly better in general and it is faster in computing embeddings. A limitation of an approach that involves the use of pre-trained embeddings is that we would not be able to generalise this task with the same performances when using a completely different medical dataset. In that context, an embedding like ClinicalBert would probably perform better. However, it would have a high computational cost to the embedding computation.

In the future, we would like to validate our method on a larger quantity of clinical notes and exploit our methodology to generate explanations expressed by natural language.

In addition, we would like to test the methodology to understand if the semantically enriched explanation could improve the interpretability of the explanation. Lastly, we plan to investigate the opportunity to exploit our methodology to generate explanations expressed by natural language.

Acknowledgments. This work is supported by: the EU NextGenerationEU programme under the funding schemes PNRR-PE-AI FAIR (Future Artificial Intelligence Research); the EU - Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 - Integrating Activities for Advanced Communities" (G.A. n.871042) "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu); PNRR-"SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013; TAILOR (G.A. 952215), HumanE-AI-Net (G.A. 952026) and EU H2020 project XAI (Grant Id 834756).

References

1. Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Proceedings AMIA Symposium, pp. 17–21 (2001). ISSN 1531-605X. eprint 11825149. https://pubmed.ncbi.nlm.nih.gov/11825149
2. Barnett, A.J., et al.: IAIA-BL: a case-based interpretable deep learning model for classification of mass lesions in digital mammography. arXiv preprint arXiv:2103.12308 (2021)
3. Bhatt, U., et al.: Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 648–657 (2020)
4. Bodria, F., et al.: Benchmarking and survey of explanation methods for black box models. CoRR abs/2102.13076 (2021)
5. Boominathan, S., et al.: Treatment policy learning in multiobjective settings with fully observed outcomes. In: ACM SIGKDD 2020, pp. 1937–1947 (2020)
6. Caruana, R., et al.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721–1730 (2015)
7. Celino, I.: Who is this explanation for? Human intelligence and knowledge graphs for eXplainable AI. arXiv preprint arXiv:2005.13275 (2020)
8. Chari, S., Seneviratne, O., Gruen, D.M., Foreman, M.A., Das, A.K., McGuinness, D.L.: Explanation ontology: a model of explanations for user-centered AI. In: Pan, J.Z., et al. (eds.) ISWC 2020. LNCS, vol. 12507, pp. 228–243. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_15
9. Cheerla, A., et al.: Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics **35**(14), i446–i454 (2019)
10. Chekroud, A., et al.: The perilous path from publication to practice. Mol. Psychiatry **23**(1), 24–25 (2018)
11. Chen, Q., et al.: BioSentVec: creating sentence embeddings for biomedical texts. In: 2019 IEEE IICHI (2019)
12. Choi, E., et al.: Doctor AI: predicting clinical events via recurrent neural networks. In: Machine learning for healthcare conference. PMLR (2016)
13. Confalonieri, R., et al.: Trepan reloaded: a knowledge-driven approach to explaining artificial neural networks (2019)
14. EU General Data Protection Regulation. European Commission (2018). https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf. Accessed 17 June 2019
15. European Parliament. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Antelligence Act) and amending certain union legislative acts (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206. 11 June 2021
16. Guidotti, R., et al.: A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) **51**(5), 1–42 (2018)
17. Huang, K., et al.: ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv:1904.05342 (2019)

18. Johnson, A.E., et al.: MIMIC-III, a freely accessible critical care database. Sci. Data **3**, 160035 (2016)
19. Kraljevic, Z., et al.: Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit (2020)
20. Lamy, J.-B., et al.: PyMedTermino: an open-source generic API for advanced terminology services. Stud. Health Technol. Inform. **210** (2015)
21. Longo, L., Goebel, R., Lecue, F., Kieseberg, P., Holzinger, A.: Explainable artificial intelligence: concepts, applications, research challenges and visions. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2020. LNCS, vol. 12279, pp. 1–16. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57321-8_1
22. Lundberg, S.M., et al.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017)
23. Ma, X., et al.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognit. **110**, 107332 (2021)
24. Panigutti, C., et al.: Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In: ACM FAccT (2020)
25. Panigutti, C., Guidotti, R., Monreale, A., Pedreschi, D.: Explaining multi-label black-box classifiers for health applications. In: Shaban-Nejad, A., Michalowski, M. (eds.) W3PHAI 2019. SCI, vol. 843, pp. 97–110. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-24409-5_9
26. Panigutti, C., et al.: FairLens: auditing black-box clinical decision support systems. Inf. Process. Manage. **58**(5), 102657 (2021)
27. Ribeiro, M.T., et al.: "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
28. Rožanec, J.M., et al.: Semantic XAI for contextualized demand forecasting explanations. arXiv preprint arXiv:2104.00452 (2021)
29. U. T. Services. SNOMED CT International Edition. https://www.nlm.nih.gov/healthit/snomedct/international.html
30. Signoroni, A., et al.: BS-net: learning COVID-19 pneumonia severity on a large chest X-ray dataset. Med. Image Anal. **71**, 102046 (2021)
31. Zhang, Y., et al.: BioWordVec: improving biomedical word embeddings with subword information and MeSH. Sci. Data **6**, 52 (2019)