Enhancing Data Quality and Semantic Annotation by Combining Medical Ontology and Machine Learning Techniques

Zina Nakhla[1,*] and Manel Sliti[2,†]
[1]*Université de Tunis, Institut Supérieur de Gestion, BESTMOD Laboratory, Tunis, Tunisia*
[2]*Université de Manouba, Institut Supérieur des Arts et Multimedia de Manouba, Manouba, Tunisia*

Keywords: Interoperability, Ehr, Ontology, Machine Learning, NLP.

Abstract: Effective management of electronic health records (EHR) is a major challenge in the modern healthcare sector. Despite technological advances, the interoperability of medical data remains a crucial challenge. This complex problem is manifested by the diversity of data formats, the presence of multiple standards and the heterogeneity of Information Technology (IT) systems used in health- care establishments. However, the diversity of IT systems and the complexity of medical terminologies often make data interoperability and semantic annotation in the healthcare domain difficult. To address this challenge, our study proposes an innovative approach to standardize the representation of medical concepts, to automate the detection of medical abbreviations and to improve the contextual understanding of medical terms. We developed an ontological model to harmonize the representation of medical data, thus facilitating their exchange and integration between different health systems. In parallel, we used advanced machine learning techniques for automatic detection of medical abbreviations in medical texts, and applied Natural Language Processing to improve contextual understanding of medical terms. The results of our study demonstrate the effectiveness of our approach in solving challenges related to medical data management. By combining different advanced techniques, our approach helps overcome barriers to medical data interoperability and paves the way for better healthcare system integration and improved patient care.

1 INTRODUCTION

An Electronic Health Record (EHR) is a digital version of a patient's paper chart (Sachdeva and Bhalla, 2022). EHR systems facilitate the collection, storage, and sharing of patient information in a structured manner to enhance healthcare delivery and clinical decision-making. They are designed to provide real-time, centralized, and secure medical information accessible to authorized users. While an EHR contains a patient's medical history, diagnoses, prescriptions, and other clinical data, its role goes beyond mere data archiving. It allows access to evidence-based tools that help healthcare professionals optimize their medical decisions (Fennelly and Moroney, 2024).

One of the key features of an EHR is that health information can be created and managed by authorized providers in a digital format capable of being shared with other providers across more than one health care organization. EHRs are built to share information with other health care providers and organizations such as laboratories, specialists, medical imaging facilities, pharmacies, emergency facilities, and workplace clinics so they contain information from all clinicians involved in a patient's care. However, sharing the information needed is a very complicated problem.

Doctors and specialists are suffering from collecting distributed data spread over different locations, and lack of the interoperability among all the healthcare information systems. Also, the domain of healthcare produces a huge quantity of data from various disparate sources. The single patient's data can be dispersed over diverse EHRs with various representation ways (Begoyan, 2007).

Data in EHRs could be presented in many different types of formats: structured data as

* Corresponding author
† Contributing author

Nakhla, Z. and Sliti, M.
Enhancing Data Quality and Semantic Annotation by Combining Medical Ontology and Machine Learning Techniques.
DOI: 10.5220/0013750200004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 2: KEOD and KMIS*, pages 140-150
ISBN: 978-989-758-769-6; ISSN: 2184-3228

(database) (Schloeffel et al, 2006) unstructured data such as (documents, images,...) (Kiourtis et al, 2017), and semi-structured data (XML files) (Mylka et al, 2012).

The EHR interoperability problem refers to the ability of different systems to seamlessly exchange, interpret, and utilize patient data across various healthcare providers and settings. This challenge arises from differences in data formats, adopted standards, and proprietary systems, leading to inefficiencies, medical errors, and fragmented care. Interoperability can be categorized into three levels, firstly Technical interoperability, which ensures the physical connection between systems and data transfer. Secondly Syntactic interoperability, which enables data exchange through standardized formats (HL7, XML), (Sartipi and Dehmoobad ,2008). Thirdly Semantic interoperability, which ensures a uniform understanding of exchanged data by using standardized medical terminologies. Semantic interoperability is essential for improving clinical decision-making, enhancing care coordination, and ensuring that all healthcare professionals have access to consistent and reliable medical information. To address these challenges, several standards and terminologies have been developed. Among the standards enabling structured clinical content exchange are Health Level Seven (HL7) Digital Imaging and Communications in Medicine Structured Reporting (DICOM SR), (Begoyan, 2007) , ISO EN 13606 (Costa et al, 2011) , openEHR [(Kalra, 2006), (Schloeffel et al, 2006),(Da Costa, 2019),(Roehrs et al, 2018), (Begoyan, 2007)], GEHR (Celesti et al, 2016). Nonetheless,semantic interoperability cannot be achieved without the adoption of standardized medical terminologies, such as the Systematized Nomenclature of Medicine Terms (SNOMED CT), which is the most comprehensive medical terminology system used worldwide. SNOMED CT enables precise encoding of clinical information, facilitating medical document annotation, clinical decision support, and EHR interoperability.

Despite continuous efforts, limitations persist, particularly concerning the variety of medical data formats, variability of collection protocols, confidentiality concerns, and the absence of uniform standards for semantic tagging. These challenges hinder the automatic understanding of medical data, slowing down interoperability advancements. To overcome these limitations, this study proposes an innovative approach that integrates ontologies, machine learning, and Natural Language Processing (NLP). This combination allows for standardizing medical concept representation, automatically detecting medical abbreviations, and improving the contextual understanding of medical terms.

Unlike previous approaches, our method introduces a novel integration of structured ontology-based representations with advanced machine learning and NLP models, enhancing data standardization, medical entity recognition, and abbreviation expansion. This paper details each phase of our proposed approach, demonstrating how it helps overcome barriers to medical data interoperability and facilitates seamless integration within healthcare systems, ultimately improving patient care.

This paper presents an integrated approach to solving challenges related to medical data interoperability by combining ontology, machine learning and NLP. We discuss in detail the different phases of our approach by combining different advanced techniques, our approach helps overcome barriers to medical data interoperability and paves the way for better healthcare system integration and improved patient care. The rest of this paper is structured as follows. Section 2 presents the related work and previous studies. Section 3 describes the used dataset. The proposed architecture with its internal phases are described in Section 4. Section 5 contains the results of experiments, and we discuss our evaluation of the proposed solution. Section 6 presents a comparison study. Finally, the conclusion and future work.

2 RELATED WORK

One of the most consistent themes across the literature is the potential of EHR to revolutionize multiple aspects of healthcare. Many of these papers describe how an EHR system can improve how patients are diagnosed, treated and improve healthcare (Gunter et al , 2005). McClanahan describes how quick access to patient information through a universal EHR system can save the lives of thousands of emergency room patients each year by reducing medical errors (McClanahan, 2008).

Santos et al. explained the importance of EHR due to its ability to integrate various user interfaces and programs. While their findings are promising, they do not fully address the significant infrastructural and financial challenges of integrating diverse systems at a national or international level. Many of the proposed solutions are theoretical and lack real-world validation, which raises questions about their scalability and long-term feasibility (Santos et al, 2010).

Berges et al. explore the use of ontologies in improving interoperability in heterogeneous EHR systems. First, they used a reasonable ontology that EHR-related concepts focus on meanings and perspectives. Second, they tracked modules that enable the acquisition of rich ontological descriptions of EHR data guided by constraint models of medical knowledge frameworks. Third, it considers necessary mapping maxims between the concepts that mappings update. While their model demonstrates the potential for more structured and accurate data representation, it overlooks the practical challenges of implementing such models across diverse healthcare environments. The author's approach to aligning heterogeneous EHR descriptions is promising but requires further exploration into how these mappings can be dynamically updated in real-time across various clinical settings (Berges et al, 2011).

Ferrer et al. represented and integrated patient data from many different heterogeneous data sources and encouraged the integration of patient data into a rule-based Clinical Decision Support System (CDSS). Their work represents a significant step in bridging standards such as ISO/CEN 13606 and HL7 through OpenEHR approach. They proposed a Personal Health Records (PHR) combining OpenEHR and HL7 supported by a service-oriented information exchange system (González-Ferrer et al, 2012).

Liyanage et al proposed a method for improving semantic interoperability in healthcare by using ontologies. They present an ontology toolkit that facilitates the development of ontologies for chronic disease management using heterogeneous health data. Nevertheless this approach is useful but limited by its focus on specific diseases and may not be easily generalized to other domains (Liyanage et al, 2015).

Also, Kiourtis et al. proposed a generic ontology-based semantic architecture to solve EHR interopability problem. This architecture used an ontology language to transform heterogeneous medical data into a generic schema, called CHL that can be used to represent health data in a way that is consistent and understandable to machines. The CHL also makes it possible to merge different medical ontologies with similar relationships. The authors also pointed out that this method is not lossless, as data transformation can lead to loss of information. This is a significant limitation, as the loss of data could compromise patient care and treatment decisions (Kiourtis et al, 2017).

Futhermore, Hajjamy et al. developed a semiautomatic approach to integrate classical data sources into an ontological database. Their method involves transforming data sources into ontologies using measures of syntactic, semantic, and structural similarity, to generate a global ontology. The authors suggested that their approach could be enhanced by incorporating other information retrieval techniques and big data methods to handle larger ontologies. Although their method is promising, its reliance on traditional data fusion methods may hinder its ability to handle the large and complex datasets common in modern healthcare (El Hajjamy et al,2018).

In 2019, Li Chen et al. developed an ontology to represent knowledge and relationships in the field of diabetes and used this ontology to build a reasoning model for medical decision making. Their method involved collecting data from various sources, defining Semantic Web Rule Language (SWRL) rules to define the reasoning rules of the ontology, and implementing the SWRL rules in an inference engine to create a help system for decision called Ontology-based Medical Diagnostic and Treatment Platform (OMDP). OMDP used OWL ontology and SWRL rules to analyse patient's symptoms, make diagnosis and recommend treatment by integrating different types of diabetes knowledge. Despite this, the limited scope of their model focused only on diabetes raises concerns about its broader applicability across other medical conditions. Furthermore, while the integration of different knowledge sources is valuable, the lack of real-time data integration limits its practical utility (Li Chen et al, 2019).

In 2020, Sreenivasan and chacko. proposed an approach to map heterogeneous EHR data to ontologies, to create a semantically annotated knowledge base that can be queried. The approach involved semantically annotating the data using an ontology, building a new of knowledge and the use of a semantic inference module to infer knowledge from data. The proof of concept showed that health data mapped into a relational database using ontologies can be used for inference and for rapid decision making by healthcare professionals. The proof of concept is promising but remains at a pilot stage, and further validation is required before it can be widely adopted in clinical practice (Sreenivasan and Chacko, 2020).

In 2022, Adel et al. proposed an ontological model to integrate patient health data from heterogeneous data sources at a centralized point to improve the quality of care. The authors unified five different healthcare data formats into an unified ontology. The results showed that the proposed model made it possible to integrate and collect all patient data from heterogeneous data sources, improving the quality of

care and reducing medical errors. However, there are limits to the model, including its specificity to a particular area, the use of limited test data, the need for close collaboration between health professionals and computer scientists, and the need for more in-depth validation on a larger population in real clinical environments (Adel et al, 2022).

Overall, most of the existing approaches faced limitations, especially the variety of medical data formats, the variability of collection protocols, the problem of confidentiality of medical data, and the lack of uniform standards for semantic tagging, which have hindered the automatic understanding of data. These challenges have hindered the implementation of comprehensive solutions and highlighted the need for innovative approaches to overcome these limitations.

3 DATASET DESCRIPTION

This section details the data sources and the construction of the patient file, outlining the process used to gather and structure the datasets necessary for our study. The Web offers access to an enormous quantity of documents, in several forms (texts, images, videos, sounds) and in different languages. Most of these documents are freely available, easy to access, and in electronic format. To identify relevant resources, we used the Google search engine with targeted queries such as “medical dataset Excel”, “EHR data sample xls”, and “clinical records spreadsheet». From the search results, we first considered the most popular web pages (the top-ranked pages in the results list). These pages were then filtered according to qualitative criteria:

- Representativeness of the medical domain,
- Audience targeted by the site (general public or healthcare professional),
- Author of the page (health professional or not),
- Language used (easily understandable or specialized terminology).

Based on these criteria, we selected web pages that hosted structured datasets and, more specifically, Excel files. While Excel is not a standard format for EHR representation, its tabular structure makes it an accessible, modifiable, and easily shareable medium for organizing medical information. These files provided a practical way to compile representative terms used by healthcare mediators while maintaining data consistency.

Thus, the dataset consists of Excel files retrieved from public health websites and open research portals, each containing structured medical records in English and French.

The Excel model is organized into five main classes:

1. Patient ID: an anonymized identifier assigned to each patient.
2. Healthcare Professional: the physician, nurse, or medical specialist associated with the patient record.
3. Disease: the main diagnosis for the patient.
4. Medicine: the prescribed drug or treatment.
5. Symptom: the clinical signs reported by the patient or observed by the healthcare professional.

In addition to these core classes, the dataset also contains supplementary attributes such as medical exams, therapeutic procedures, and allergies, which are linked to the five main classes. This structure ensures both flexibility and consistency in representing patient-related information

We started with data cleansing which is a crucial step in the process of managing healthcare data, including EHR. This phase aims to guarantee the quality, consistency and reliability of medical information. During data cleansing, several aspects are taken into account including the detection and removal of duplicates, errors, and missing data. In general, data cleansing plays a major role in creating high-quality EHRs, ensuring that medical information is accurate, complete and consistent, which contributes to better quality healthcare and efficient management of health data.

4 THE PROPOSED ARCHITECTURE

Our proposed approach combines the structuring power of a medical ontology with the flexibility of machine learning and NLP. The motivation for this approach is based on identifying the greatest challenges to medical knowledge. This information is the cornerstone of clinical decision-making, biomedical research, and health management. However, medical data is often unstructured, heterogeneous, and subject to semantic ambiguity, creating significant obstacles for healthcare professionals and researchers who need to extract relevant insights. To address these challenges, our method integrates ontology-based structuring with advanced machine learning techniques and Natural Language Processing (NLP) to enhance semantic annotation and interoperability.

The architecture of this approach is shown in Figure 1. It consists of two main layers: the construction of local ontologies and the unification into a global ontology.

1. In the first layer: Local Ontology Construction
   - A local ontology is generated for each data source, and then converted into an OWL ontology representation.
   - The input EHR data sources used in this process are Excel files.
   - The output of this layer is a structured local ontology in OWL format.
2. In the second layer: Global Ontology Construction
   - The objective of this layer is to generate a unified global ontology that consolidates heterogeneous data sources into a single structured representation.
   - This step provides semantic alignment and ensures consistency across different healthcare information systems.

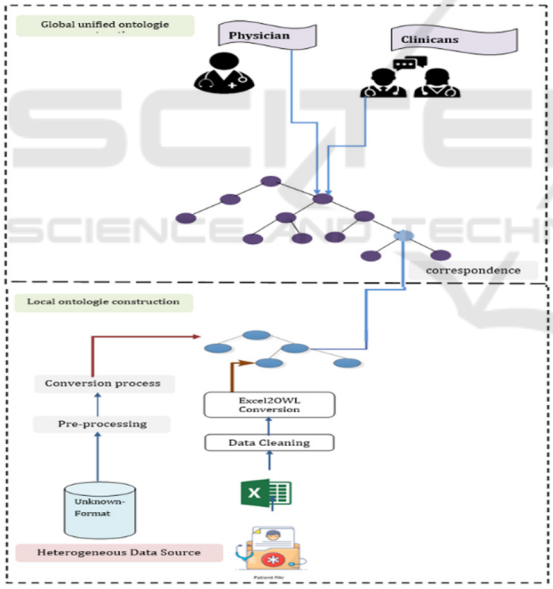

Figure 1: Architecture of the proposed approach.

4.1 Local Ontologies Construction

Converting an Excel file into an ontology is a fundamental process for transforming tabular data into a rich and interpretable semantic structure. This process is essential when we want to create a formal and structured representation of the concepts and relationships present in the data. Excel files are commonly used to store and manage data in tabular form, making them easy to use but less suitable for in-depth semantic analysis. In contrast, an ontology is a formal representation of concepts and relationships providing rich semantics for data. The conversion process begins with defining an appropriate ontology structure for the data domain. This involves identifying the classes, properties and relationships that characterize the data. For example, in an Excel file containing patient information, an ontology could define classes such as ‘‘Patient‘‘,‘‘Disease‘‘, ‘‘Treatment‘‘, etc., and specify properties. Then, each row in the Excel file is transformed into a class instance in the ontology, and the values of the table cells are associated with the corresponding properties.

4.2 Global Ontologies Construction

Ontology integration is a key step for ensuring semantic interoperability across heterogeneous medical datasets. One of the main challenges is that ontologies may be developed under different conceptual frameworks, which complicates their integration. In our approach, we focused on domain-specific ontologies that provide complementary coverage of structural, administrative, and semantic aspects of medical data. We selected three ontologies for integration: EHR EXTRACT RM.owl, which provides a formal representation of electronic health record structures; RIMV3OWL.owl, which models the HL7 Reference Information Model (RIM) and its use in clinical systems; and OGMS.owl (Ontology for General Medical Science), which defines core concepts of medical science such as disease, symptom, and diagnosis. The choice of these ontologies was motivated by their complementary scope: EHR EXTRACT RM and RIMV3OWL address the structural and administrative aspects of patient records, while OGMS provides the semantic layer for core medical concepts.

Although BFO (Basic Formal Ontology) is a well-known top-level ontology, we did not explicitly adopt it in this work. Integrating BFO would have required substantial re-engineering of the selected ontologies, which was beyond the scope of our study. Instead, we aimed to construct a unified mid-level ontology that bridges EHR structures and medical semantics in a pragmatic way, directly addressing interoperability challenges. It is also worth noting that OGMS is itself grounded in BFO, which indirectly provides our integration with a degree of top-level alignment.

Furthermore, during the integration process we identified a lack of sufficient information on medical abbreviations within the existing ontologies. To address this gap, we introduced a dedicated Medical Abbreviations class, in which each abbreviation

instance is systematically organized in alphabetical order and explicitly linked to its full meaning. This addition enhances accessibility and improves the semantic clarity of medical records for healthcare professionals, researchers, and practitioners.

The use of abbreviations is frequent, users of the global ontology could thus quickly access detailed information on medical abbreviations, thereby strengthening the quality and accuracy of medical research and data analysis.

4.3 Machine Learning to Ensure Correspondence

Faced with the complexity of medical terminologies and the diversity of standards, semantic annotation is essential for a precise understanding of medical information. Nevertheless, applying this annotation at scale requires a scalable approach, which is where machine learning comes in. By leveraging machine learning models, our solution aims to automate the semantic annotation process, thus allowing precise identification of medical terms and their contexts within the data. At the same time, machine learning helps overcome interoperability barriers by enabling systems to dynamically adapt to different standards and data formats. The following steps in Figure 2 illustrate a comprehensive approach to solving the automatic abbreviation detection problem.

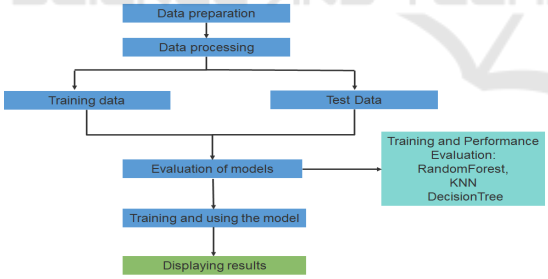

Figure 2: Abbreviation detection model using ML.

4.4 Natural Language Processing to Ensure Correspondence

The use of NLP and the SpaCy library represent an essential dimension of our approach for semantic annotation in the context of EHR. While machine learning offers powerful solutions, NLP stands out as a complementary method, exploiting advanced understanding of human language to extract rich semantic information. By integrating SpaCy, a modern NLP library, our research aims to go beyond the traditional limits of semantic annotation. The NLP techniques and SpaCy significantly contribute to our goal of semantically enriching medical data, thereby improving the understanding of complex medical concepts and supporting informed decision-making in healthcare. Unlike standard implementations, we enhance SpaCy capabilities by integrating domain-specific medical ontologies, custom rule-based phrase matchers, and pre-trained transformer models to improve entity recognition. This enhancement allows for high precision annotation and a better contextual understanding of medical texts, making it more adaptable to the complexities of EHR systems.

To validate our approach, we conduct quantitative performance evaluations, including:

- Precision, Recall, and F1-score for entity recognition and abbreviation expansion.
- Comparative analysis with alternative NLP libraries to justify model selection.
- Error analysis to refine NLP-based entity resolution

This method's approach performs several steps to detect and process abbreviations in text using SpaCy, the steps of this method are illustrated in the following Figure 3.

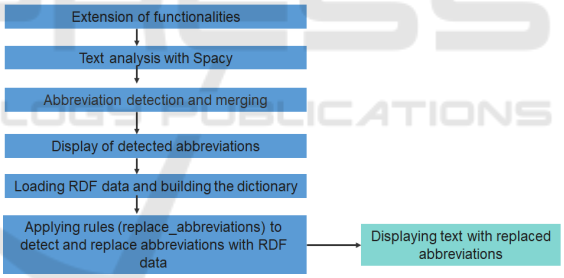

Figure 3: Model using NLP (SpaCy).

5 EXPERIMENTATION AND RESULTS

This section discusses the gathered results for each phase of the proposed system.

5.1 Result of Data Cleansing

Figure 4 shows the missing values in all columns, the 1st column 'class' shows that it has the highest number of values displaying NaN with a number of NaN is equal to 34, on the other hand the column 'properties' has no NaN missing values, and the other two columns 'subclass' and 'instances' have almost

the same number of missing values with a NaN count less than 10. Figure 4 illustrates the distribution of missing NaN values in all columns before the cleansing data. Then we replaced the NaNs with zeros, and then we replaced with a specific chain under the name 'Inconnu'.

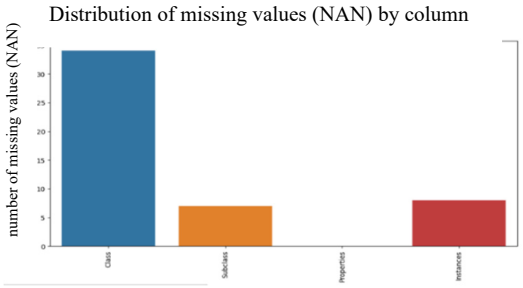

Figure 4: Screenshot of the distribution of missing NaN values in all columns before the cleansing data.

5.2 Results of Ontologies Construction

In this section we discuss the results of converting data from Excel to our local ontology. This phase constitutes the foundation of our approach, giving our medical data a rich and precise semantic representation. Figures 5 and 6 show an extract of the constructed ontology.

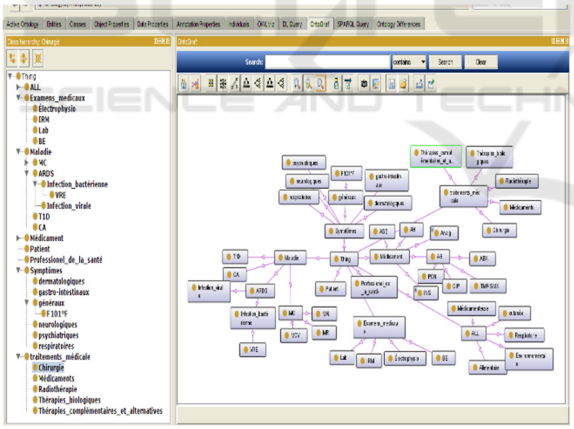

Figure 5: Ontology generated from Excel file

Figures 5 illustrate the ontology generated from an Excel dataset, showing how tabular attributes are transformed into semantic classes and organized hierarchically. Some classes are instantiated with terms in French, while others appear in English, reflecting the multilingual nature of the source data. This bilingual alignment allows the ontology to support annotation in both languages, enabling health records to be processed without ambiguity and ensuring interoperability across multilingual datasets.

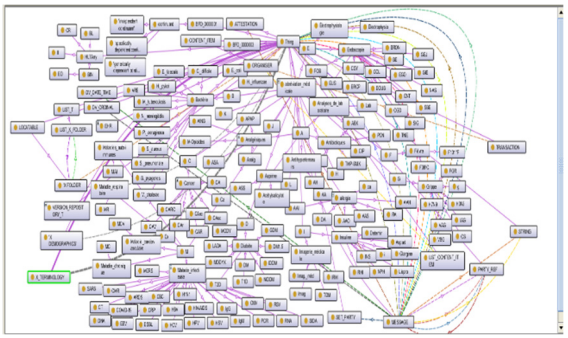

Figure 6: OntoGraf of the global ontology constructed.

Figure 6 OntoGraf of the global ontology constructed from integrated sources (EHR EXTRACT RM.owl, RIMV3OWL.owl, OGMS.owl) presents the graph of the unified ontology produced by merging local ontologies with domain ontologies. It highlights aligned classes, merged properties, and the extensions introduced (notably the Medical Abbreviations class and its links to full forms). This mid-level ontology acts as a semantic bridge between EHR structural schemas and medical domain concepts.

5.3 Abbreviation Detection Using Machine Learning

The following steps illustrate a comprehensive approach to solving the automatic abbreviation detection problem:

- Data preparation: Begins loading data containing phrases, abbreviations, meanings and binary functions. Then created the abbreviation to id dictionary to map the abbreviations to unique numeric indexes and added the abbreviation encoded column to encode the abbreviations and use them for the subsequent machine learning model.
- Data processing: For data processing, we have used scikit-learn to create a word vector using CountVectorizer. These word vectors are then combined using hstack with other properties such as uppercase letters, alphanumeric letters, and dot endings. The dataset is divided into a training and a test set to evaluate the performance of the models.
- Evaluation of the model: Several classification models such as RandomForest, KNN and DecisionTree are trained on the data. The performance of each model is evaluated using various metrics such as accuracy, mean absolute error, mean squared error and confusion matrix. These estimates are then visualized using Plotly Express to facilitate model operation.

The comparison shows that the Decision Tree classifier achieved the best performance, with an accuracy of 1.0 and average absolute and squared errors of 0.00. In second place, the Random Forest model obtained an accuracy of 0.93, an average absolute error of 0.06, and an average squared error of 0.26. Finally, the K-Nearest Neighbors (KNN) model achieved an accuracy of 0.83, an average absolute error of 0.17, and an average squared error of 0.41.

Although Decision Tree reached perfect accuracy on the dataset, accuracy alone is not always a sufficient metric for evaluating classifiers in the medical domain, where robustness and generalization are critical. For this reason, we also considered complementary metrics such as precision, recall, and F1-score, which provide a more balanced view of performance. Based on these metrics, Decision Tree remained highly effective, but Random Forest offered more stable generalization results.

The final choice of Random Forest as the selected model was motivated by several factors. First, unlike a single Decision Tree, Random Forest aggregates multiple trees, which reduces overfitting and improves robustness on unseen data. Second, in practice, the implementation of a Decision Tree with high complexity required more computational resources than were available in our environment, as well as additional expertise in fine-tuning and optimization. In contrast, Random Forest required less fine-tuning, was easier to implement, and offered strong performance while maintaining computational feasibility.

5.3.1 Training and Using the Model

The implementation of the Random Forest model for abbreviation detection involves a structured approach comprising several key steps. Initially, a pipeline is established, integrating a transformer (CountVectorizer) and a RandomForestClassifier model, facilitating efficient data preprocessing by converting text into word vectors. Subsequently, the dataset is partitioned into training and test sets, and the RandomForest model is trained on the training set. Following training, the model generates predictions on the test set, and its performance is evaluated based on accuracy. Additionally, the model is applied to new sentences to detect abbreviations, assigning labels to indicate their presence. This systematic approach showcases the integration of Random Forest into a natural language processing pipeline for abbreviation detection, emphasizing its role in data preparation, training, and prediction

5.3.2 Displaying Results

We indicated whether an abbreviation was detected in the new sentence and also displayed information about the detected abbreviation by replacing the abbreviation with its meaning using the replace abbreviations function.

```
La phrase "le médecin a demandé une EGD pour le patient." contient une abréviation : 1
********** ok abreviation detecte*********
Endoscopie .**********.......
Phrase: EGD, Abréviation détectée!

La phrase "le médecin a demandé une test IDDM pour le patient." contient une abréviation : 1
********** ok abreviation detecte*********
Diabète .**********.......
Phrase: IDDM, Abréviation détectée!
```

Figure 9: The results given by the Random Forest model.

Figure 9 shows our result given by the Random Forest model trained to predict if there is an abbreviation. If detected, the result displays the detected abbreviation and its meaning. Take as an example the first screenshot which shows that the abbreviation 'EGD' is detected and changes with their meaning. which is endoscopy. This section shows the resolution of the automatic abbreviation detection problem, integrating text processing, feature creation, model training and performance evaluation. It offers a complete and extensible solution for detecting abbreviations in similar contexts.

5.4 Method of Abbreviation Detection Using SpaCy a NLP Library

The method of automatic abbreviation detection was implemented using SpaCy, a natural language processing (NLP) library, combined with rule-based heuristics and semantic enrichment.

SpaCy was applied for text preprocessing and linguistic analysis, including tokenization, part-of-speech tagging, and dependency parsing. Abbreviation candidates were then identified through regular expression–based rules, focusing on patterns such as:

- uppercase sequences of up to five characters,
- contextual forms like "long form (short form)" (e.g., Blood Pressure (BP)),
- and co-occurrence of potential abbreviations with their expansions within the same text window.

To improve accuracy, the system performed frequency analysis of abbreviation–expansion pairs and validated them against ontology concepts. When

multiple expansions were possible, the one corresponding to an existing ontology class was selected.

Furthermore, we used RDF data to enrich the abbreviation dictionary. Each abbreviation was annotated with semantic information such as domain, synonyms, and related ontology class.

In the final step, abbreviations were automatically replaced by their expanded forms, improving both readability and semantic clarity. This process is particularly beneficial for medical record analysis, clinical reporting, and advanced biomedical research.

Figures 10 illustrate the overall workflow, showing how SpaCy, regular expressions, and RDF-based enrichment were combined to detect, expand, annotate, and replace abbreviations in the corpus.

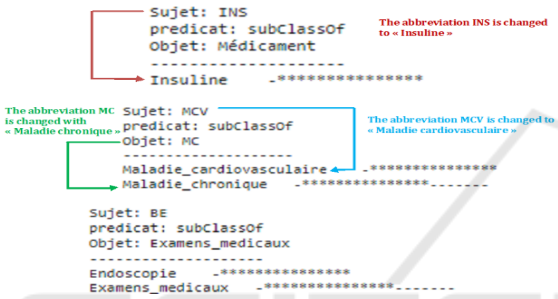

Figure 10: Screenshot of some abbreviations with their text replaced given by SpaCy.

6 COMPARISON STUDY

We have done a comparison between the proposed framework and other previous frameworks.

The machine learning-based approach is more suited to complex tasks requiring generalization from labeled data, while the NLP approach with SpaCy is more transparent, easily adaptable, and well suited to specific tasks like abbreviation detection. We have done a comparison between the proposed framework and other previous frameworks. Table 1 summaries that comparison. This comparison highlights the different approaches, their strengths and weaknesses, allowing a comparative assessment of the methodologies adopted for medical data interoperability and semantic annotation. Our approach stands out significantly from others on several fronts, attesting to its reliability and effectiveness. In terms of semantic annotation, our methodology relies on advanced NLP and machine learning techniques, thus surpassing the competing approach which does not offer similar semantic functionalities. The flexibility of our system is demonstrated by its ability to adapt to the nuances inherent in medical data, thus surpassing the average flexibility offered by other approaches. Regarding model training, our approach excels by integrating a robust process, unlike the other method which neglects this crucial dimension. In terms of ontological interoperability, our approach confirms its advantage by guaranteeing semantic consistency between different data sources. And in terms of complexity, our approach maintains a careful balance between sophistication and practicality, thus positioning itself favorably compared to other approaches, characterized by high complexity. Finally, from the performance point of view, our approach excels with a high evaluation, while the other approaches show moderate performance. These substantial differences illustrate the increased reliability and overall superiority of our approach.

Although the proposed approach achieves syntactic interoperability in distributed EHRs, it contains many perspectives. First, integration of new data sources, we are expanding our approach to include a variety of medical data sources, such as medical images, imaging reports and genomic data. This would contribute to a more complete presentation of medical information. Second, much of the knowledge in the field of medicine is unclear, so we have to deal with incomplete and unclear problems in the field. Therefore, the presence of a specialist in this field is a plus.

7 CONCLUSIONS

The complex terminology used by specialists in the medical field often creates a barrier for anyone seeking to understand health-related information. Health users need to decipher this ‘‘jargon‘‘to better manage their situation. In the context of EHR, hospitals and physicians face major challenges in effectively sharing the information needed for quality, timely, and costeffective care. In this approach, we set out to solve the complex challenge of EHR interoperability by combining approaches based on machine learning and NLP. Our main objective was to improve the semantic annotation of medical data, thus facilitating interoperability between different healthcare systems. To achieve this goal, we followed an integrated approach, which built a local and global ontology that helped improve interoperability by standardizing the representation of medical concepts. Our approach demonstrated notable efficiency in semantic annotation, providing

Table 1: A Comparison Study of the Proposed Framework and Other Previous Ones.

Criteria	Jaleel et al. [7]	EL hajjamy et al [21]	Plastiras et al [22]	Kiourtis et al [23]	Proposed approach
Year	2020	2018	2014	2017	2023
Methodology	Medical Data Interoperability through collaboration of healthcare devices (MeDIC)	Global crisp ontological	ontology based on the HL7 Message Information model and extended it to include essential PHR requirements	Ontology mapping	Integration of machine learning techniques into ontology
Data formats	JSON, XML,Text	UML , XML, RDB	XML	Different EHR standard (e.g.HL7 v2,HL7 v3, HL7 FHIR ,etc)	Excel file(CSV, xlxs)
Semantic annotation	No	No	No	No	Yes
Flexibility	High	Medium	Medium	Medium	High
Model training	No	No	No	No	Yes
Ontological interoperability	Yes	Yes	Yes	Yes	Yes
Complexity	High	Medium	Medium	High	Medium
Performance	High	Medium	Low	High	High

increased flexibility and better adaptation to variations in medical data. Machine learning models helped with accurate abbreviation detection, while the use of SpaCy enhanced contextual understanding of medical terms. The results show that Our proposed ontological model offers a significant contribution to understanding and resolving the challenges surrounding semantic annotation of electronic health records. Also it showed significant benefits in data cleansing and medical data interoperability. By combining robust ontological elements, machine learning and NLP approaches, our approach aspires to improve the efficiency and accuracy of health systems through smarter management of medical data, paving the way for significant advances in understanding and management medical data, reducing medical errors and guaranteed interoperability. For future work, we need to create a graphical user interface to easily use the implemented framework and Further development of the framework will concentrate on the limitations of this work, as discussed previously.

REFERENCES

Adel, E., El-Sappagh, S., Barakat, S., Kwak, K., Elmogy, M. (2022). Semantic architecture for interoperability in distributed healthcare systems. *IEEE Access*, **10**, 126161–126179.

Begoyan, A. (2007). An overview of interoperability standards for electronic health records. *Society for Design and Process Science*, USA.

Berges, I., Bermúdez, J., Illarramendi, A. (2011). Toward semantic interoperability of electronic health records. *IEEE Transactions on Information Technology in Biomedicine*, **16**, 424–431.

Celesti, A., Fazio, M., Romano, A., Villari, M. (2016). A hospital cloud-based archival information system for the efficient management of HL7 big data. In *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 406–411.

Chen, L. Li, Lu, D., Zhu, M., Muzammal, M., Samuel, O., Huang, G., Li, W., Wu, H. (2019). OMDP: An ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems. *International Journal of Distributed Sensor Networks*, **15**, 1550147719847112.

Costa, C., Menárguez-Tortosa, M., Fernández-Breis, J. (2011). Clinical data interoperability based on archetype transformation. *Journal of Biomedical Informatics*, **44**, 869–880.

Costa, C. Da, Wichman, M., Rosa Righi, R., Yamin, A. (2019). Ontology-based model for interoperability between openEHR and HL7 health applications. In Proceedings of the International Conference in Health.

Fennelly, O., Moroney, D., Doyle, M., Eustace-Cook, J., Hughes, M. (2024). Key interoperability factors for patient portals and electronic health records: A scoping review. *International Journal of Medical Informatics*, 105335.

González-Ferrer, A., Peleg, M., Verhees, B., Verlinden, J., Marcos, C. (2012). Data integration for clinical decision support based on openEHR archetypes and HL7 virtual medical record. In *International Workshop on Process-oriented Information Systems in Healthcare*, pp. 71–84.

Gunter, T., Terry, N. (2005). The emergence of national electronic health record architectures in the United

States and Australia: models, costs, and questions. *Journal of Medical Internet Research*, **7**, e383.

Hajjamy, O. El, Alaoui, L., Bahaj, M. (2018). Integration of heterogeneous classical data sources in an ontological database. In Big Data, Cloud and Applications: Third International Conference, BDCA 2018, Kenitra, Morocco, pp. 417–432.

Kalra, D. (2006). *Electronic health record standards*. Schattauer GMBH-Verlag.

Kiourtis, A., Mavrogiorgou, A., Kyriazis, D. (2017). Aggregating heterogeneous health data through an ontological common health language. In *2017 10th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 175–181.

Liyanage, H., Krause, P., De Lusignan, S. (2015). Using ontologies to improve semantic interoperability in health data. *BMJ Health & Care Informatics*, **22**.

McClanahan, K. (2008). Balancing good intentions: protecting the privacy of electronic health information. *Bulletin of Science, Technology & Society*, **28**, 69–79.

Mylka, A., Kryza, B., Kitowski, J. (2012). Integration of heterogeneous data sources in an ontological knowledge base. *Computing & Informatics*, **31**.

Roehrs, A., Costa, C., Rosa Righi, R., Rigo, S., Wichman, M. (2018). Toward a model for personal health record interoperability. *IEEE Journal of Biomedical and Health Informatics*, **23**, 867–873.

Santos, M., Bax, M., Kalra, D. (2010). Building a logical EHR architecture based on ISO 13606 standard and semantic web technologies. In *MEDINFO 2010*, pp. 161–165.

Sartipi, K., Dehmoobad, A. (2008). Cross-domain information and service interoperability. In *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, pp. 25–32.

Sachdeva, S., Bhalla, S. (2022). Using knowledge graph structures for semantic interoperability in electronic health records data exchanges. *Information*, **13**, 52.

Schloeffel, P., Beale, T., Hayworth, G., Heard, S., Leslie, H., et al. (2006). The relationship between CEN 13606, HL7, and openEHR. In *HIC 2006 and HINZ 2006: Proceedings*, p. 24.

Sreenivasan, M., Chacko, A. (2020). A case for semantic annotation of EHR. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1363–1367.

Plastiras, P., O'Sullivan, D., Weller, P. (2014). An ontology-driven information model for interoperability of personal and electronic health records