UNIVERSITY OF GENOVA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

Medical Data Management to enable the use of Machine Learning-based systems

by

Sara Mora

Thesis submitted for the degree of *Doctor of Philosophy* (35° cycle)

June 2023

Mauro Giacomini	Supervisor
Gabriele Arnulfo	Supervisor
Paolo Massobrio	Head of the PhD program

Thesis Jury:

Jaime Delgado, *Universitat Politècnica de Cataluny, Spain*	External examiner
Lenka Lhotská, *University in Prague, Czech Republic*	External examiner
Laura Pastorino, *University of Genoa, Italy*	Internal examiner

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

Abstract

Objective. Differential diagnosis is the process that formulates a precise and accurate diagnosis on a patient. However, in most of the cases it is performed at first admission in a primary-care unit and it mainly relies on physician's experience. Delays in the diagnosis formulation and mistakes may lead to serious complications. The aim of my PhD project is to improve differential diagnosis by developing a system able to automatically extract both structured and unstructured data to enable the use of Machine Learning (ML) methods on Real World Data (RWD), i.e., data collected during daily clinical practice outside traditional interventional controlled clinical trials.

Approach. I used standard-based novel architectures to extract structured data from the hospital Laboratory Information System (LIS) and transfer them into a SQL Server database using the already existing architecture of the Ligurian Infectious Diseases Network (LIDN). Then I used NLP-based methods to build the most appropriate numerical representations for textual data, manually extracted and anonymized from Electronic Medical Records. To test the efficacy of the proposed pipeline, I used both structured and unstructured data coming from two different medical scenarios as input of the developed ML-pipeline to support diagnosis process.

Main results. The first main result has been building an intelligent system able to extract a dataset of 285 features based on structured RWD of a selected group of patients with a diagnosis of candidemia/bacteremia. As I obtained each feature performing a re-elaboration of stored data, the outcome of the rules-based system needed to be validated. Specifically, clinicians manually validated results of 381 patients randomly selected from the cohort and attesting that each of the selected features presented an error $< 1\%$. The second main result has been developing an NLP and ML-based pipeline able to transform free texts into the most appropriate numerical representation, using Bag of Words or Word Embedding techniques, to enable text classification or information extraction tasks. The first use case aimed at localizing the Epileptogenic Zone in drug-resistant epilepsy patients using the textual data of the semiological descriptions of seizures. I proved that all the numerical representations built by the pipeline accurately (F1-score up to 0.78 on blind set) localized the seizure onset zone. The second use case aimed at extracting information related to the possible presence of Central Venous Catheter (CVC) implanted at the diagnosis of candidemia to build a more complete picture of the patient. To do that I used the clinical notes written by medical staff in a limited time span around the diagnosis. The developed pipeline reached mean values of F1-score up to 0.92 in determining if a patient had CVC implanted and up to 0.84 in determining if CVC was removed, both results are obtained on a blind test set. The third main result derives from the features selection applied to the complete dataset, composed by structured and unstructured data related to the use case candidemia/bacteremia, involved in a majority voting process. My results confirm that CVC feature has a great impact (selected 100% of times, mean coefficient value in LASSO matrix is 0.12) on the outcome infection of invasive candidiasis.

Significance. The developed NLP and ML-based pipeline accurately identifies EZ location and the presence of CVC from text alone. The main advantage is that it does not contain any specific information about the medical discipline, so it can be easily used in other scenarios, and it is based on Italian text. In general, the complete architecture exploits the paradigm of data reuse to support differential diagnosis, so in the future an always growing amount of data will be available.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AI	Artificial Intelligence
ARCA	Antiretroviral Resistance Cohort Analysis
CDA R2	Clinical Document Architecture Release 2
CDM	Clinical Data Management
CDMS	Clinical Data Management System
CISAI	Italian Coordination for the Study of Allergies and HIV Infections
CRF	Case Report Form
CTS2	Common Terminology Service Release 2
CVC	Central Venous Catheter
DPCM	Decree of the President of the Council of Ministers
DRS	Drug-Resistant Seizures
EHR	Electronic Health Record
EMR	Electronic Medical Record
EZ	Epileptogenic Zone
HCV	Hepatitis C Virus
HIV	Human Immunodeficiency Virus
HL7	Health Level 7

HSSP	Healthcare Services Specification Project
ICONA	Italian Cohort Naive Antiretrovirals
IE	Information Extraction
ISS	Istituto Superiore di Sanità
KRLS	Kernel Recursive Least Squares
LASSO	Least Absolute Shrinkage and Selection Operator
LIDN	Ligurian Infectious Diseases Network
LIS	Laboratory Information System
LOINC	Logical Observations Identifiers Names and Codes
LR	Logistic Regression
ML	Machine Learning
NEHR	National Electronic Health Record
NHSN	National Healthcare Safety Network
NLP	Natural Language Processing
NLTK	Natural Language ToolKit
OMG	Object Management Group
RF	Random Forest
RWD	Real World Data
SNOMED-CT	Systematized Nomenclature of Medicine—Clinical Terms
SVM	Support Vector Machines
TB	Tuberculosis
VEEG	Video-electroencephalography
WHO	World Health Organization

Part I

Introduction

Chapter 1

Background and Significance

Differential diagnosis is the process that formulates a precise and accurate diagnosis on a patient by distinguishing between common clinical signs and diseases with similar manifestation. It is mainly conceived as a subjective process primarily relying on physician's experience. At first, clinicians must list all possible diseases that explain patient's clinical signs and symptoms, then multi-modal and multi-scale data are collected, e.g., historical information and laboratory exams results. Often differential diagnosis is performed at first admission in primary-care units, mainly relying on observational information by general/non-specialized medical doctors. However, delays in the diagnosis formulation and misdiagnosis, reported in literature as ranging from 10% to 15% [16], can lead to serious complications in several medical domains [135, 34, 66, 206].

During the last years, *Machine Learning* (ML) approaches have been used to build diagnosis supporting systems and encouraging results have been obtained in several areas, such as early diagnosis, prognosis and development of new therapies [132].

However, the phenomena of misdiagnosis is even more evident in those branches of medicine where the conduction of clinical trials is somehow hindered by the high presence of meaningful information in the textual sections of the *Electronic Medical Records* (EMR)s [20]. This made necessary to investigate techniques that allow to deal with natural language text [59, 8, 113, 177, 44].

So, the development of an advanced system able to support the process of differential diagnosis at different stages, using heterogeneous data, could significantly impact patients' quality of life by reducing the time necessary to achieve the best individual treatment.

1.1 How can I obtain data necessary to run ML models?

1.1.1 Where does clinical data come from?

Research in medical field mainly relies on clinical data, which can derive from either:

- **Formal clinical trial programs**. This class of data is collected through ad-hoc *Case Report Form* (CRF), which can be electronic or paper-based and results are later reported on a computer.
- **All procedures that concern patient care**. This class of data is called *Real World Data* (RWD), i.e., data collected during daily clinical practice outside traditional interventional controlled clinical trials. It includes administrative data, claims data, patient/discharge registries, health surveys, and of course the most rich source of patient's data which is the EMR.

Data collected according to the first modality have the following main advantages: only the necessary data are available; the same set of data is collected for each patient; they are ready to use as they already are in the correct usable format. While, the main disadvantages are: they are manually entered, so they may contain errors; they are limited in dimension as their collection depends on human and timing resources; the sample may not be representative of the whole population as patients are selected so it may not entirely represent the complexity of real scenario; necessary features should be present from the beginning in the CRF [139, 160, 157].

On the contrary, the EMRs are a collection of multidimensional and heterogeneous retrospective data that are often used to conduct clinical research with the purpose of diagnostic analysis. From EMRs it is possible to extract demographic factors, clinical variables, information related to the drugs treatments, and patient's morbidity and mortality [43, 65, 116, 130, 146]. This kind of data source presents some advantages: data availability is limited only by the number of patients that is cured in the specific hospital and not by any temporal or human resource; data can be easily retrieved even in a second moment (after a new approval by the ethics committee). However, it presents also some disadvantages: data are in a format that could be not directly usable (semi-structured and non-structured); they may present a low quality or being incomplete; data collection performed not following the same settings and criteria for all patients [109, 191].

1.1.2 Why is Clinical Data Management an important step?

Clinical Data Management (CDM) is a critical phase in clinical research, which leads to the generation of high-quality, reliable, and statistically sound data from and for clinical trials [92, 195]. Specifically, the professional in the clinical data management is actively involved in all stages of a clinical trial right from inception to completion. Various procedures are included in CDM, among all: CRF designing and/or features definition, database designing, data-entry, data validation, discrepancy management, medical coding, data extraction, and database locking. Some of these procedures are performed periodically in order to maintain the appropriate level of quality.

1.2 Which kind of data am I looking for?

Over the last three decades EHRs have been extensively adopted as great source of valuable information for both patient care and biomedical research [36, 56]. They facilitate the storage of data that can be queried and processed in an automatic way. Laboratory exams results are an example of such data as they are produced by medical machines and automatically stored in the *Laboratory Information System* (LIS).

So, certainly the advent of EHRs constituted a turning point in the conduction of retrospective clinical trials, compared to the same process of information search on paper based reports [36, 193]. However, the secondary use of above the half of meaningful information stored into the EMRs is limited by the high presence of unstructured text, *e.g.,* discharge letters, nurses and physicians notes [108, 201]. This kind of data are often discarded for research because they need manual inspection and so it would require a huge effort from a human and timing point of view. It represents even a bigger issue for those branches of medicine that make an extensive use of free text. This problem, which is not only restricted to the clinical scenario[1] [149], contributed to increase the interest in the field of text mining and *Natural Language Processing* (NLP) [39].

[79] describes NLP as follows: "Natural language processing is the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content". So, NLP is a research and application area that investigate how natural language text can be understood and manipulated by computers. It is involved in several applications, such as text summarization, text translation, sentiment analysis, information extraction, speech recognition, etc.

[1] https://www.forbes.com/sites/forbestechcouncil/2019/01/29/the-80-blind-spot-are-you-ignoring-unstructured-organizational-data/

However, it should be considered that the application of NLP tools developed for other more common scenarios is not straightforward on clinical text [39]. First of all, because clinical texts are written using a highly specialized terminology and they also may contain acronyms and shared specific abbreviations, often recognised only at local level. Then, even though texts are mainly written through a user interface they may contain misspellings [162, 183] and the grammatical structure of the text may be incomplete to speed up the writing process, as the clinical practice occupies the majority of a healthcare employ time.

1.3 Aim of the thesis

The main aim of my thesis is to explore the potential of the usage of unstructured data in the process of differential diagnosis and the need to combine multiple source of information to obtain a more complete picture of the patient. The two use cases that I considered for my PhD project involve the Infectious Diseases and Neuropsychiatry medical areas. This choice was guided by the need to improve research in these two fields as they both have a consistent portion of meaningful information available in an unstructured format and this represents an obstacle in their usage for research purposes. So, within the project I dealt with RWD, both structured and unstructured, derived from EMRs. One of the main contributions of my work is a ML and NLP-based pipeline aimed at assigning a specific label to each piece of text in input. A very important aspect is that, as I did not include any specific medical knowledge, therefore, it could be used in multiple tasks. Specifically, I used it in two ways:

- **Information Extraction task.** It aimed at building a more complete patient's picture by extracting from the unstructured sections of the EMRs information that could not be found in any analytical field of the database.
- **Text Classification task.** It consisted in assigning each patient to a certain group considering as input only natural language text.

One of the main challenges that I faced during the development of my PhD project is that the majority of text processing tools and libraries are specific for English language, which is overall more widespread, while the availability of such tools is more restricted for text written in other minor languages, e.g., Italian. In addition to that, no Italian ontologies on the specific topics that I faced are available. An ontology is an "explicit specification of a conceptualisation" [74] and it provides links between concepts and relations that are precious in an information processing task [53]. In particular, an ontology would be useful in the text normalization process, first because it is a collection of concepts and words strictly linked to the specific context and then it also provides properties of these concepts, for example known synonyms.

Part II

Methods

Research is a term that can be divided in two sub-words: $Re + Search$ where *Re* indicates a repetition and *Search* to examine closely. So, research can be described as the process that leads to the formulation of a conclusion through the analysis of data collected during the repeated observation of a phenomena.

According to George J. Mouly [129] research can be defined as:

"The systematic and scholarly application of the scientific method interpreted in its broader sense, to the solution of social studiesal problems; conversely, any systematic study designed to promote the development of social studies as a science can be considered research."

Research can also be described from a philosophical point of view and the physical sciences are generally associated with positivism [97, 33]. Science is defined as:

Objective, reality can be described from an objective point of view and the researcher should not interfere with it or introduce any bias.

Deterministic, reality follows causality laws, so events are connected by a cause-effect relationship.

Mechanistic, reality can be explained through a hypothetico-deductive approach.

Using methods, the hypothesis can be operationalized, *i.e.,* the sample is selected, measures and analysis are performed, conclusions about the phenomena and the hypothesis are formulated.

According to this approach, the hypothesis can not only explain a phenomenon but also predict it [181].

The criticisms leveled at the positivist vision that emerged in the 1900s do not influence this specific study as the object under consideration does not deal with issues to which the mechanistic view is not applicable [78].

Present research work is designed a deductive process, it starts with a hypothesis, called *Research Question*, and aims at testing it using data [175, 89]. It is configured as a quantitative experiment exploiting a top-down approach [178].

The research question can be summarized as "What the potential of the usage of unstructured data in the process of differential diagnosis?" and it is declined in two use cases. It is addressed using observations and, specifically, RWD extracted from patients' EMRs.

The overall project is configured as an analytical, observational and retrospective study [159, 150]. Within each use case, the research question aims at discriminating between two groups of patients based on an outcome which is established and has already occurred.

As sampling strategy, both use cases involve a random sample from a target population. As a large number of samples is required but not available, other mathematical strategies are used to evauate the generalisability of findings.

Chapter 2

Structured Data Management

2.1 Data collection: Medical informatics and standards support for Clinical Research

It is 30 years since evidence-based medicine became a great support for individual clinical expertise in daily practice and scientific research [165]. However, this required the need to conduct several clinical trials [55] and to collect a huge amount of data. To achieve this objective, the development of systems able to support first the coordination of clinical trials and then the interoperability of heterogeneous data originating from different medical centers [102], was necessary. These kinds of systems are indispensable not only to control the huge amount of data produced by health facilities but also to build clinical pathways, research systems, and effective public health management policies. Since the beginning of the 2000s, the ability to describe heterogeneous data through the use of different kinds of models and to connect them to the available web forms has led to the setup of web applications that facilitate the exchange of research data in many fields [71, 111, 138, 84, 35]. Multi-center research networks and CDM systems (CDMSs) have played an important role in data storing and management within a varying range of medical domains [52].

2.1.1 Automatic LIS data export

In multi-center or long-lasting clinical trials, the use of CDMSs has become essential to handle the huge amount of data and several international projects started working on this topic. Specifically, in order to collect the necessary sample to conduct the study, usually physicians manually insert data through a dedicated web user-interface or collect them into predefined spreadsheets. However, during the years the rise of the well-know problems caused by manual data input [75, 100, 104] and the huge amount of time and human effort necessary to complete the data collection phase made an evolution of this process necessary. Specifically, structured data such as laboratory test results are already stored in the hospital LIS in digital format and so they can be easily read and automatically transferred towards other databases. Thus, the paradigm of data reuse allows the collection of patients' clinical data without human intervention. However, this process of data transfer is not straightforward as it depends on the specific hospital level of informatization:

1. The hospital does not have a LIS which stored data in a digital format, or it have a LIS, but external agents can not obtain access to data.

2. The hospital have a LIS and external authorized agents can access a specific subset of data.

Centers that belongs to the first scenario can take part in clinical trials only collecting data manually, e.g., coping them through a web-user interface. On the contrary, centers belonging to the second scenario can exploit the data reuse paradigm. However, the automatic transfer of data also depends on other factors, e.g., the different ways according to patients' data are exposed and to the national regulation scenario. Specifically, hospital may expose patients' data through services or an ad-hoc view on the hospital LIS that can be reached only through Virtual Private Network. Both methods are regulated by authentication phase. Nowadays, Italian regulations, specifically the *Decree of the President of the Council of Ministers* (DPCM) of September 2015, titled "Regulation on National Electronic Health Records (NEHR)" [1], identify the Health Level 7 Clinical Document Architecture release 2 (HL7 v3 CDA r2) as the standard schema that defines the structure and semantics of clinical documents and messages. This standard is used at national level to stored clinical data into the National Electronic Health Record (NEHR).

2.1.2 Use of medical standards to enable interoperability

One of the main problems when multi-center or long lasting clinical trials are executed is that nomenclature may be different. This is due to the fact that each medical center developed its own nomenclature which may also change during the years for several reasons, e.g., new diagnostic machines or changes in the way exams are reported. This lead to the need of introducing international standard coding system, e.g., vocabularies and ontologies. Together

[1] https://www.agid.gov.it/sites/default/files/repository_- files/linee_guida/dpcm_178_2015.pdf

with the necessity of collecting comparable results, the use of standard coding systems has become also mandatory at national level in Italy. Specifically, one of the requirements of the aforementioned DPCM of 2015 is the univocal interpretation of clinical data.

So, I adopted:

- The international vocabulary *Logical Observation Identifiers Names and Codes* (LOINC) to translate the local laboratory procedures names

- The national and international coding systems *Italian Clinical Microbiologists Association* (AMCLI), *Systematized Nomenclature of Medicine—Clinical Terms* (SNOMED-CT), *National Healthcare Safety Network* (NHSN) to identify unequivocally the the name of the microorganism found in the microbiological cultures results.

In order to maintain this mapping update as new versions of international coding system are released, it is necessary to involve terminology services, e.g., services built according to the standard *Common Terminology Service Release 2* (CTS2) developed within the *Healthcare Services Specification Project* (HSSP). An example of this kind of services can be found in [64].

2.1.3 Example of a ten years-old CDMS: the Liguria Infectious Diseases Network

The *Ligurian Infectious Diseases Network* (LIDN) [72, 128] started in 2011 as a new research network at a regional level during a scientific collaboration between medical groups and bioengineers. At the very beginning, the LIDN was a web platform aimed at enabling the easy collection of *Human Immunodeficiency Virus* (HIV)-infected patient data in order to conduct multicenter clinical trials at a regional level. The first one involved the drug Maraviroc [163]. Medical experts manually copied data from the Electronic Health Records (EHRs) to the LIDN through a web user interface [2]. During the first year, the platform supported several regional and national studies. However, frequent use of the platform led to the rise of two main issues: first, the huge amount of time required to insert even a minimal set of data necessary to conduct a clinical trial and, second, the randomness of errors induced by manual data imputation. Therefore, system architecture evolved in order to exploit the data reuse paradigm. So, laboratory test results started to be automatically transferred from hospital LISs towards the LIDN database. Thus, this allowed the daily update of patients' clinical data without human intervention. The bioengineers involved in the project decided

[2] https://www.reteligurehiv.it/

to start from laboratory data since they are the largest dataset used in scientific research about infectious diseases. Once the LIDN obtained a completely updated database, it was planned to exploit this system also to avoid manual data entry in other databases such as *Antiretroviral Resistance Cohort Analysis* (ARCA) [3], *Italian Cohort Naive Antiretrovirals* (ICONA), and *Italian Coordination for the Study of Allergies and HIV Infections* (CISAI) [4]. These are the most important cohorts for clinical studies concerning HIV at a national level. The architecture modular structure allowed during the years the expansion of the group of the monitored infectious diseases. Specifically, it started with HIV, then *Hepatitis C Virus* (HCV) and Tuberculosis (TB) were added. Finally, when the *COrona VIrus Disease 19* (COVID-19) pandemic arose, the architecture was rapidly expanded to store data for this new class of patients as well [187, 123]. Specific approval by the Ligurian Ethics Committee was requested and obtained (163/2020–10475).

2.2 Missing Data Management

The first step that needs to be addressed when dealing with a dataset, especially when derived from RWD, is data pre-processing in order to clean the data and make them useful for any experiment associated with ML or data mining. Usually, one of the main issues is the presence of missing values, which can be faced in both in features and in observations (size reduction). This problem could be addressed in two ways: list-wise deletion (or complete case analysis) by removing all the samples with missing values from the dataset, or imputation methods by estimating the value of missing data. At both row and columns levels, it is a common practice to exclude from the analysis those features and observations with a percentage of missing values higher than the 50%. However, when the percentage of missing values is lower than this threshold, to keep the sample as large as possible, the second option can be considered. According to literature, a generally effective algorithm in this scenario of data imputation is based on K-Nearest Neighbor, which imputes each sample's missing values using the mean value from *K* nearest neighbors found in the dataset [115].

[3] ARCA (Antiretroviral Resistance Cohort Analysis): https://www.fondazioneicona.org/

[4] CISAI Coordinamento Italiano Studio Allergie e Infezioni da HIV: https://www.cisai.it/

Chapter 3

Unstructured Data Management

3.1 Data collection

Together with the increasing adoption of data extracted from EMRs also patient privacy concerns arose, especially when dealing with unstructured data [172]. Natural language texts contain more sensitive information comparing to structured data as a consequence of their nature, but they could also contain information that unequivocally identify a person, such as first name, last name, date of birth, etc. So, in order to use them in clinical research but at the same time preserve patient privacy and be compliant with existing regulatory laws, such as the Health Insurance and Portability and Accountability Act (HIPAA), it is necessary to perform a de-identification process [103, 40]. However, it is not possible to exclude the sensitive information at the extraction level from EMRs as it happens for structured data, e.g., excluding the corresponding columns from database tables. So, data should be extracted as they are and then undergo a de-identification process, but there are several issues linked to the execution of this task automatically. First, currently, to the best of my knowledge, only scattered studies have been performed [25] and there are not open source tools able to perform de-identification task on texts written in Italian language. Second, even if a such system existed, it should be run locally because data containing sensitive information should not be stored on other external services. Last, to build an ad-hoc pipeline, a new information consent should be signed by patients and collected by clinicians. So, in order to preserve patients privacy and to go on with the research, data were collected already de-identified manually by medical staff involved in the projects. Then, this kind of data was stored in SQL server databases created ad-hoc for each project. According to my experience, even if this process requires human intervention at the beginning to extract data from the textual files where they are contained and store them into database fields, it is worth. Using databases to store unstructured data allows to manage them more easily. For example, through the use of simple queries it is possible to include or exclude sections of narrative text from the analysis or apply the same pipeline in two different tasks keeping the same code but only changing the query.

3.2 Data pre-processing

3.2.1 Data cleaning

Data cleaning is usually the first step and it is performed with the aim of improving data quality, e.g., by correcting the detected misspellings and substituting the abbreviations with default values. Currently, common agreement defining data cleaning is not available, and this is due to the different requirements of the projects [87]. Considering the available sample, data were cleaned looking for regular expressions and 're' python module was used. First all text was made lowercase and then from the sentences the following elements were removed: patterns containing numbers (e.g.: dates or names of electrodes), punctuation, text in brackets. Then shared specific abbreviations used by clinicians in their daily practice were extended (e.g.: 'aass' means upper limbs, 'aoo' means open eyes).

3.2.2 Tokenization

Consists in dividing the sequences of characters in minimum units of analysis called tokens, which include various categories of parts of the text [192]. To perform this task, *'Natural Language ToolKit'* (NLTK) library was used as it is the most used to perform text analysis in multiple languages [17].

3.2.3 Lemmatization

Consists in assigning to each word its base form (called lemma). This process contributes to make the context of the text uniform, it can be considered as a normalization process [94]. To perform this task, two different tools in Python language were tested: 'spaCy'[1] and 'TreeTaggerWrapper'[2]. The second one was chosen because it obtained a better performances on texts written in Italian language. It is worth to perform this step if the syntax of the sentence is complete.

[1] https://spacy.io/models/it

[2] https://treetaggerwrapper.readthedocs.io/en/latest/

3.2.4 Stop-words removal

Consists in removing from the texts the most common words with the aim of increasing the discrimination between the considered text [185]. In the Italian language examples of stop words are articles, prepositions, etc., as they are not very informative and may alter other processes, e.g. most common patterns identification.

3.3 Information Extraction

Information Extraction (IE) is the sub-field of NLP aimed at retrieving specific information from unstructured natural language texts by automatically processing them [39, 125, 196, 121]. This NLP technique is frequently used in the clinical research scenario to extract meaningful information from the narrative texts contained in the EMRs.

3.3.1 Rule-based methods

Historically, one of the first approaches to text analysis is the use of rule-based methods. They consist in a set of a-priori known rules imposed to a program to make it behave in a specific desired way. There is no specific format for rules, they can be for example grammars used to parse text or regular expressions [39].

Grammars

Grammar is set of rules that define the structure of perfect well-formed sentences within a language. It specifies the linguistic elements which constitute a sentence and the structural relations between them [30]. Parser is a computational elements which executes the grammar rules and creates the syntactic tree, which describes the role of each element in a sentence.

A base form of parsing is chuncking [1]. Specifically, according to this method, each sentence can be divided in constituents. They are groups of words acting as a single unit. They can be: noun phrases (NP), verb phrases (VP) and preprosition phrases (PP). Each sentence S can be described as:

$$S = NP + VP$$

where a NP can be for example a pronoun or a proper-noun, while a VP can be, for example, a single verb or a verb and a NP.

Regular Expressions

Among them, Regular Expressions allow the extraction of predefined patterns from text [60]. Specifically, regular expressions, also called regex, are a powerful instrument able to perform string matching considering a combination of several elements in the text, *e.g.,* words, characters, numbers, etc.

In the clinical scenario, they can be used to extract personal information of a patient that follow a syntactic rule, such as identification numbers or anamnestic information [82, 131]. For example, the length of an Italian fiscal code is 16 characters, composed by 9 letters and 7 numbers. Its structure is: 3 letters extracted from the last name; 3 letters extracted from the first name; 2 digits corresponding to the last two numbers of year of birth; 1 letter corresponding to the month of birth; 2 digits corresponding to a transformation of the day of birth; 4 characters (1 letter and 3 digits) corresponding to the municipality of birth code; 1 letter corresponding to the control character.

So, it can be extracted using the following regex:

```
[A-Z]{6}\d{2}[A-Z]\d{2}\w{4}[A-Z]
```

However, since regex are based on predefined and fixed patterns, their usage is limited when the complexity of text is high, especially in the clinical scenario [133].

Fuzzy Logic

Fuzzy Logic is a computational-based approach which involves the concept of "degrees of truth". So, the truth value is not a Boolean restricted to 0 and 1, but it can assume each decimal position between them [205, 204].

Fuzzy Wuzzy[3] is the Python library used within this project to manage the comparisons between strings. Specifically, it is used to compute the distance between two strings with the same number of characters or not, taking into account the order of words and the allowed maximum frequency of a string. This comparison is based on the Levenshtein distance [107].

[3] https://pypi.org/project/fuzzywuzzy/

$$lev_{a,b}(i,j) = \begin{cases} max(i,j), & \text{if min(i,j) = 0} \\ \\ min\begin{cases} lev_{a,b}(i-1,j)+1, \\ lev_{a,b}(i,j-1)+1, \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq a_j)} \end{cases} & \text{otherwise} \end{cases} \tag{3.1}$$

where i and j are the indexes of the last character of the substring.

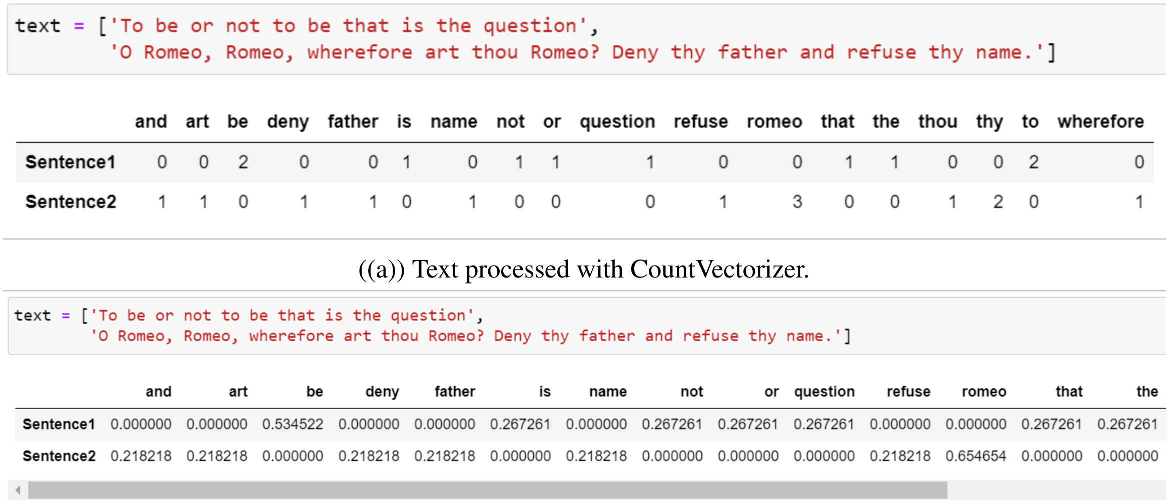
According to a review performed in 2018 [190], the 65% of article on clinical text mining used rule-based methods. However, their application has two main disadvantages. First of all, rules need to be constructed by experts in the specific clinical domain. Then, as they are domain specific, it is necessary to have a knowledge on the input text and their performance is prone to overfitting (high precision and low recall) [39, 143].

Therefore, NLP started to be used in combination with ML models [133, 147, 200].

3.4 Text Classification

During the last years, text classification has been widely studied, not only in the clinical scenario [99, 105]. The problem of text classification can be defined as follows. Given a set of documents $D = d_1, d_2, d_3, ..., d_N$ and a set of fixed classes, also called labels, $C = c_1, c_2, c_3$, the aim of text classification is to assign each document the corresponding class $< d, c >$ [118].

The main problem when dealing with unstructured data is that machine learning models can not directly use them, as they need vectors of numerical features as input values. For this reason it was necessary to investigate text embeddings. In this thesis two main categories are discussed and used, the first one is a count-based method while the second one is a prediction-based method.



```
text = ['To be or not to be that is the question',
        'O Romeo, Romeo, wherefore art thou Romeo? Deny thy father and refuse thy name.']
```

	and	art	be	deny	father	is	name	not	or	question	refuse	romeo	that	the	thou	thy	to	wherefore
Sentence1	0	0	2	0	0	1	0	1	1	1	0	0	1	1	0	0	2	0
Sentence2	1	1	0	1	1	0	1	0	0	0	1	3	0	0	1	2	0	1

((a)) Text processed with CountVectorizer.

```
text = ['To be or not to be that is the question',
        'O Romeo, Romeo, wherefore art thou Romeo? Deny thy father and refuse thy name.']
```

	and	art	be	deny	father	is	name	not	or	question	refuse	romeo	that	the
Sentence1	0.000000	0.000000	0.534522	0.000000	0.000000	0.267261	0.000000	0.267261	0.267261	0.267261	0.000000	0.000000	0.267261	0.267261
Sentence2	0.218218	0.218218	0.000000	0.218218	0.218218	0.000000	0.218218	0.000000	0.000000	0.000000	0.218218	0.654654	0.000000	0.000000

((b)) Text processed with TfidfVectorizer.

Figure 3.1 **Different embedding outcome using CountVectorizer and TfidfVectorizer.** Sentences included in the examples are from two renowned William Shakespeare's plays [170, 171].

3.4.1 Bag of Words

A standard sparse representation of text which describes the occurrence of patterns within a document [76]. It involves two things: a vocabulary of tokens and a measure of their presence in the text. Specifically, patterns can be single tokens or n-grams of tokens, *i.e.*, a sequence of neighbouring characters and/or words within a document [41]. Any information about the order or structure of words in the document is discarded. So, the use of n-grams of words allows the Bag of Words model to capture a little of context from the document [93]. The shape of a numerical representation obtained with this technique is: number of samples x number of n-grams of tokens (usually a selection of the most frequent ones). I used two vectorizers from the Python library 'Scikit-learn' [140] named 'CountVectorizer' [4] and 'TfidfVectorizer' [5]. An example of outcome the above mentioned embedding methods is shown in figure 3.1.

Subfigure 3.1(a) shows that CountVectorizer is a mere measure of the frequency of the patterns in a text, e.g., 'be' has frequency = 2 in Sentence1, while 'romeo' has frequency = 3 in Sentence2. On the contrary, subfigure 3.1(b) shows that TfidfVectorizer is weighted measure, and it is based on the *Term Frequency-Inverse Document Frequency* (TF-IDF)

[4] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

formula:

$$(tf-idf)_{i,j} = \frac{n_{i,j}}{\mid d_j \mid} * log_{10} \frac{\mid D \mid}{\mid \{d:\, i\, \varepsilon\, d\} \mid} \tag{3.2}$$

The main difference compared to the previous measure is that TF-IDF is a numerical statistic aimed at reflecting word importance to a document within a collection or a corpus. Specifically, it increases with a direct proportion to the frequency within the document but it is offset by the frequency across all documents. This measure is important because it adjust the fact that, in general, some words can be more frequent than others in speaking or in that specific topic.

It is important to highlight that first the vocabulary could be very large, as it may include every word of the document (it depends on parameters set-up) and all the other n-grams defined by the user. Second the resulting matrix could contain a lot of uninformative elements, some patterns could appear only few times in the documents. For this reason it is important to evaluate possible vectors lengths, which in turn is the number of best selected features, and investigate how this affects the outcome [88, 22, 197, 143].

3.4.2 Word Embedding with Word2Vec

It is an approach to provide a dense vector representation of words that capture something about their meaning. Word embeddings work by using an algorithm to train a set of fixed-length dense and continuous-valued vectors based on a large corpus of text. Each word is represented by a point in the embedding space and these points are learned and moved around based on the words that surround the target word. The output is a model in which each word is mapped in an array of fixed size. As I could not find any pretrained Word2Vec model on Italian texts I decided to train my own model using a part of the clinical text contained in the available sample. To do that, the main steps are:

- Collect a big corpus
- Use a sliding window to go over the text, considering one word at a time as the central word and the other words in the window as context words
- For each central word, compute the probabilities for each context words

- Adjust vectors in order to increase probabilities and minimize the loss function, that is an averaged Negative Log-Likelihood:

$$Loss = J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m \leq j \leq m,\\ j \neq 0}} logP(w_{t+j}|w_t,\theta) \tag{3.3}$$

The first summation allows to iterate over all the words in the text while the second one iterates over the words in the window (context words $-m \leq j \leq m$) excluding the central word ($j \neq 0$). θ represents the variables to optimize.

Each probability $P(w_{t+j}|w_t)$ is computed as follows:

$$P(o|c) = \frac{exp(u_o^T v_c)}{\sum_{w \in V} exp(u_w^T v_c)} \tag{3.4}$$

For each word w in the text, two vectors will be involved: v_c when w is a central word and u_w when it is a context word. So, for each couple of central word c and context word o the probability is computed as the measure of similarity between o and c normalized over the entire vocabulary V. In order to speed up the training process, it is possible to use the Negative Sampling technique which allows to select only a subset of K randomly words instead of all the words in the vocabulary V.

One of the main differences between the two considered approaches is that the Bag of Words model directly returns a representation of the whole document, while the Word Embedding model works at word level. Therefore, when using Word Embedding model, I performed a preliminary analysis of the quality of the words' representation before building the whole document representation. As suggested in [169, 189], I used the following intrinsic evaluators:

Words similarity also called *cosine similarity*, which is defined as:

$$\cos(w1,w2) = \frac{w1 * w2}{|| \ w1 \ || * || \ w2 \ ||} \tag{3.5}$$

where $w1$ and $w2$ are the two word vectors and $|| \ w1 \ ||$ and $|| \ w2 \ ||$ are L_2 norms.

Words analogy given a pair of words (a and a^*) and a third word (b), the analogy relationship between a and a^* can be used to find the word b^* that corresponds to b.

$$a\colon a^* = b\colon b^* \tag{3.6}$$

Outliers detection given a group of words, the objective is to find the one that does not match the context. This is used to evaluate the semantic coherence in words' clusters.

Chapter 4

Data Analysis

4.1 Data Normalization

Considering that the numerical values associated with each feature may belong to different scales, the second mandatory step in data pre-processing is normalization.

The chosen algorithm for structured data normalization is based on the Z-score:

$$Z = \frac{x - \mu}{\sigma} \tag{4.1}$$

where μ is the mean and σ is the standard deviation. The Z-score gives an idea of how far from the mean a data point is. I chose this method for the structured database because it derives from real world data so it is necessary to deal with outliers. Z-score is a perfect strategy because it measures how many standard deviations are below or above the population mean and raw score [26].

While considering unstructured data, I only scaled each feature by its maximum absolute value.

4.2 Supervised learning problem

Supervised Learning is a branch of machine learning that deals with labelled examples. Specifically, the aim of supervised learning is to learn a model that represents the available data and use it to predict an outcome from unseen data. Depending on the problem that needs to be solved, three main methods can be used: classification, regression and forecast. Data Classification deals with categorical labels, while regression takes into consideration also the order of the input measures and so it deals with a concept of distance. Finally, forecast involves also the concept of time.

4.2.1 Regression

Considering a dataset composed by p-points $(x_1, y_1), ..., (x_p, y_p)$, there are many ways of approximating the regression function $f(x)$ as combination of the input, in general it is defined as a function of degree p:

$$f(\underline{x}) = w_0 + w_1 x^1 + w_2 x^2 + ... + w_p x^p = \sum_{i=0}^{p} w_i x^i \tag{4.2}$$

The vector of coefficients $\underline{\mathrm{w}}$ (of dimension p) needs to be learned from the training data and the optimal combination is the one that minimize the error which is defined as the difference between the true function and the estimated one. A well-known method to define the loss function to penalize the errors in the prediction is "Least squares":

$$LossFunction = \hat{L} = \sum_{i=0}^{p} (y_i - f(x))^2 \tag{4.3}$$

The simplest but effective approximation of the regression function is a line, which is called "linear model":

$$f(\underline{x}) = w_0 + w_1 x_1 + w_2 x_2 + ... + w_p x_p = \sum_{i=0}^{p} w_i x_i \tag{4.4}$$

4.2.2 Regularization

One of the main issues in supervised learning is overfitting. It happens when a model explains too precisely the known data, fitting the noise as well as the signal, and therefore losing the ability to generalize on unknown data. This could be due to a low number of training examples or to a too high complexity of the model. To address this issue a penalty term called "regularizer" is introduced.

A well-know model that belongs to the category of supervised learning is the *Kernel Recursive Least Squares* (KRLS). Its structure can be described as a linear function in a space induced by $\underline{\phi}(x)$:

$$f(\underline{x}) = \underline{w}^T \cdot \underline{\phi}(x) \tag{4.5}$$

Where $\underline{x} \in R^d$ and y may belong to: R (*Regression*), $\{\pm 1\}$ (*Binary Classification*) or $\{1...c\}$ (*Multiclass Classification*).

$\underline{\phi}(x)$ is defined by a kernel, the one with good properties is the Gaussian kernel:

$$K(\underline{u}, \underline{v}) = e^{-\gamma ||\underline{u} - \underline{v}||^2} \tag{4.6}$$

And the optimal value of $\underline{w}$ is defined as:

$$\underline{w} : {}^{ARGMIN}_{\underline{w}} ||X\underline{w} - \underline{y}||^2 + \lambda ||\underline{w}||^2 \tag{4.7}$$

where the first term is the Loss function ($\hat{L}$), which measures the cost of the model, and the second one is the Regularizer (λReg), which governs the complexity of the model. The sum of the two is an approximation of the True Error:

$$L \approx \hat{L} + \lambda Reg \tag{4.8}$$

λ is used to control the compromise between the two terms, specifically if λ is high the model tends to underfit while if it is low it tends to overfit.

In order to find the best values of the hyper-parameters γ and λ it is necessary to perform Model Selection and Error Estimation by splitting the data into three chunks: training, validation and test.

From KRLS structure it is possible to derive two of the ML models that I used by changing one term in the expression: *Least Absolute Shrinkage and Selection Operator* (LASSO) and *Support Vector Machines* (SVM).

Least Absolute Shrinkage and Selection Operator (LASSO)

The Least Absolute Shrinkage and Selection Operator, whose acronym LASSO was introduced by [180], is a penalised least square regression model using L_1 penalisation function [58, 124]:

$$\lambda ||\underline{w}||_2^2 \rightarrow \lambda ||\underline{w}||_1 \tag{4.9}$$

This type of regularization makes LASSO suitable to perform features selection, as its coefficients' matrix is sparse. This means that a decimal positive/negative weight in the matrix indicates the importance of the specific feature for the outcome while 0 is assigned to features that are not involved in the prediction. Even though LASSO is a regression model, so its output is a decimal number, it can be used also in classification task, by imposing a threshold.

Support Vector Machines (SVM)

A supervised learning method based on maximum margin linear discriminants [186, 37], that can be cast as an optimization problem in the regularization framework [54]. The goal is to find a function that, according to a given value of the regularization parameter, minimizes a functional composed of a fitness term measured according to the Hinge Loss [188] and a regularization term that controls the smoothness of the function f according to a kernel function K. In binary classification, to compute the maximum margin search, SVM uses the same concept or KRLS but the involved loss function is the Hinge Loss [188]:

$$\hat{L} = HingeLoss = MAX[0, 1 - yf] \tag{4.10}$$

I evaluated SVM model performances on the specific tasks considering three possible kernels: linear, Gaussian and polynomial.

Sparse Logistic Regression (SLR)

It is a classification technique that directly models the probability to belong to a class as a discriminative function based on explanatory variables [199, 96]. Considering the class variable as binary, the outcome variable is Bernoulli distributed:

$$f(y_i, \pi_i) = \begin{cases} \pi_i^{y_i}(1-\pi_i)^{1-y_i} & for \quad y_i = 0, 1 \\ 0 & otherwise \end{cases} \tag{4.11}$$

where π_i is the probability of the positive class. The aim is to define a π_i that depends on a vector of covariates. It could not be directly a linear function, but a monotone transformation allows to do that, it is called "logit":

$$log(\frac{\pi}{1-\pi}) = w_0 + w^T x = w_0 + w_1 x_1 + w_2 x_2 + ... + w_p x_p = f(x) \tag{4.12}$$

where p is the number of predictor variables and w_i the parameters to estimate. Therefore, the probability π_i is defined as a sigmoid:

$$\pi = \frac{1}{1+e^{-f(x)}} \tag{4.13}$$

The optimal set of parameters w^T is found by minimizing the negative log likelihood:

$$\hat{L} = \sum y_i log(\pi_i) + (1-y_i) log(1-\pi_i) \tag{4.14}$$

In order to prevent overfitting, the concept of penalized logistic regression has been introduced by adding a regularization penalty term [202]. The regularized can be:

- L_1 norm, which is the same introduced in LASSO, is used to enforce sparsity which allow to build a predictive model based only on a subset of meaningful input variables.
- L_2 norm, which is the same used in KRLS, is used to assign lower weights to features that contribute less but differently from LASSO it includes all variables.
- *Elastic Net*, it is a combination of the above mentioned two norms L_1 and L_2. In this case some coefficients are shrinked towards zero and other are exactly zero.

4.2.3 Ensamble methods

Ensamble methods are a subclass of supervised learning algorithms that build a set of classifiers and the final prediction is the weighted average of the predictions of each classifier [47].

Random Forest (RF)

Random Forest is based on the technique of bootstrap aggregation which aims at reducing the variance of the prediction function by averaging many noisy models but approximately unbiased [77]. Contrary to the other classifiers described above, RF is an ensemble method. It is composed by K decision tree classifiers created from a different bootstrap sample. The trees are built by sampling a random subset of the attributes at each internal node in the decision tree. The random sampling of the attributes reduces the correlation between the trees in the ensemble [106]. The main advantage in using trees is that they can detect complex interactions among features and the reaches a low level of bias if they are sufficiently deep.

4.3 Performance evaluation metrics

Another important aspect linked to machine learning is model performances evaluation. In order to measure how accurately the model can predict the outcome, in a classification task, multiple evaluation metrics are available [77, 28, 21, 81]. Note that it is crucial to chose the most appropriate ones based on the specific case study and usually it is better to evaluate the performance considering more than one of them.

A graphical representation of the performances can be obtained through the Confusion matrix, which in the case of binary classification has two dimensions.

True values

		0	1	
Predicted values	0	TN	FN	$TN+FP$
	1	FP	TP	$FP+TP$
		$TN+FP$	$FN+TP$	N

Accuracy

Accuracy is the simplest evaluation metrics and it is measured as the number of correct predictions on the total number predictions.

$$Accuracy = \frac{TN+TP}{N} \tag{4.15}$$

Accuracy is an effective indication of model performance when the two class are balanced.

Precision

It measures the number of true positives over the total number of positives predicted. It is useful in the case of unbalanced datasets and it decreases with the increase of false positives.

$$Precision = PositivePredictiveValue = \frac{TP}{TP+FP} \tag{4.16}$$

Recall

It measures the number of true positive over the total number of positives in the dataset. It gives a feedback on model ability to detect the positive samples and it decreases as the number of false negative increases.

$$Recall = Sensitivity = \frac{TP}{TP + FN} \tag{4.17}$$

F1-score

It is the harmonic mean of Precision and Recall. It is a complete measure of model performance because it reaches high values only if both involved metrics are high. This metric is especially used in the clinical scenario where class are usually unbalanced and it is important to detect the highest number possible of positive cases, which could be for example patients with a specific disease.

$$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4.18}$$

4.4 Statistical Analysis

Finally, it is possible to investigate statistically significant differences between performances obtained with different combinations of classifiers and numerical representations of text.

4.4.1 Hypothesis Testing

First of all, it is necessary to formulate a hypothesis, which should be testable through an experiment or observation. Then, it is necessary to test the results in order to evaluate the possibility that they happened by chance, and therefore returning an indication on repeatability of the experiment. A statistical test is a procedure which allows to decide, with a certain degree of certainty (p-value = 1 - α), if the null hypothesis can be rejected and, as a consequence, accept the alternative hypothesis [49]. Common threshold levels for α are: 0.05, 0.01, 0.001. Each test has its validity conditions and it is more or less robust depending on the violation of such conditions. It is important to identify the most appropriate test, *i.e.,* the most robust, based on the scale and characteristics of the data.

2 Groups

Kolmogorov-Smirnov Test It aims at assessing the proximity of two distributions, *e.g.,* in terms of mode and symmetry. They can be one theoretical and one sample, e.g., to determine if a sample is normally distributed, or two samples. It is suitable both for continuous and discrete data [119].

Wilcoxon Test It is the non-parametric equivalent of Student's t-test, therefore it is used on paired data but when they are not Gaussianly distributed. By comparing the means of two samples, it aims at determining if they belong to the same population [194].

Mann-Whitney U Test It differs from the previous one because it is used when data are considered independent, *i.e.,* two samples are randomly selected from two populations and they may have different size [117].

At least 3 groups

Kruskal-Wallis Test It is a non-parametric test, which means that there are no assumptions about data distribution, the only requirement is that data should be ordinal scale. The objective is to determine if the medians of three or more independent groups are statistically different. In this case the null hypothesis (H0) is that all groups have the same median while the alternative hypothesis (Ha) is that the median is not the same [101].

Friedman Test As the previous one, it is a non-parametric test, used to compare at least three "paired" groups. It is used on repeated measures in order to investigate the effect of a specific factor [61].

4.4.2 Corrected Significance Level

The main assumption is that by testing an increasing number of hypotheses (k) then the probability of obtaining at least one significant result increases [18]. Therefore, correction methods are used in order to address type I errors, so to reduce the probability of obtaining false-positive results.

Bonferroni correction

It is the simplest type of correction, the values of α corrected can be obtained by dividing α for the number of tested hypothesis (k) [46].

$$\alpha^* \approx \frac{\alpha}{k} \tag{4.19}$$

However, it is used when k is relatively small, otherwise it can be too conservative, and other methods are more suitable [51].

Dunn-Sidàk correction

It is another method to address the problem of multiple comparisons less stringent than the previous one [173].

$$\alpha^{*} = 1 - (1-\alpha)^{\frac{1}{k}} \tag{4.20}$$

Part III

Results

Chapter 5

Localization of the epileptogenic zone in patients with drug-resistant focal epilepsy

5.1 Research Question

Epilepsy is one of the most common neurological disorders. According to the *World Health Organization* (WHO) [137], epilepsy affects almost 50 million individuals worldwide and up to 10% of people have one seizure during their lifetime. About one third of people with epilepsy continue to be treatment resistant despite reasonable trials of at least two antiepileptic medications. People with focal onset drug-resistant seizures (DRS) may be cured or significantly improved with epilepsy surgery. The aim of surgery is to remove or disconnect the EZ, from which the seizures originate [164]. A delayed or incorrect diagnosis of EZ location limits the efficacy of surgery, impairing the final outcome and reducing the quality of life of people with drug-resistant epilepsy (DRE). Notwithstanding, not enough people receive epilepsy surgery [176, 166, 23]. Epilepsy surgery continues to be underutilized for several reasons, including patient misunderstandings regarding surgery and economic gaps that restrict access to care. Moreover, physicians consistently lack the information necessary to recognize candidates for surgery. This could be partially attributable to the difficulty of presurgical assessment, which entails classifying the kind of seizure, localizing and lateralizing the EZ, and determining the safety of the desired surgical procedure in light of any potential defects (motor, cognitive, etc.) [176]. Long-term Video-electroencephalography (VEEG) monitoring is a diagnostic technique commonly used to objectively capture both clinical manifestations and brain activity during seizures [98]. It is crucial to properly interpret seizure-related subjective and objective manifestations in order to develop a solid hypothesis on the potential location of the EZ. Usually epileptologists review multiple seizure manifestations obtained from long-term VEEG recordings and write a detailed report about the characteristics of the semiological manifestations (e.g., motor/non-motor) and their chronological appearance [179]. A hypothesis regarding the location of the EZ is created as a result, and it may direct the planning of a surgical intervention (when supported by EEG, MRI, and/or functional data), or it may guide additional presurgical evaluation phases (including invasive procedures). However, this process is very complex and relies on special skills that require many years of experience to form [63]. With the advent of advanced ML algorithms, modern decision support systems reach high accuracy in the interpretation of clinical data [112, 50, 24, 4, 136] as well as provide support for the formulation of optimal therapeutic options [14].

Hence, ML-based systems able to automatically analyze clinical reports of seizure manifestation could represent an important tool to support clinical diagnosis of people with DRS. However, the collected reports of seizure descriptions are usually text-based unstructured data that cannot be trivially analyzed with ML alone. NLP is the branch of Artificial Intelligence that deals with the analysis of natural human language performed by computers. The applications of NLP in the clinical field are multiple [190, 112, 50] and, together with ML methods, they can be involved in the diagnostic process.

The increasing interest in the combined use of ML and NLP techniques in the clinical neuroscience field led to several research projects focused on its use to support differential diagnosis and management in epilepsy syndromes [203, 57, 144]. Among them, there are few examples of applications in the specific task of predict the localization of the EZ and they mainly rely on expert clinicians that identify meaningful keywords later extracted manually or using regular expressions [95, 5]. However, these attempts are based on pre-defined rules that require additional work to be adopted at larger scale and biased because they rely only on clinicians' experience. So, the automatic analysis of semiology description for EZ location is still an open question [3].

5.2 My contribution

As time plays a key role in the EZ localization process in DRE patients, the aim of the work is building a cost-effective support system based on text, that is the simplest and always available type of information. Therefore I aim at transforming the EMR from a mere collection of clinical information into a powerful instrument for clinical investigation. The developed system takes as input the raw text containing the description of the seizures and it predicts the probability of the EZ to be in the temporal/extra-temporal lobes and left/right hemispheres. One of the main challenges that I faced during the development of this pipeline is that not only most of the NLP-based tools are specifically built for the English language but also no Italian ontology on epilepsy is available. To the best of my knowledge, this is the first work that uses the semiological descriptions of seizures written in Italian language to identify the EZ.

5.3 Sample Characteristics

A large number of subjects with an epilepsy diagnosis was recruited from the "Claudio Munari" Epilepsy Surgery Centre, Niguarda Hospital in Milan (Italy). Among them, the group of focal DRE patients who resulted seizure-free after a surgical intervention (minimum follow-up of two years) was selected so that the origin of seizures (epileptogenic zone) could be exactly identified. This resulted in a cohort of 129 patients, whose age and sex are distributed according to the pyramid chart in Fig. 5.1 and localization and hemisphere of EZ distribution are summarized in Table 5.1.

Table 5.1 *Number of patients in the sample for each combination of Localization and Side.*

Localization	**Side**	**Patients (*n*)**
Frontal	Right	14
	Left	17
Temporal	Right	33
	Left	30
Central	Right	2
	Left	1
Insulo-opercular	Right	6
	Left	4
Posterior	Right	10
	Left	10
Hemispheric	Right	1
	Left	1

All participants gave informed consent for data collection and usage for scientific research (ID 939-12.12.2013). This is an anonymous and retrospective study that complies with the principles outlined in the Declaration of Helsinki [9].

For each patient, we collected the EZ localization/lateralization label and two groups of textual data written in Italian language consisting of the description of all available seizures and an excerpt of the patients' EMR. In particular:

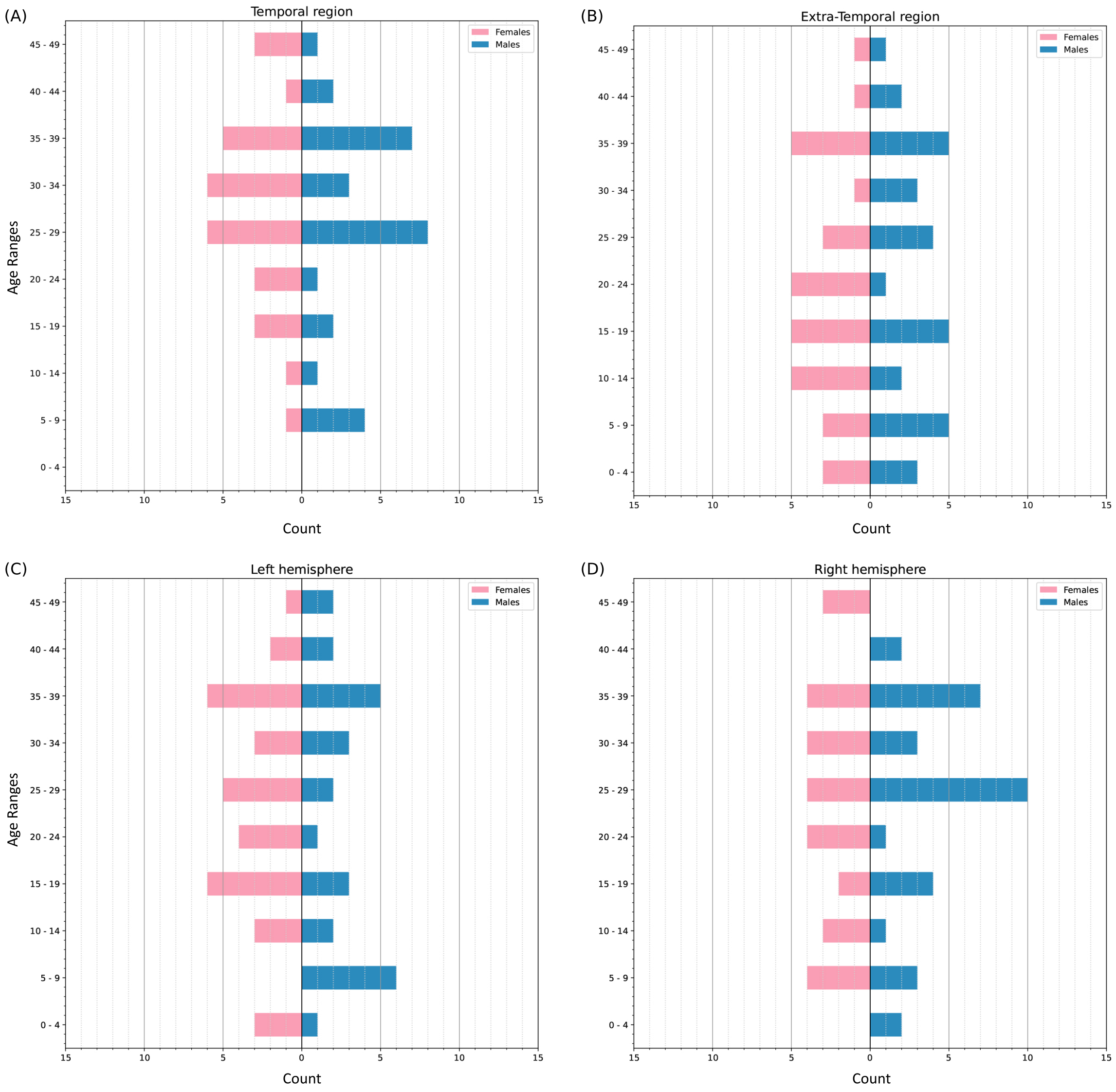


Figure 5.1 **Population distribution.** All patients included in the analysis are under 50 years old. In particular, for both tasks temporal/extra-temporal and right/left the majority of the population have an age between 25 and 39.

Seizure descriptions are texts describing the semiology of seizures. Specifically, medical experts revised the recorded videos showing the patient undergoing seizure events and they described the patient's seizure manifestations and evolution. We revised all the seizure descriptions (N=562) and we discarded the ones that: i) were labelled as "subjective manifestations"; ii) referred to previous seizures (e.g., the sentence reads as *"Seizure similar to the previous ones including the automatisms of the right hand brought to the face"*); iii) were composed of less than 20 words; iv) had no mention of physical movements or sensations (*i.e.*, ictal EEG interpretations). In total, we excluded 53 texts resulting in a final dataset of 509 descriptions from 121 patients out of the 129 included. The average number of seizures per patient is 4.21 ± 3.51, range: 1 - 17. We considered single seizure description as independent events because ictal events occurred at different time, the observer who dealt with the patients was not necessarily the same as well as the clinician who wrote the semiological descriptions.

Electronic Medical Records (EMRs) are the de-identified extractions from EMRs from 129 patients. These texts contained *anamnestic information* such as patient's history, previous treatments, drug-dosage, etc.

In order to preserve the morphological structure of the sentences, as well as to de-identify all texts before pre-processing, Protected Health Information (PHI) were manually removed. First "direct identifiers", i.e. those words (or group of words) that could directly identify the patient, such as first and last names, were removed and substituted them with "ragazza/ragazzo" *("girl"/"boy")* by matching gender. Then other "quasi-identifiers", i.e. those entities that, if combined together, may detect a small group of people, such as locations and exact birth-date, were removed and substituted with generic "città"*("city")* the locations. Finally, day and year from dates were removed. I stored the de-identified texts in a SQL Server database located in a server that could be only reached through Virtual Private Network. I used a unique identifier for each patient and retained only minimal personal information, such as sex and year of birth, as prescribed by the international and national regulations on data protection [153, 152].

The goal of the study is to build predictive models based only on seizure descriptions represented according to some embedding criterion. The problem is cast into a supervised learning setting where each seizure is associated to a label. In my cohort, expert epileptologists identified two types of labels: the region and the side of the brain where the epileptogenic zone is located. This information is available as all patients underwent a surgical intervention that resolved the pathological condition. The first type defines whether the seizure onsets from the temporal ($n_{temporal}$=58) or extra-temporal ($n_{extra-temporal}$=63) brain region, where the extra-temporal label also includes patients whose EZ do not cover only the temporal region. The second label type differentiates the EZ based on which hemisphere they are located in (right (n_{right}=61) or left (n_{left}=60)). Considering that, as described above, each patient may suffer from more than one seizure, the dataset is composed of: 39% of seizures labeled in the temporal region and 56% of seizures associated to the right hemisphere.

5.4 Experimental Design

Fig. 5.2 shows the sequence of steps I performed. First, I had to pre-process the collected textual data to anonymize, standardize and filter it *(Sec. 3.2)*. Then, I extracted three possible numerical representations of the seizures descriptions based on Bag of Words *(Sec.3.4.1)* and Word Embedding methods *(Sec.3.4.2)*. Finally, I applied supervised machine learning models, using the numerical representations as inputs to predict in a first step the brain region and then the lateralization of the EZ *(Sec. 4.2)*.

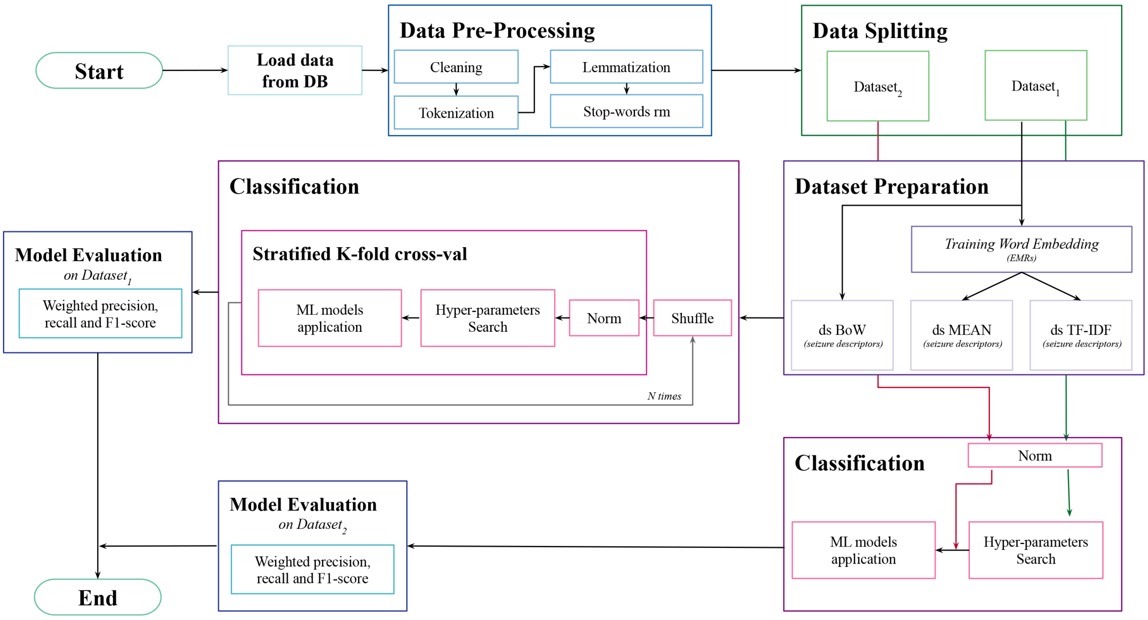

Start
Load data from DB
Data Pre-Processing
Cleaning
Tokenization
Lemmatization
Stop-words rm
Data Splitting
Dataset$_2$
Dataset$_1$
Classification
Stratified K-fold cross-val
ML models application
Hyper-parameters Search
Norm
Shuffle
N times
Dataset Preparation
Training Word Embedding (EMRs)
ds BoW (seizure descriptors)
ds MEAN (seizure descriptors)
ds TF-IDF (seizure descriptors)
Model Evaluation on Dataset$_1$
Weighted precision, recall and F1-score
Classification
Norm
ML models application
Hyper-parameters Search
Model Evaluation on Dataset$_2$
Weighted precision, recall and F1-score
End

Figure 5.2 **Pipeline Schema.** The pipeline can be divided into 5 main sections: data loading, data pre-processing, dataset preparation, classification and model evaluation.

Dataset

After the preprocessing phase, the dataset was split into two chunks: $Dataset_1$ consisting of 445 seizures from 106 patients and $Dataset_2$ consisting of 64 seizures from 15 patients. The splitting is required to properly assess the generalization properties of the representation/predictive models pairs, as described below.

NLP

Considering the Bag of words technique, to identify the best size for the numerical representation, tested three choices for the total number of features: 100, 200, and 300. I obtained best results with 200 features in the first task and with 300 features in the second task. From now on I will refer to this numerical representation with *bw*.

Considering the Word Embedding representation, I tested different vector dimensions and combinations of some word embedding parameters for the vector representation, according to reference ranges described in [29]. In particular I ranged: vector dimension between 10, 30, 100; negative sampling between 5, 10, 15; epochs between 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000. I determined the best combination performing the intrinsic evaluation: vector dimension = 100; negative sampling = 10 and number of epochs = 300. Then I fixed: the minimum words' occurrence in the text = 2, in order to exclude too rare words or misspellings of frequent words, and number of context words = 3.

Once I obtained the Word Embedding model built with *Word2vec*, I used two approaches to build the representation of seizure descriptions:

In the first one, all the arrays of the words have been averaged and the first matrix was obtained. From now on, I will refer to it with ***mean representation***.

$$mean = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.1}$$

In the second approach, the TF-IDF formula 3.2 has been applied and the second matrix obtained. From now on, I will refer to it with ***tf-idf representation***.

In conclusion, I obtained a total of 3 matrices per dataset: the first one using the Bag of Words *(bw representation)* and the other two using the Word Embedding model *(mean and tf-idf representations)*.

ML methods

For each data representation, I solved a binary classification problem for two different tasks, that are the prediction of the brain region and the prediction of the brain hemisphere where seizures originate. Specifically, I adopted and compared two different machine learning classification methods, that are SLR with L_1 penalty and SVM with three different kernels: *linear*, *rbf* (gaussian), and *poly* (polynomial with degree equal to 3). Hence, in total I obtained 4 models per each input and task.

For all experiments, I performed a stratified k-fold cross-validation, with k = 10, to iteratively split the $Dataset_1$ into ten different training and validation sets. At each split, the following steps were performed:

1. Data normalization

2. A 10-fold cross-validation to identify the best hyperparameters

3. Model training on the training set with the best hyperparameters

4. Model evaluation on the validation set

All the aforementioned steps have been executed $N = 3$ times, each time shuffling the data. In order to guarantee the reproducibility of the results I set the random state used to shuffle data equal to the iteration index (i.e. in order 0, 1, 2). Then I performed an overall evaluation of the model over multiple trials calculating the median performance per trial and the mean performance across the N trials.

It is important to specify that I pre-processed seizures of patients belonging to Dataset$_2$ together with Dataset$_1$ used for cross-validation but I did not use samples from the Dataset$_2$ for training the word embedding and the bag of words models. So, when testing generalization ability I looked for the best model hyperparameters using the whole Dataset$_1$, while during cross-validation I looked for model hyperparameters only on the training set.

All experiments are evaluated in terms of the following weighted metrics on each fold and on average: precision, recall (which is equal to accuracy), and F1-score. *Sections 5.6 and 5.7* describe and graphically represent results in terms of F1-score while the graphs related to the other two metrics are available in the supplementary material.

5.5 Word Embedding model evaluation

Natural language represents a high-dimensional space whose mathematical representation can vary significantly depending on the adopted embedding. Therefore, before building the numerical representations of seizures descriptions based on Word2Vec model, I tested the performances of the devised model with three different evaluators. First, I tested if the Word Embedding correctly recognized semantic and syntactic meaning of random words. I chose five target words and extracted the most similar words from the Word Embedding according to the Word Similarity measure. I graphically visualized the five target words and the five nearest ones using *T-distributed Stochastic Neighbor Embedding* (T-SNE), which allows the visualization of high dimensional data by locating the datapoints in a two/three-dimensional map [184]. Our model correctly associates words with syntactic and semantic meanings similar to the target words in all selected cases *(Fig. 5.3, Table 5.2)*. In particular, as we can see in *Fig. 5.3*, the word “sollevamento *(lift)*" is one of the most similar words to “movimento *(movement)*" but it is also a very similar to “elevazione *(elevation)*" and this property is recognized by the model, as the two words are close together in the two-dimensional projection made by T-SNE. Then, I tested our model with the Words Analogy evaluator, so I wanted to find the word that satisfied the following relation: “braccio *(arm)*" + “gamba *(leg)*" - “piede *(foot)*". I obtained the expected word “mano *(hand)*". Finally, I evaluated the ability of outliers detection, so I tested if the Word Embedding model could recognize words out of their context. Specifically, I chose a quadruplet of random words: three within the same context and one outlier. I repeated this experiment three times, finding that the model is always able to detect the outlier *(Table 5.3)*. Results suggest that the model properly identifies relations between words.

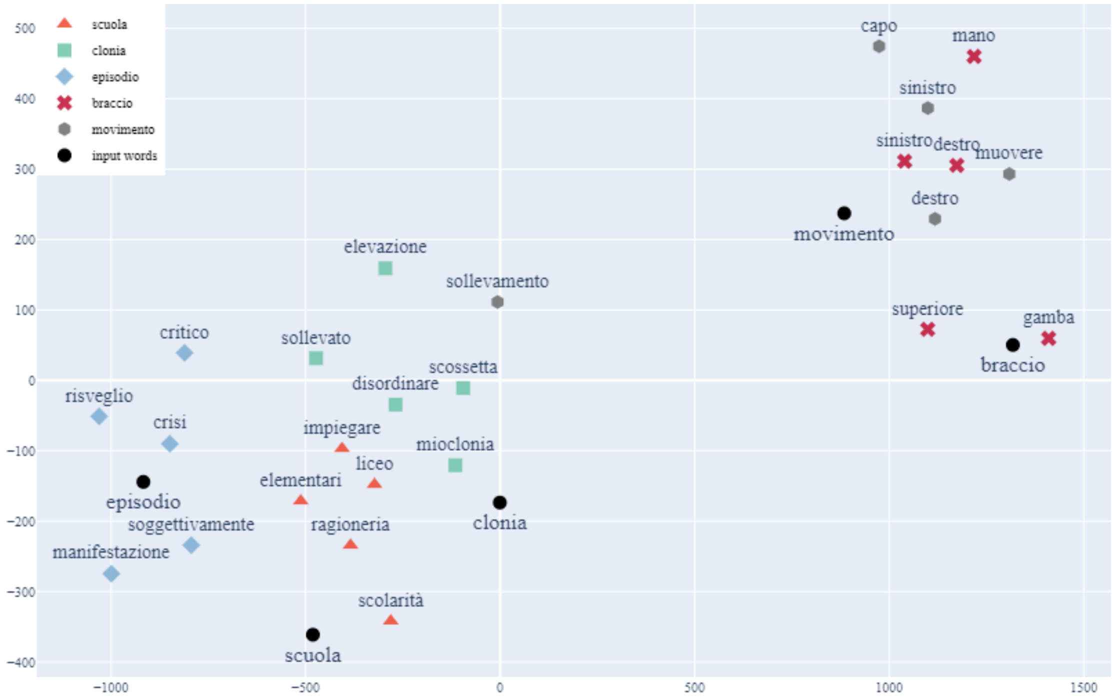


Figure 5.3 **Examples of clusters of Word Embedding, visualized in a two-dimensional space using T-SNE.** Each cluster is composed of one input word and the top 5 most similar words. Two main clusters can be distinguished in the figure, words in the upper-right part of the figure are linked to the seizure descriptions (“braccio *(arm)*" and “movimento *(movement)*") and words in the lower-left part of the figure linked to anamnestic information (“episodio *(episode)*", “scuola *(school)*" and “clonia *(jerk)*", which is mentioned most of the times in that section).

Table 5.2 *Examples of the words similarity evaluation. The first column contains the five target words arbitrarily selected and the second column contains the corresponding five most similar words induced by the Word Embedding model. Most similar words are written in descending order considering the similarity coefficient based on the cosine similarity.*

Target word	**Most similar words**
Braccio	Gamba, mano, sinistro, destro, superiore
(arm)	*(leg, hand, left, right, upper)*
Episodio	Crisi, soggettivamente, risveglio, critico, manifestazione
(episode)	*(seizure, subjectively, awakening, critical, manifestation)*
Clonia	Mioclonia, scossetta, disordinare, elevazione, sollevato
(jerk)	*(myoclonia, jerk, clutter, elevation, raised)*
Scuola	Ragioneria, liceo, elementari, scolarità, impiegare
(school)	*(accounting, high school, primary school, schooling, employ)*
Movimento	Muovere, sinistro, destro, capo, sollevamento
(movement)	*(move, left, right, head, lift)*

Table 5.3 *Examples of the outliers detection evaluator. Given 4 words the trained model could correctly identify the word that did not match the context.*

Input words	**Outliers**
Dormire, cuscino, letto, bicchiere	Bicchiere
(sleep, pillow, bed, glass)	*(glass)*
Scuola, braccio, gamba, mano	Scuola
(school, arm, leg, hand)	*(school)*
Fratello, sorella, madre, crisi	Crisi
(brother, sister, mother, seizure)	*(seizure)*

5.6 Temporal vs. Extra-temporal seizure onset sites

The first learning task I faced aimed at predicting the temporal or extra-temporal origin for a given seizure.

In *Figure 5.4* I report the weighted F1-scores for all combinations of trained model, data shuffle (red, light green, and light blue) and data representation (*mean*, *tf-idf*, and *bw*).

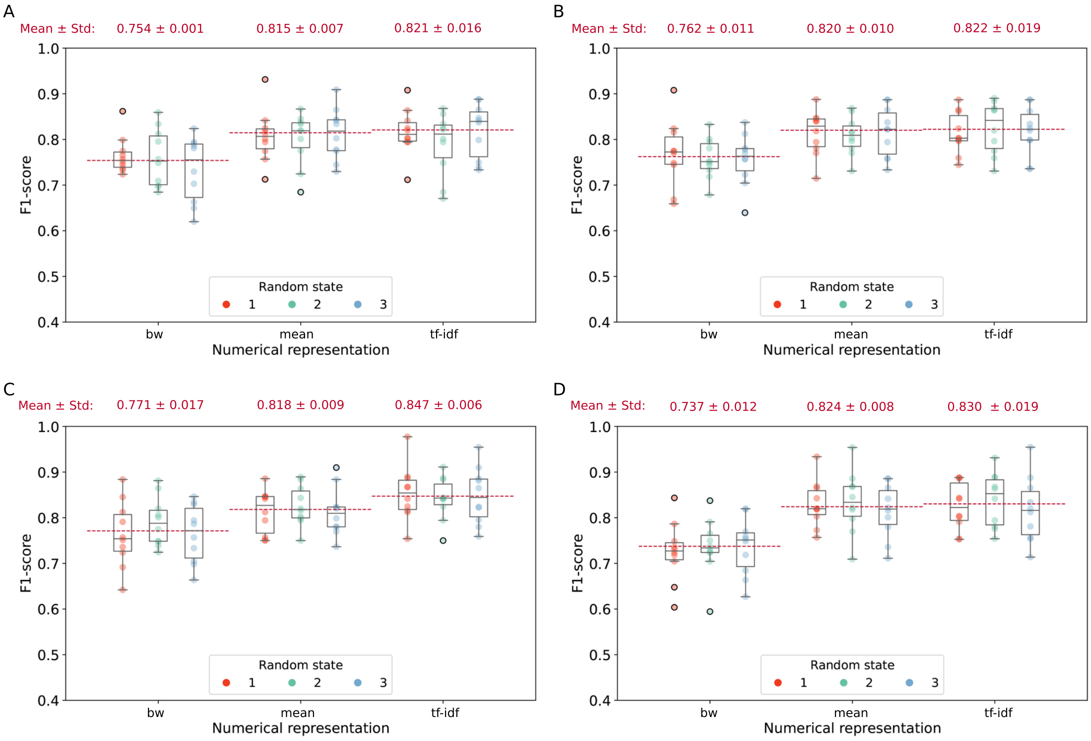


Figure 5.4 **Weighted F1-scores for the localization task** of (A) Sparse Logistic Regression, (B) SVM with linear kernel, (C) SVM with rbf kernel, and (D) SVM with poly kernel over the three fixed random states (red, light green, and light blue) and the three numerical representations (*bw*, *mean*, and *tf-idf*). For each representation and random state, the weighted F1-score values of the K-folds are showed. The red dotted lines identify the mean of second quartiles over the three random states. Numbers at the top of each panel represent $\mu \pm \sigma$ of the second quartiles over the three random states.

Sparse Logistic Regression obtained the best performance using *mean representation* while SVM with *tf-idf representation* considering all kernels. Among all possible combinations, the devised pipeline obtained its best performance (F1-score = 0.847 ± 0.006) in correctly identifying the EZ location using SVM with rbf kernel with *tf-idf* representation.

Therefore, using *Dataset*$_1$, all predictive models showed the best performance (≥ 0.80) when relying on the Word Embedding-based representations respect to those generated by Bag of Words *(Fig. 5.4)*.

This difference was also confirmed as significant by statistical analysis. Specifically, the impact of the different numerical representations of text on the outcome was first evaluated with Kruskal-Wallis Test [101] (null hypothesis: same performance) comparing, for each classifier, the thirty F1 scores obtained from 10-fold cross-validation repeated for three shuffles. Then, as a statistical significant difference was observed, multiple comparisons were performed using Mann-Whitney U Test [117] and the obtained p-values are displayed in Table 5.4. The level of significance $\alpha = 0.05$ was corrected with Bonferroni correction [46]. So, $\alpha = 0.05/k = 0.01667$, with k = 3 (number of comparisons).

Table 5.4 *P-values derived from multiple comparisons performed on Dataset$_1$ to assess the impact of different numerical representations of text on the outcome of localization of the EZ in temporal vs Extra-Temporal region (null hypothesis: same performance). Considering each classifier, the results obtained with the three numerical representations are compared in couples.*

	mean vs tf-idf	**mean vs bw**	**tf-idf vs bw**
SLR	0.70069	**0.00079**	**0.00053**
SVM linear	0.54933	**0.00032**	**0.00029**
SVM rbf	0.03147	**0.00443**	**2.35415e^{-5}**
SVM poly	0.94107	**1.02773e^{-6}**	**7.59915e^{-7}**

An analogous analysis was performed to determine if the choice of the classifier had an impact on the outcome (null hypothesis: same performance). So, first, considering each numerical representation of text, the results obtained by the four classifiers are compared using Friedman Test [61]. As a statistical significant difference was observed, then, for each numerical representation, classifiers performances have been compared using Wilcoxon Test [194] and p-values are displayed in Table 5.5. The values of α corrected are: $\alpha \times (Bonferroni) = 0.00833$ and $\alpha(Dunn - Sidk) = 0.00851$.

Eventually, the choice of the classifier did not show an impact considering *mean* and *bw*, while considering *tf-idf* there is a statistically significant difference in using the SVM with rbf kernel.

To assess level of generalization of both classification and data representation models, I tested the pipeline also on *Dataset*$_2$ and I obtained performances above 0.70. According to the results reported in *Table 5.6*, the F1-score on Dataset$_2$ obtained with SLR is 0.721 with *mean*, 0.743 with *tf-idf*, and 0.728 with *bw*. Considering the SVM classifier I analyzed the

Table 5.5 *P-values derived from multiple comparisons performed on $Dataset_1$ to assess the impact of the choice of the classifier on the outcome of localization of the EZ in temporal vs Extra-Temporal (null hypothesis: same performance). Considering each numerical representation of text, the results obtained with each classifier are compared in couples.*

	SLR	**SLR**	**SLR**	**SVM linear**	**SVM linear**	**SVM rbf**
	vs	vs	vs	vs	vs	vs
	SVM Linear	**SVM rbf**	**SVM poly**	**SVM rbf**	**SVM poly**	**SVM poly**
mean	0.19430	0.31988	0.07570	0.63844	0.09645	0.46883
tf-idf	0.30550	**0.00178**	0.02253	**0.00547**	0.19963	0.02302
bw	0.3235	0.04042	0.05193	0.48221	0.05849	0.01370

F1-score obtained with each kernel. The linear kernel obtained 0.733 with *mean*, 0.749 with *tf-idf*, and 0.731 with *bw*. The rbf kernel showed a value of 0.781 with *mean*, 0.762 with *tf-idf*, and it presented 0.777 with *bw*. The poly kernel showed a performance of 0.701 with *mean*, 0.716 with *tf-idf*, and 0.745 with *bw*. In general almost all combinations of classification models and numerical representations return a F1-score higher than 0.70 but SVM with rbf kernel obtained best performances.

Table 5.6 *Models performances on $Dataset_2$ in the brain region identification. For each combination of numerical representation and classifier the weighted F1-score obtained on the blind test set is reported.*

	mean	**tf-idf**	**BW**
SLR	0.721	0.743	0.728
SVM linear	0.733	0.749	0.731
SVM rbf	0.781	0.762	0.777
SVM poly	0.701	0.716	0.745

5.7 Left vs. Right hemisphere seizure onset sites

The second learning task aimed at defining a predictive model capable of determining the lateralization (*i.e.*, left vs right hemisphere) of the EZ.

As in the previous task, mean performances of all combinations of representations and classifiers are presented in *(Fig. 5.5)*. Differently from the previous task, Word Embedding-based representations yielded to the best performances on only half of the predictive models. Indeed, Sparse Logistic Regression and SVM with linear kernel obtained their best performance using *bw* representation, and only SVM with rbf and poly kernel using *tf-idf*

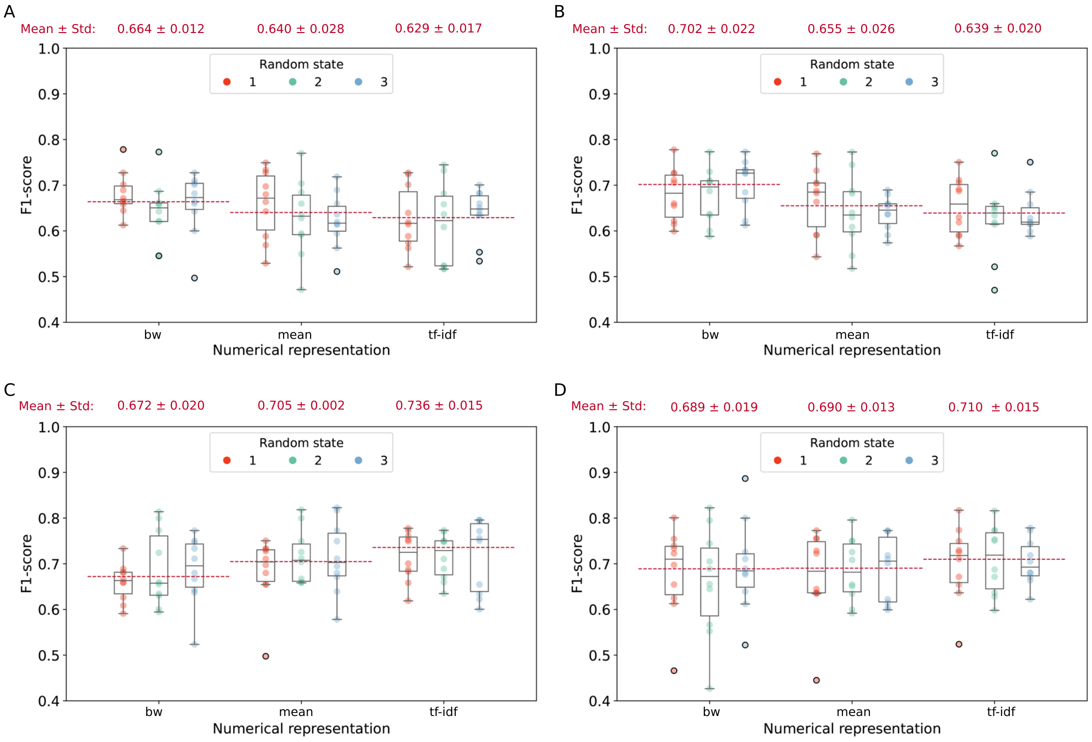


Figure 5.5 **Weighted F1-scores of classification model for the lateralization task** of (A) Sparse Logistic Regression, (B) SVM with linear kernel, (C) SVM with rbf kernel, and (D) SVM with poly kernel over the three fixed random states (red, light green, and light blue) and the three numerical representations (*bw*, *mean*, and *tf-idf*). For each representation and random state, the weighted F1-score values of the K-folds are showed. The red dotted lines identify the mean of second quartiles over the three random states. Numbers at the top of each panel represent $\mu \pm \sigma$ of the second quartiles over the three random states.

representation. Altogether these results confirm that while attempting to predict the lateralization of seizure onset zone the combination of *tf-idf* representation and SVM with rbf kernel yielded the best performance (F1-score = 0.736 ± 0.015).

The same statistical analysis performed in the previous task was executed also in the lateralization task in order to highlight significant differences between all the combinations of classifiers and numerical representations. Differently from the previous task, but in accordance with the expectations, the choice of the numerical representation had no significant impact on the outcome (null hypothesis: same performance). On the contrary, a statistically significant difference was observed considering the two numerical representation built with Word Embedding. Specifically, as shown in Table 5.7 and confirmed by 5.5, the choice of

SVM with rbf or poly kernel instead of SLR and SVM with linear kernel is statistically different.

Table 5.7 *P-values derived from multiple comparisons performed on Dataset*$_1$ *to assess the impact of the choice of the classifier on the outcome of lateralization of the EZ in left vs right hemisphere (null hypothesis: same performance). Considering each numerical representation of text, the results obtained with each classifier are compared in couples. The values of* α *corrected are:* $\alpha^*(Bonferroni) = 0.00833$ *and* $\alpha^{**}(Dunn - Sidk) = 0.00851$.

	SLR	**SLR**	**SLR**	**SVM linear**	**SVM linear**	**SVM rbf**
	vs	vs	vs	vs	vs	vs
	SVM Linear	**SVM rbf**	**SVM poly**	**SVM rbf**	**SVM poly**	**SVM poly**
mean	0.17421	**5.43239e^{-5}**	0.00927	**0.00033**	0.01245	0.04042
tf-idf	0.31351	**1.63944e^{-5}**	**7.51366e^{-5}**	**7.87148e^{-5}**	**0.00019**	0.25488
bw	0.00861	0.29894	0.16311	0.44201	0.65757	0.84508

5.8 Discussion

The study shows that a NLP-based tools may help localizing the origin of the seizures in drug-resistant patients candidates to surgery, suggesting a right/left and temporal/extratemporal origin, thus representing a useful tool, especially for those epileptologists without specific skills in the interpretation of seizures-related semeiological data.

A great motivation for the research in this sector derives from the will to speed up the process that leads the patient to be considered a candidate for surgery and also to find alternative types of information that healthcare professionals can collect more easily without necessarily requiring highly specialized diagnostic machinery. Recently, many research projects aimed at developing decision support systems that could provide a more accurate classification of seizures relying on questionnaires [14]. Studies based on self-reported questionnaires are highly dependent on patients' willingness to personally contribute to the study. Therefore the trend of many research projects moved towards the use of Real World Data, so information already collected and available in electronic format into the EMRs. The main problem with EMRs is that they are mainly composed by unstructured free-text. Hence, the extraction of meaningful information requires a heavy human effort. In order to deal with textual data, many projects introduced NLP techniques and the majority of them focused on information extraction tasks. ExECT [57] is an information extraction system that combines rule-based and statistical techniques to automatically fill the routinely collected data with missing information from epilepsy clinic letters. EpiFinder [136] is a decision support system that extracts keywords to predict the probability to have epilepsy. The pipeline that we presented in this manuscript uses anonymized EMRs and semiological descriptions of seizures, supporting the diagnosis of epilepsy by attempting to ease the process of localization of the epileptogenic zone. In particular, we used NLP methods to pre-process texts and to build numerical representations of the semiological descriptions of individual seizures. We built both count-based and embedding-based models. In particular, we used supervised ML methods to classify the location of seizures origin into right/left hemisphere and temporal/extra-temporal region. Our analysis outlined that, accordingly with literature [11], embedding models obtain best performances on the learning set ($Dataset_1$). While on unseen data ($Dataset_2$) good performances are obtained by count-based model as well but this is coherent with the nature of the two numerical representations. The word embedding model is based on deep learning and when applied to $Dataset_2$ it undergoes an additional process of generalization, compared to when applied to $Dataset_1$ and so its performances decrease.

Encouraging results considering both localization tasks can be observed on the testing set of $Dataset_1$ while only the temporal/extra-temporal task obtained results well above chance level on $Dataset_2$. In this regard, it should be noted that the lateralization (left vs right) task presents some additional complexities. Specifically, some clinical signs that possess high lobe-localizing value may lack lateralization value (e.g. epigastric aura in mesial temporal lobe epilepsy), while some clinical signs (e.g., head version) may address ipsilateral or contralateral localization depending on which neuronal network is being activated. Moreover, some focal seizures may occur with bilateral signs, in which the detection of asymmetries with lateralization value is particularly challenging and consequently clinician-dependent (e.g., hypermotor seizures in frontal lobe epilepsy). Finally, some clinical and potentially lateralizing signs such as ictal/postictal aphasia may have not been always tested. Moreover, the lateralizing value of these signs may be relative to hemispheric dominance (dominant vs non-dominant hemisphere), thus not expressing an exact left-right distinction value.

The main advantage of our pipeline is that we did not adopt specific operation or information linked to the epilepsy context, e.g. we do not rely on ontologies to support models building phase. This renders our pipeline suitable for different clinical scenarios beyond epilepsy. According to our knowledge, our pipeline represents the first NLP-based diagnostic tool for drug-resistant focal epilepsy dedicated to Italian centers. Indeed our project was particularly challenging also because no pre-trained embedding models exist for biomedical applications or other existing works on this topic dealing with Italian language.

Present work also presents some limitations. The most important one is that the syntax of EMRs and the semeiological descriptions of seizures are highly dependent on the specific clinician writing them. The syntax and choice of terms of each clinician, e.g., the use of different synonyms, impact the building of both the two representations of text. In particular, the training of the word embedding model because each word depends on its context (other near words) and the most relevant features extracted by the count-based model are selected depending on their frequency.

In conclusion, identifying the EZ represents a challenging task in drug resistant focal epilepsy patients' assessment. Collectively, the results establish a starting point in the development of a non-invasive, cost-effective tool which can both enhance the pre-surgical evaluation carried out in highly specialized centers and provide a useful support in primary-care units, where many diagnostic procedures may not be available. In both cases this could diminish the delay between epilepsy onset and surgery, with a significant impact on patients' quality of life and healthcare expenditures.

Chapter 6

CDMS aimed at improving the early diagnosis of bloodstream infections

6.1 Research Question

The recent COVID-19 pandemic highlighted even more the worrying and widespread increasing circulation of pathogenic microorganisms in hospitals, sheltering for elderly, and assisted residences. The Italian *"Istituto Superiore di Sanità"* (ISS) [1] identifies *Hospital-Acquired Infections* (HAI) as the most frequent and serious complications of healthcare. A possible definition of HAI is "infections that first appear 48 h or more after hospital admission or no later than 30 days after discharge following inpatient care" [155]. So, HAIs constitute a real health emergency and require decisive action from both local and national health organizations. Candidemia, i.e. *Bloodstream infection* (BSI) caused by Candida spp., is one of the most frequent infections that affects hospitalized patients. Specifically, it is highly associated with mortality in the case of critically ill patients or those presenting septic shock [155, 83, 2, 114]. The are no specific signs and symptoms exclusively correlated to candidemia. For this reason it is impossible to distinguish it from bacteremia, which is an overall most frequent event [12, 134], unless it is executed a microbiological culture. However, this kind of exam could take up to 48-72 hours and clinicians can not wait for the results to start the most appropriate pharmacological treatment. At this point, two concerns need to be done. First, an early antifungal treatment should be given to patients that will be later confirmed as infected by Candida spp., and, second, only empiric antibacterials are more appropriate for patients

[1] https://www.epicentro.iss.it/

with bacteremia [90, 68, 151]. To define the pharmacological treatment, clinicians rely on clinical scores and bio-markers but the accuracy of these methods is still inadequate.

Over the last years, there is a growing interest in ML-based applications for early differential diagnosis [156, 69, 7], but this requires a considerable amount of data in terms of both samples and features. Thus, it is important to exploit the data reuse paradigm and avoid the manual data collection, first because it would require many human and time resources and then because it would also introduce many errors. Therefore, efficient systems able to automatically extract structured data already present into hospital LISs and EHRs should be involved.

Overall, the process of diagnosis, management and treatment of candidemia is highly complex. Guidelines and standardized treatments supporting clinicians in their clinical practice are available, but their daily application is not straightforward. To feedback this process, the EQUAL Candida score [120] was developed with the aim of measuring the adherence to guidelines in patient's management and antifungal stewardship. As a consequence it also represents a measure of the quality of candidemia diagnostic and therapeutic management and it was demonstrated that a high EQUAL score is linked to an improved survival [80]. The score is composed by a list of quality indicators, identified by the European Society for *Clinical Microbiology and Infectious Diseases* (ESCMID) and the *Infectious Diseases Society of America* (IDSA) guidance, and each of them is linked to a number of points. The score is calculated by summing the points assigned on the basis of the behavior of medical staff and it is distinguished for patients with central venous catheter (CVC) or without. The maximum score for patients with CVC is 22 while for patients without CVC is 19. For example, performing an echocardiography is moderately recommended by guidance and it is assigned 1 point for both patients with and without CVC. On the contrary the CVC removal within 24 hours from the diagnosis is assigned 3 points as it is highly recommended (of course it applies only to patients with CVC implanted).

However, a serious issue related to this medical branch is that although for more than 20 years modern LISs [73, 6] managed laboratory analysis and in last decade the management of patients related clinical data evolved from a paper-based approach towards the introduction of Electronic Medical Records (EMR)s, microbiology is one of the field where computerized systems faced many problems. Because of the lack of ad hoc and appropriate analytical fields, many meaningful information flew into textual sections of the EMRs or into the field "note" of the microbiological culture results. Accordingly, clinical texts became an important source of information but the necessity of manual inspection made the use of unstructured data expensive in terms of personnel effort and time. This phenomenon obstacle their use not only in biomedical research but it has an impact also on patient management, and care, e.g., on the daily usage of the EQUAL score because most of the needed information are contained in clinical text.

Over the last years, many research projects started exploiting the branch of AI called Natural Language Processing in the clinical scenario in order to extract embedded information from the texts written in natural language into the EMRs. In many cases the information extraction task is performed using regular expressions or target textual markers identified by experts, e.g., surgeons and infectious diseases specialists [13, 62, 182]. However, this method is limited, first, because of the necessity of prior knowledge of the pattern or words that the system needs to identify, and then because in many cases also the context (words around keywords) has an impact on the information.

The aim of my work has been building a system able to extract both from structured and unstructured data meaningful information about a patient and used this knowledge to support the differential diagnosis and management of candidemia.

6.2 Features automated extraction systems from both structured and unstructured data

6.2.1 My contribution

First I aimed at developing a decision support system able at extracting and organizing a set of features identified by clinicians as of interest from structured patients' data. These data were collected in a standard and automatic way from hospital LIS.

Then, I aimed at developing an NLP and ML-based system able at extracting from clinical diaries (free text), written by clinicians and nurses, additional information about the possible CVC implanted in a patient.

Finally, I used a ML-based system to analyse a subset of data and to find out the most influential features on the binary outcome candidemia/bacteremia (which corresponds to Candidemia Yes/No).

6.2.2 Sample Characteristics

This retrospective study has been conducted at IRCCS Ospedale Policlinico San Martino (Genoa, Italy) and it represents the first phase of the project titled "AUTO-CAND". The study has been approved by the pertinent local ethics committee (Liguria Region Ethics Committee, registry numbers 71/2020 and 159/2022). The sample is composed by the complete list of all single observations of candidemia and/or bacteremia that occurred between 1 January 2011 and 31 December 2020.

6.2.3 Experimental Design

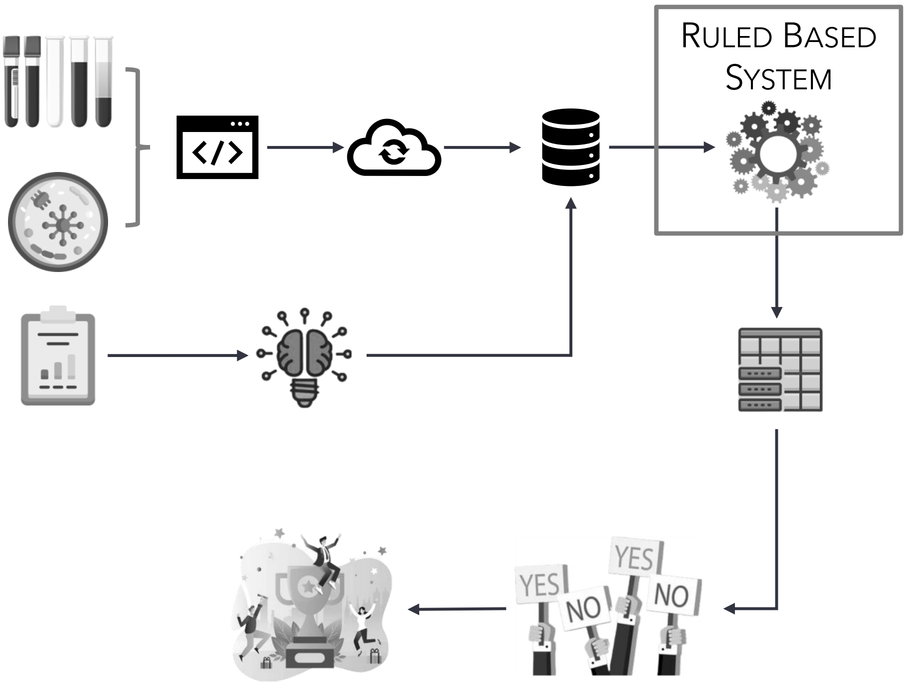


Figure 6.1 **Graphical representation of the system.** It has three main aims: first, achieving a complete knowledge about a patient merging both structured and unstructured data; second, extracting and organizing the desired features into a dataset; third, using the available knowledge to identify the most important features on the outcome "Candidemia Yes/No".

Figure 6.1 graphically presents the system macro-architecture, that is composed by 4 main actors: **Structured Data extraction and transfer system (I)**, an upgraded version of the already existing automatic data transfer from hospital LIS towards the LIDN; **Rule-based System (II)**, which extracts from the research database and organize data into the desired features; **NLP-based pipeline (III)**, which extracts from EMRs specific information that were not available or enough accurate if inferred from structured data; **ML-based system**

(IV), which uses LASSO and a majority voting process to identify the most influential features on the outcome.

Structured Data extraction and transfer system (I)

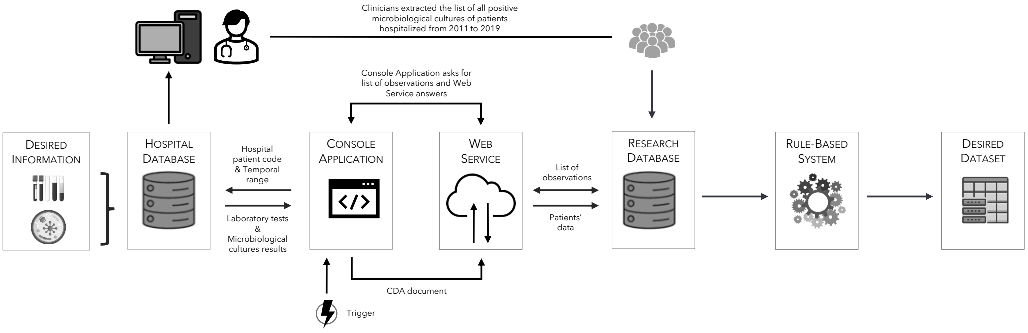

Clinicians extracted the list of all positive microbiological cultures of patients hospitalized from 2011 to 2019
Console Application asks for list of observations and Web Service answers
DESIRED INFORMATION
HOSPITAL DATABASE
Hospital patient code & Temporal range
Laboratory tests & Microbiological cultures results
CONSOLE APPLICATION
WEB SERVICE
List of observations
Patients' data
RESEARCH DATABASE
RULE-BASED SYSTEM
DESIRED DATASET
CDA document
Trigger

Figure 6.2 Elements and events chain constituting the process that leads to structured data transfer from the hospital database to the research one.

The necessary data that the system needs to retrieve can be distinguished in:

- **Laboratory tests results**, which main characteristics are the name of the exam, the result (which is usually a number), the unit of measure and the reference range of normal values.

- **Microbiological culture results**, which main characteristics are the result of the culture (positive/negative), and, if positive, the name of the microorganism and the corresponding antibiotic testing results (if performed).

At the beginning of the study all designated features derived from structured data, which is a type of data that can be easily queried and processed as it is already organized in a structured format into the hospital database. Structured data has been extracted from the Oracle views on the hospital database created ad-hoc for the 10 years-old project "The Liguria Infectious Diseases Network" [127]. After having obtained the specific approval of the Liguria Ethics Committee, the views on San Martino hospital LIS could be used to read the necessary data.

I performed the transfer and storing of data by upgrading the already existing system of data extraction from San Martino LIS towards the LIDN database. I created a console application that, for each patient's episode, reads data from the LIS and organizes them into a *Clinical Document Architecture Release 2* (CDA R2). The structure of the console application is similar to the one that I used to connect and transfer laboratory tests results from the latest center involved in the LIDN project, i.e., Sanremo hospital, towards the LIDN database. The CDA R2 is sent to the listening Windows Communication Foundation service of the LIDN that receives the standard document, validates it, and stores the information into the target database.

The structure of the CDA R2 has been built in accordance with the Implementation Guide "Referto di Medicina di Laboratorio" by HL7 Italy.

So, San Martino connection has been updated in two fundamental aspects:

1. This new connection is compliant with national regulation, which guarantees the secure transfer of the clinical data.

2. The updated architecture enables the management of a wider range of information. Before this work the CDA R2 contained only laboratory tests results, while now its structure is suitable also for containing complete microbiological culture results.

Rule-based system for features extraction (II)

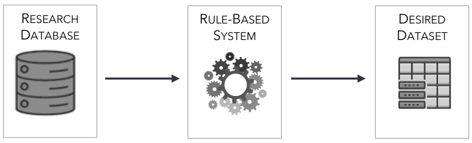


Figure 6.3 Elements involved in the process that leads to the data extraction from the research database towards the desired dataset.

The second step, after data collection, is the features extraction task. So, I built a ruled based system that, among the complete list of all patients' blood cultures, aims at: (i) identify the origin of each episode of candidemia and/or bacteremia; (ii) recognize different episodes occurring within the same patient; (iii) recognize mixed episodes of candidemia and bacteremia; (iv) differentiate bacteremia episodes by coagulase-negative staphylococci or other common skin colonizers from contamination. Then, the system extracts the designated features according to the definition given by clinicians. They can be grouped into 3 main sections:

- **Anamnestic information:** sex, age at the time of the hospitalization, hospital area of hospitalization (wards have been grouped in areas by clinicians).
- **Information derived from microbiological cultures:** name of the microorganism identifying the episode, presence of Candida spp. colonization in the 30 days prior the hospitalization, number of explored body sites during the hospitalization (divided into 3 main sections: respiratory, urinary and gastrointestinal area) and possible colonization, presence of multifocal Candida spp. colonization.
- **Information derived from a fixed list of laboratory tests:** values at the moment of the hospitalization and values at fixed endpoints from 1 to 7 days prior the hospitalization.

NLP-based pipeline for unstructured data extraction (III)

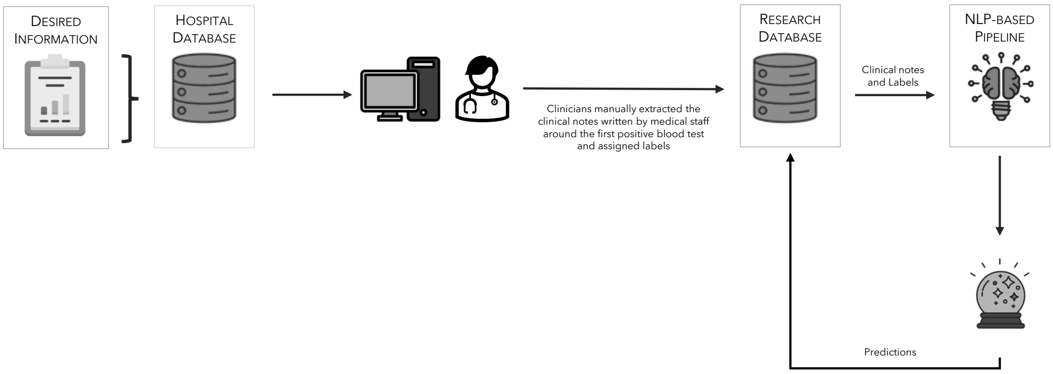

DESIRED INFORMATION
HOSPITAL DATABASE
Clinicians manually extracted the clinical notes written by medical staff around the first positive blood test and assigned labels
RESEARCH DATABASE
Clinical notes and Labels
NLP-BASED PIPELINE
Predictions

Figure 6.4 Elements involved in the process that leads to the automatic extraction and storage of information from clinical unstructured data in the research database.

During the extraction system validation phase, that will be deeply described below, the feature "Presence of CVC? Yes/No" showed a low value of accuracy if derived from structured data. Specifically, the true value of this feature was assigned if the patient showed at least one positive results of the culture exam performed on blood collection from CVC. As clinicians validated manually all the extracted features for the randomly selected observations, this process demonstrated that the low value of accuracy was not due to any technical issues. Indeed, clinicians noticed that this information was correctly reported in medical charts but not in the LIS. Therefore, I addressed the problem by developing an NLP-based pipeline that performs an *Information Extraction* (IE) task from the unstructured data present into patient's EMRs to build a more complete picture of the patient. The complete set of information stored in the database is used to support both diagnosis and management. Specifically, I developed an NLP-based pipeline aimed at answering to the following questions:

- "Presence of CVC? Yes/No" *(Task 1)*
- "CVC removed? Yes/No" *(Task 2)*
- "CVC removed within 24/48/72 hours?"

I considered as input only text contained into medical charts collected between the time of the diagnosis and the following 3 days.

The available sample for the analysis is a subset of the aforementioned one, specifically, it was composed by the Candidemia episodes that occurred between January 2018 and December 2020 at IRCCS San Martino hospital. This was related to two different aspects: first only candidemic patients in *Intensive Care Units* (ICUs) had sufficient information in their EMRs, as clinical charts of this kind of patients are very detailed; second, EMRs are computer-based only since 2018, while previously they were paper-based. So, the sample was composed by 108 patients who have been hospitalized in the ICUs at the time of Candidemia. For each patient, only the first episode was considered, and this resulted in 4236 notes extracted from EMRs (mean number of notes per patient was 38 ± 11). I excluded from the analysis 66 notes which length was less than 25 characters. So, the total number of notes considered for the analysis is 4177.

Expert clinical staff labeled the dataset at two stages and with two labels, one for each task:

At patient level:

- “Presence CVC? Yes/No”: Label = 1 was assigned if the patient had CVC implanted in the time span between the hospitalization and the following 72 hours, otherwise label = 0 was assigned.
- “CVC removed? Yes/No”: Label = 1 was assigned if the CVC was removed within 72 hours from the diagnosis of candidemia, otherwise label = 0 was assigned. For each patient, medical staff reported also the date and time of the first note containing the information of CVC removal.

At note level:

- “Presence CVC? Yes/No”: Label = 1 was assigned if the note cited positively the CVC, otherwise label = 0 was assigned.
- “CVC removed? Yes/No”: Label = 1 was assigned if the note cited a substitution, removal or insertion of a new CVC, otherwise label = 0 was assigned.

I decided to perform a classification task because the information about CVC could be written in the notes as an abbreviation, or in extended form or using the proper name of the medical device, or the words could contain misspellings. In addition to that, I considered also the case where CVC (or another homologous patterns of characters) could also be used in the sentence but with a negative meaning, e.g., "Tentivo infruttuoso di posizionamento di via Venosa centrale in ecoguida" *"Unsuccessful attempt to position the venous central via in eco-driving"*. Also the information about the removal of CVC could be written in several ways and some keywords alone without a little context could lead to misclassifications, e.g., "Sostituita medicazione CVC" *“CVC dressing replaced”* has not the same meaning as "Sostituzione CVC" *“CVC replacement”*. For this reason, the use of regular expressions alone was considered as not effective.

The other important aspect that should be considered is that the dataset is very unbalanced at note level.

Considering Task 1, only the 9% of the dataset had label = 1, therefore I decided to down-sample the notes with label = 0. So, for each patient, I considered for the analysis only the 40% of the notes labeled with 0. They were randomly extracted and the random state was set to guarantee the repeatability. The final sample was composed by 1893 notes, 19,76% with label = 1 and 80,24% with label = 0.

Considering Task 2, dataset was even more unbalanced, the 95,71% of dataset was labeled as 0. For each patient, I considered for the analysis the 10% of the negative notes, they were randomly extracted and the random state was set. The final sample for Task 2 was composed by 577 notes, 31,02% with label = 1 and 68,98% with label = 0.

In order to evaluate the performance of the pipeline but also its ability to generalize, the sample at patient level was divided into two chunks: Learning set ($n_{patients} = 97$) and Test set ($n_{patients} = 11$).

Then, as machine learning methods require in input a matrix of numbers, I searched for the most suitable numerical representation for the available textual data. I chose the NLP-based technique named "Bag of Words" and I considered the two vectorizers "CountVectorized" and "TfidfVectorizer", presented above in section *“Methods. Numerical representation of text”*. Two important parameters that needed to be defined were: size of the vectors and balance between patterns of characters and patterns of words. Specifically, I considered all the possible combination of the following:

- Sizes among 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300.
- Proportions between n-grams of characters and n-grams of words from 10% and 90% to 90% and 10%, respectively.

I performed this evaluation, on the learning set of Task 1, using a SVM with gaussian kernel and both numerical representations of text. Then I used the resulting best combination of parameters to learn the classifiers and perform predictions on the two tasks, I compared the performances of SVM with gaussian kernel with those of SLR and Random Forest. This comparison was executed both for the learning set and the blind test set.

ML-based system for features importance evaluation (IV)

As last step of the analysis, I merged all the features extracted from structured and unstructured data and build a ML-based system aimed at determining the most influential features on the outcome candidemia/bacteremia. The sample of data was composed by a sub-sample of randomly selected candidemia episodes ($n_{candidemia}$ = 44) and a list of randomly selected bacteremia episodes for which also the feature "CVC present? Yes/No" was manually extracted by clinicians and added ($n_{bacteremia}$ = 61). So, the complete sample was composed by 105 randomly selected episodes and the corresponding 43 features listed below in table 6.1.

Before performing missing data imputation and normalization, I removed from the sample the features that showed a low presence percentage in the dataset. Specifically, I set the threshold at 40%, and I kept only features whose presence level was equal or greater than the threshold. This resulted in a list of 15 removed features. Then I performed missing data imputation using the K-Nearest Neighbours and normalized data with Z-score. To investigate the most influential features on the outcome, I performed features selection using LASSO in a majority voting process. Specifically, I ran LASSO in a 10-fold cross-validation process and stored, for each fold, the features with a not null coefficient in the matrix. This process has been repeated three times, shuffling the data each time, and setting the random state equal to the iterator in order to make results reproducible. Then, I ran, in a 10-fold cross validation process, SVM classifier with Gaussian kernel for each value of N (ranging from 0 to 30, which is the number of times each feature has a not null value in the LASSO coefficients matrix) and training the SVM using all features voted at least N times. As for LASSO, I repeated the process three times shuffling the data each time in a reproducible way. Finally, I

Table 6.1 *Complete list of features considered for the analysis. Note that: sex and features extracted from microbiological cultures are binary. While age, number of explored body sites and all the laboratory exams results are discrete variables.*

Sex	Prothrombin time
Age at hospitalization	Uric acid
Presence of CVC?	Alkaline phosphatase (ALP)
Candida colonization in the previous 30 days	Alanine aminotransferase (ALT)
Multifocal Candida colonization	Aspartate aminotransferase (AST)
Number of explored body sites for colonization	Direct bilirubin
Respiratory colonization by Candida spp.	Total bilirubin
Urinary colonization by Candida spp.	Creatinine
Gastrointestinal colonization by Candida spp.	Gamma-glutamyl transferase
Basophil cells count	Lactate dehydrogenase
Eosinophil cells count	Urea
Lymphocyte cells count	Glycated hemoglobin
Monocyte cells count	Glucose
Neutrophil cells count	Albumin
Hematocrit	Beta-D-glucan
Hemoglobin	C-reactive protein (PCR)
White cells count	Procalcitonin
Red cells count	Total proteins
Platelet count	Lactate from arterial blood
Activated partial thromboplastin time (APTT)	Lactate from venous blood
International normalized ratio (INR)	Triglycerides

evaluated models performances in terms of mean values of accuracy, weighted precision and weighted f1-score (weighted recall is equal to accuracy) across the ten folds and across the three shuffles.

6.2.4 Features extracted from Structured Data

The complete sample of all positive blood cultures for Candida spp. and/or bacteria collected in the time span from 1st January 2011 to 31st December 2019 at San Martino hospital was composed by 65,767 records. First a performance validation of the extraction system has been done. Specifically, a subgroup of 381 randomly selected observations (derived from different patients) was considered for manual review. This process involved two clinicians that compared the information extracted for each feature of each observation with the original data present in the hospital database. All the features obtained accuracy values above 99%. The only exception was the feature named "Presence of CVC? Yes/No", which obtained a value <80% despite improvements and targeted modification of the automated extraction code. So the feature has been excluded temporarily from the analysis.

Once the validation process was completed successfully [70], I used the ruled based system to filter the list of sample blood cultures, to retrieve from hospital database only the strictly necessary data. Specifically, I excluded from the sample the observations that were not the origin of the episode, i.e., the first positive blood culture for the specific microorganism, and blood cultures that could be contaminations. The definition of contamination has been described as follows "the event that only one blood culture, which blood collection was performed through skin, tested positive for a microorganism in the list of contaminating bacteria". The resulting sample obtained from the first round of pre-processing was composed by 16,102 episodes. Then, within mixed episodes, i.e., episodes where patient tested positive for both candidemia and bacteremia, I excluded observation of bacteremia in order to keep only observations of candidemia for each mixed episode.

The final dataset was composed of 15,752 episodes, distinguished as follows:

- Bacteremia episodes: 14,112
- Candidemia episodes: 1,338
- Mixed episodes: 302

6.2.5 Feature extracted from Unstructured Data

This section reports results obtained with Actor III, presented above, which aims at extracting information from patients clinical texts about the possible presence of CVC implanted and store them in the database, in order to build a complete patient picture.

Evaluation of Numerical Representations of Text

Model performances has been compared across all possible combinations of sizes and proportions of patterns considering both numerical representations of text within the task “Presence CVC? Yes/No”. Each couple size-proportion is evaluated in terms of the mean values of weighted F1-score obtained across the three rounds. Considering the numerical representation obtained with CountVectorizer, figure 6.5 shows how model performance varies depending on both size and proportions between patterns. Highest and lowest values are summarized in Table 6.2.

Considering the numerical representation obtained with CountVectorizer, figure 6.5 shows how model performance varies depending on both size and proportions between patterns.

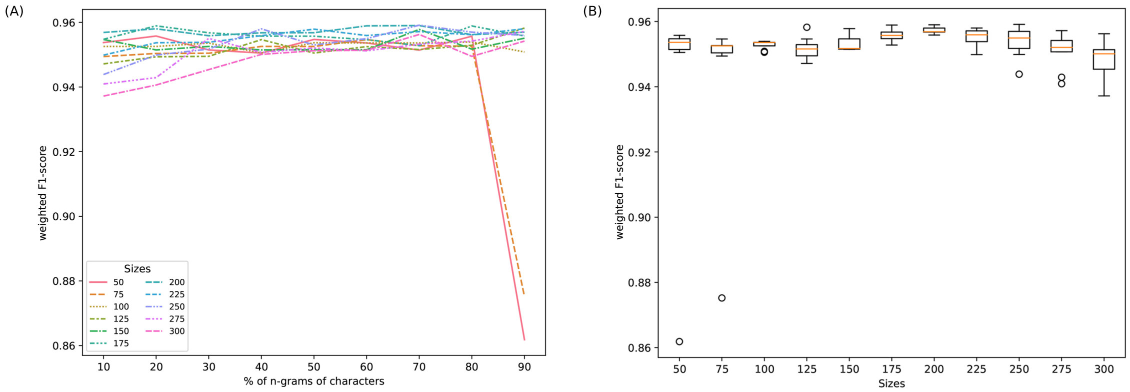


Figure 6.5 **Model performance evaluation of numerical representation built with CountVectorizer.** Subfigure (A) graphically presents performances as both sizes and proportions between patterns of characters and words vary. Sizes are graphically presented as lines with different pattern and color, while proportions are represented on the abscissa axis. Specifically, the abscissa axis corresponds to the percentage of n-grams of characters and the number of n-grams of words is obtained as 100 - number of n-grams of characters, as the sum of the two is the 100%. Subfigure (B) represents the same quantity but grouped per size. It shows that some sizes are most sensitive to changes in proportions between n-grams while others are stable.

Highest and lowest values are summarized in 6.2. It is possible to see that at lower sizes (50 and 75) model stability highly depends on the proportion between the two patterns of n-grams. Specifically, as the number of n-grams of characters becomes greater than the number of n-grams of words, the performance quickly decreases. On the contrary, the difference between best and worst performance becomes around 1% or less for middle sizes (from 100 to 225), which demonstrate the slightly low dependence of proportion between patterns. This difference increases again at the highest sizes (250, 275, 300), that depends again on the proportion of patterns and, specifically, this dependency is the opposite of lowest sizes. So, it is important to highlight that best performances are comparable, but at lowest and highest sizes it is important to consider also proportions between pattern. However, I chose to consider the best combination of size and proportions for the numerical representation build with CountVectorizer, which is: size = 250, percentage of n-grams of characters = 70% and percentage of n-grams of words = 30%.

The phenomenon observed with CountVectorizer is partially confirmed considering the numerical representation obtained with TfidfVectorizer. Performance at lower sizes is still highly dependent on the proportion between patterns and specifically it decreases

Table 6.2 *Comparison between best and worst performance in terms of weighted F1-score for each total number of features extracted with CountVectorizer. For each performance measure, the combination of patterns of characters and words that produced the result is reported.*

Sizes	Best performance	% n-grams characters (% n-grams words)	Worst performance	% n-grams characters (% n-grams words)
50	0.956	20 (80)	0.862	90 (10)
75	0.955	60 (40)	0.875	90 (10)
100	0.954	80 (20)	0.951	40 (60)
125	0.958	90 (10)	0.947	10 (90)
150	0.958	70 (30)	0.952	40 (60)
175	0.959	20 (80)	0.953	70 (30)
200	0.959	70 (30)	0.956	30 (70)
225	0.958	90 (10)	0.950	10 (90)
250	0.959	70 (30)	0.944	10 (90)
275	0.957	90 (10)	0.941	10 (90)
300	0.956	70 (30)	0.937	10 (90)

quickly when patterns of characters become grater then patterns of words. On the contrary it demonstrates a more stable performance at higher sizes.

Table 6.3 *Comparison between best and worst performance in terms of weighted F1-score for each total number of features extracted with TfidfVectorizer. For each performance measure, the combination of patterns of characters and words that produced the result is reported.*

Sizes	Best performance	% n-grams characters (% n-grams words)	Worst performance	% n-grams characters (% n-grams words)
50	0.956	50 (50)	0.853	90 (10)
75	0.955	70 (30)	0.875	90 (10)
100	0.954	20 (80)	0.948	10 (90)
125	0.957	60 (40)	0.951	10 (90)
150	0.958	50 (50)	0.946	10 (90)
175	0.958	40 (60)	0.950	90 (10)
200	0.957	50 (50)	0.942	10 (90)
225	0.956	40 (60)	0.943	10 (90)
250	0.951	70 (30)	0.945	10 (90)
275	0.950	70 (30)	0.943	50 (50)
300	0.954	20 (70)	0.941	30 (70)

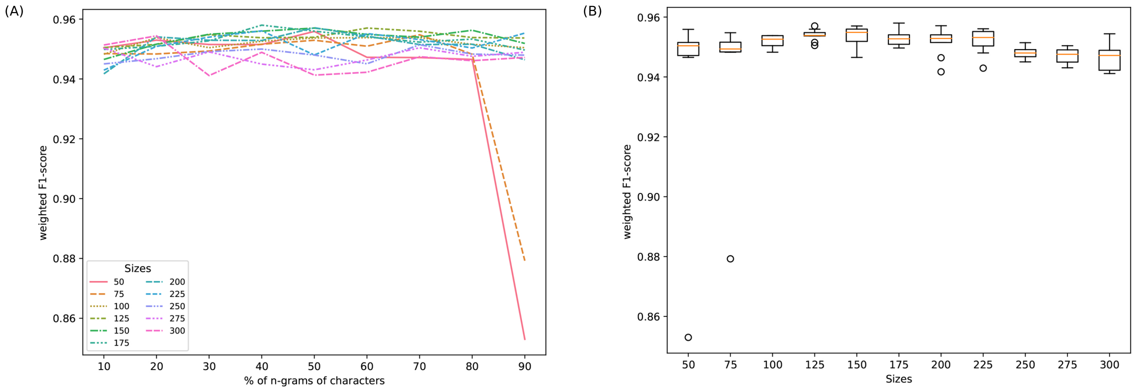


Figure 6.6 **Model performance evaluation of numerical representation built with TfidfVectorizer.** Subfigure A graphically presents performances as both sizes and proportions between patterns of characters and words vary. Sizes are graphically presented as lines with different pattern and colour, while proportions are represented on the abscissa axis. Specifically, the abscissa axis corresponds to the percentage of n-grams of characters and the number of n-grams of words is obtained as 100 - number of n-grams of characters, as the sum of the two is the 100%. Subfigure B represents the same quantity but grouped per size. It shows that some sizes are most sensitive to changes in proportions between n-grams while others are stable.

The best combination of size and proportions for the numerical representation build with TfidfVectorizer is: size = 175, percentage of n-grams of characters = 40% and percentage of n-grams of words = 60%.

The two best numerical representations have been used to train SLR and Random Forest and performances has been compared with those obtained with those of SVM classifier with rbf kernel. First I tested the stability of models in a 5-folds cross-validation process (performed at patient level) on the learning set, performed three times each time shuffling the data, and then I tested model ability to generalize using as input a set of unseen data.

Model performance on task 1: "Presence CVC? Yes/No feature extraction"

Performances at note level are summarized in table 6.4.

Table 6.4 *Task 1: Presence CVC? Yes/No. Performances evaluation in terms of weighted F1-score on the learning set at note level.*

Classifier	**CountVectorizer**	**TfidfVectorizer**
SVM with rbf kernel	0.954	0.954
SLR	0.951	0.951
Random Forest	0.924	0.932

Then, predictions have been grouped at patient level in order to assess the usability of this kind of system to extract the feature "Presence CVC? Yes/No" and include it in the analysis to improve differential diagnosis of candidemia. I assigned the feature "Presence CVC?" = True if a patient had at least one note classified as True. Results are summarized in table 6.5.

Table 6.5 *Task 1: Presence CVC? Yes/No. Performances evaluation in terms of weighted F1-score on the learning set at patient level.*

Classifier	CountVectorizer	TfidfVectorizer
SVM with rbf kernel	0.980	0.980
SLR	0.953	0.971
Random Forest	0.936	0.945

As 3 shuffles were performed, the mode of the three values was considered to obtain performances at both note and patient level. Finally, model ability to generalize was evaluated. Specifically, I learned the above-mentioned three classifiers with the best set of hyper-parameters obtained on the learning set and used them on a test set composed by unseen data. Table 6.6 summarizes obtained performances.

Table 6.6 *Task 1: Presence CVC? Yes/No. Performances evaluation in terms of weighted F1-score on the blind test set at patient level.*

Classifier	CountVectorizer	TfidfVectorizer
SVM with rbf kernel	0.758	0.837
SLR	0.916	0.916
Random Forest	0.758	0.758

The overall performance of the three classifiers is higher using the numerical representation of text built with TfidfVectorizer. However, the best performances reach the same values, specifically, only 1 patient out of the 11 of the blind test set is misclassified.

A statistical analysis was conducted to confirm the hypotheses. The repeated measures considered were the F1 scores obtained from the classification process on the learning set at note level. The total number of measures consisted in 15 values deriving from the 5-fold cross validation over 3 shuffles. Specifically, the impact of the two different numerical representations was tested with Wilcoxon Test [194] (null hypothesis: same performance) and it proved the existence a significant difference considering all three classifiers (p-value < 0.05), as shown in Table 6.7.

Then, the impact of the choice of the classifier on the outcome was tested for each numerical representation of text using first Mann-Whitney U Test [117] (null hypothesis:

Table 6.7 *Task 1: Probability that the two different numerical representations of text induce the same performance, in terms of weighted F1 score, considering each classifier (null hypothesis).*

	CountVectorizer vs TfidfVectorizer
SLR	0.03015
SVM with rbf kernel	0.01025
RF	0.00262

same performance) and it resulted significant, as shown in Table 6.8. Therefore, multiple comparisons were executed, and the use of RF resulted statistically significant considering both numerical representations of text (p-value, corrected with Bonferroni correction < 0.05).

Table 6.8 *Task 1: Probability that classifiers have the same performance, in terms of weighted F1 score, considering each numerical representations of text (null hypothesis).*

	SLR vs **SVM with rbf kernel**	**SLR** vs **RF**	**SVM with rbf kernel** vs **RF**
CountVectorizer	0.12404	6.10352^{-5}	$6.10352e^{-5}$
TfidfVectorizer	0.12054	$6.10352e^{-5}$	$6.10352e^{-5}$

Model performance on task 2: "CVC removed? Yes/No feature extraction"

As I did for Task 1, first, I evaluated performances of all combination of text representation and classifiers on the learning set. At note level, the average between the performances over the three shuffles in terms of weighted F1-score are summarized in Table 6.9.

Table 6.9 *Task 2: CVC removed? Yes/No. Performances evaluation in terms of weighted F1-score on the learning set at note level.*

Classifier	**CountVectorizer**	**TfidfVectorizer**
SVM with rbf kernel	0.893	0.901
SLR	0.896	0.896
Random Forest	0.887	0.913

Then, I grouped predictions at patient level, specifically I assigned the feature "CVC removed? Yes/No" = True if a patient had at least one note classified as True. As three shuffles were performed, I considered the mode of the three values. Results are summarized in Table 6.10.

Table 6.10 *Task 2: CVC removed? Yes/No. Performances evaluation in terms of weighted F1-score on the learning set at patient level.*

Classifier	CountVectorizer	TfidfVectorizer
SVM with rbf kernel	0.870	0.909
SLR	0.911	0.890
Random Forest	0.889	0.920

In order to assess model ability to generalize even in the second task, I learned the above-mentioned three classifiers with the best set of hyper-parameters obtained on the learning set and used them on the test set composed by unseen data. Table 6.11 summarizes the performances at note level while Table 6.12 summarizes the performances at patient level.

Table 6.11 *Task 2: CVC removed? Yes/No. Performances evaluation in terms of weighted F1-score on the blind test set at note level.*

Classifier	CountVectorizer	TfidfVectorizer
SVM with rbf kernel	0.813	0.921
SLR	0.859	0.941
Random Forest	0.836	0.853

Table 6.12 *Task 2: CVC removed? Yes/No. Performances evaluation in terms of weighted F1-score on the blind test set at patient level.*

Classifier	CountVectorizer	TfidfVectorizer
SVM with rbf kernel	0.758	0.837
SLR	0.758	0.837
Random Forest	0.758	0.758

These results partially confirm the trend identified in task 1, so in most of the cases the performance of the three classifiers is higher with the numerical representation obtained with TfidfVectorizer. An analogous statistical analysis was performed also on results of task 2. From the point of view of numerical representation of text, a statistically significant difference was observed only when using RF, as shown in Table 6.13.

While, the choice of classifiers show statistical significance when comparing performances of SLR and RF on dataset obtained with TfidfVectorizer (p-value = 0.00262 < α^*, corrected with Bonferroni = 0.01695).

The process of storing predictions into the database allowed the conduction of other deeper investigations, essential for calculating the EQUAL score. Not only I could analyze the performance of my system in answering the question “CVC removed? Yes/No” but

Table 6.13 *Task 2: Probability that the two different numerical representations of text induce the same performance, in terms of weighted F1 score, considering each classifier (null hypothesis).*

	CountVectorizer vs TfidfVectorizer
SLR	0.97797
SVM with rbf kernel	0.63867
RF	0.00763

storing the predictions together with all the other information about patient and episode allowed also to investigate the concept of time distance between the diagnosis and the CVC removal.

Specifically, considering the learning set, for each patient I performed two investigations. First, I compared the prediction of the system in the identification of the specific note that medical staff labeled as containing the information of CVC removal for the first time. I measured the level of agreement (at patient level) in terms of accuracy considering both numerical representations of text and I summarized the results in Table 6.14. As previously mentioned, I performed three times the classification process shuffling the data each time, so I considered the mode of the values for each classifier.

Table 6.14 *Performances evaluation in terms of accuracy on the learning set on correctly labelling the note identified by medical staff as the first containing the information of CVC removal.*

Classifier	**CountVectorizer**	**TfidfVectorizer**
SVM with rbf kernel	76.36%	73.61%
SLR	77.78%	77.78%
Random Forest	80.56%	81.94%

Then, I analyzed the level of agreement in the concept of “first” note containing the information of CVC removal. Specifically, I compared the first note identified by the medical staff and the first note identified by each classifier considering as time 0 the diagnosis date, results are shown in Figure 6.7.

The level of agreement on the first note containing the information of CVC removal is lower (percentage of predictions labelled as ‘Coincide’ is between 60% to 68% for all classifiers) compared to the agreement on the classification of the first note identified by medical staff and to the overall information of the removal of CVC. I analyzed the notes belonging to the categories other than “Coincide” in order to find out possible patterns in the errors.

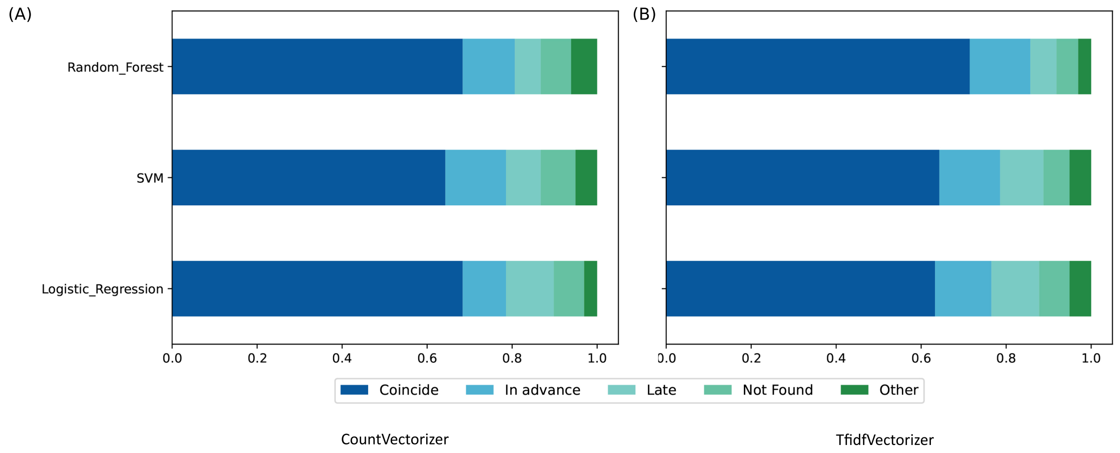


Figure 6.7 **Level of agreement** between manual labeling and the algorithm on the first note containing the information about CVC removal.

Other. It is the case where medical staff labeled the patient as 0 (CVC not removed) while one of the classifiers suggested a note as containing the information of CVC removal. Most misclassified samples highlighted the presence of many keywords linked to the information of CVC removal but together with other context words, so the meaning of the sentence was different. While two cases contained the mention of a previous removal of the CVC, so system classified correctly the note because it contained the information of a CVC removal, but it was not linked to the present episode.

Late. The classifier did not identify the note labeled as first by medical staff, but it found another note containing the information but later in time. In 63% of cases considering CountVectorizer and 50% of cases considering TfidfVectorizer, the other note detected correctly contain the mention the CVC removal and it is in the same time span of 24 hours of the note labeled by medical staff.

In advance. The classifier did not identify as first the note labeled by medical staff (they may have correctly classified it, but it is not the first one) because it found another note containing the information earlier in time. Most of the errors are due to the scheduled removal of the CVC which is not yet performed. During the review of this specific error, I found out that the system detected two human errors in the labeling process which had a negative impact on the outcome because less points are assigned according to the EQUAL score as the time between diagnosis and CVC is removed.

Not found. The classifier could not identify any note as containing the information about CVC removal.

Finally, I decided to evaluate the impact of the prediction on the temporal outcomes essential for the EQUAL Candida Score: “CVC removal $\leq$ 24 hours from diagnosis” and “CVC removal $>$ 24 and $<$ 72 hours from diagnosis”. The accuracy for each classifier and numerical representation is summarized in Table 6.15.

Table 6.15 *Performances evaluation on the learning set on correctly labeling the note identified by medical staff as the first containing the information of CVC removal.*

Classifier	**CountVectorizer**	**TfidfVectorizer**
SVM with rbf kernel	78.57%	78.57%
SLR	81.63%	75.51%
Random Forest	81.63%	83.67%

This result shows that the performance in identifying the “first” label does not effectively impact the outcome because another note in a very near time span (24 hours at maximum) is detected.

6.2.6 Most influential features on the outcome candidemia

Once features from both structured and unstructured data were organized in the desired dataset, I investigated their impact on the outcome "Candidemia Yes/No" using the ML-based system presented above as "Actor IV". Specifically, I aimed at assessing the need of extracting the feature "CVC present? Yes/No" from texts.

So, I executed a majority voting process using LASSO in order to evaluate the changes in model performances depending on the number of best selected features. This process is composed by a 10-fold cross-validation and repeated three times shufflig the data each time. The number *N* of selected features ranges from 0 (all features are considered for the classification) to 30 (only features selected as important in each fold and at each iteration are considered).

As shown in Fig. 6.8, N = 11 is the best choice as it is the value where all the three metrics trends reach a peak. So, among the complete list of features, I extracted those selected at least N = 11 times, that are: PCR, albumin, glucose, creatinine, total bilirubin, INR, APTT, platelet count, red cells count, urinary colonization by Candida spp., number of explored body sites for colonization, candida colonization in the previous 30 days, presence of CVC, and age. I analysed their importance according to LASSO coefficients matrix. Specifically, the absolute value of each feature indicates its importance in order to predict the outcome. The features with a positive weight support the prediction of the outcome when its value is 1 (patient affected by invasive candidiasis), and the features with a negative weight support

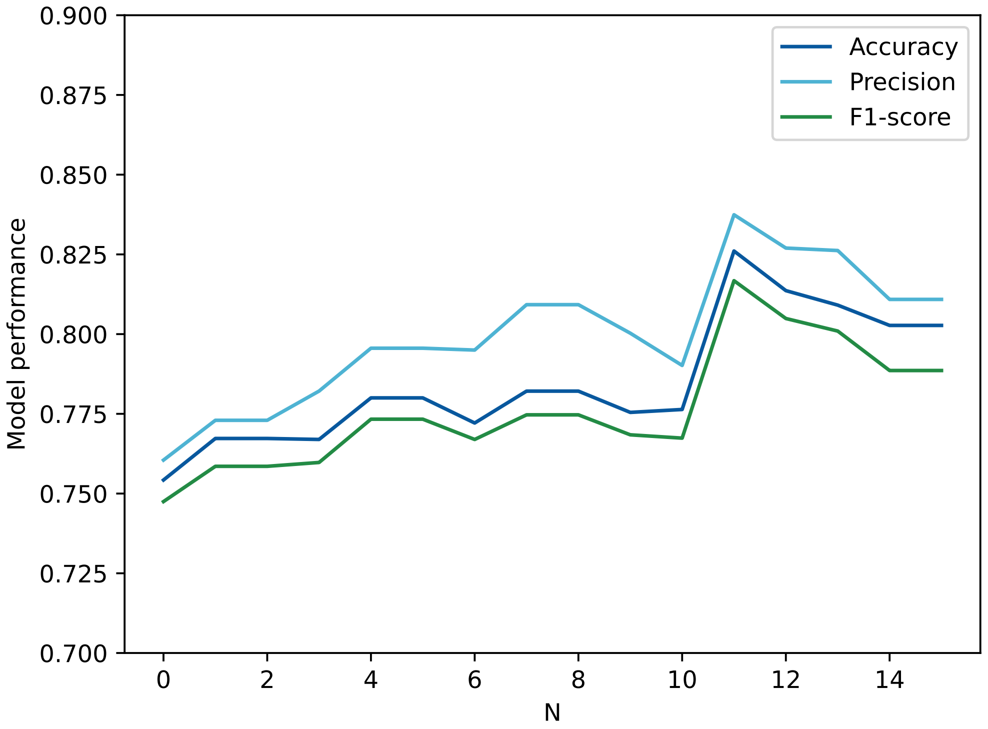


Figure 6.8 **Model performance trend as N ranges.**

the prediction of the outcome when its value is 0 (patient affected by bacteremia). Model results, showed in Fig. 6.9, confirm that CVC feature has a high influence on the diagnosis of invasive candidiasis.

6.2.7 Discussion

The work presented within this use case shows that a system able at extracting and combining both structured and unstructured information about patients hospitalized with a bloodstream infection can support the process of differential diagnosis and management of candidemia.

First of all because the automatic connection with the hospital database allowed for the creation of a large dataset, an extremely time-consuming task that would have been difficult or even impossible to achieve manually, as suggested by other studies on the topic performed within the same center that used a much smaller sample [67, 122]. Another advantage of this approach was the high quality of the data, which was evaluated by comparing the data collected automatically with those that would have been collected manually [70, 48, 174]. Indeed, despite the detection of a single technical error during manual validation, it is worth

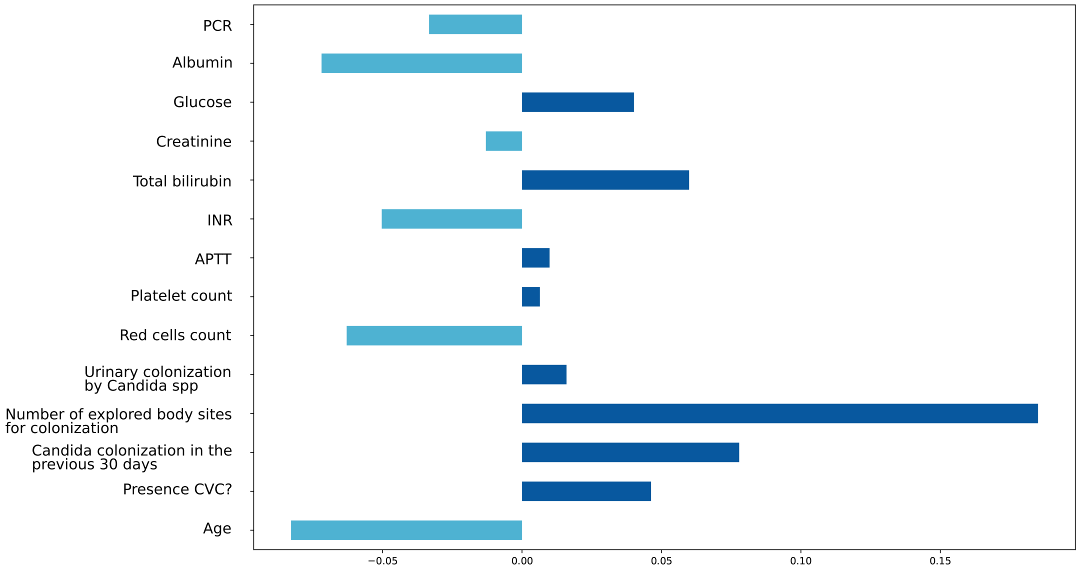


Figure 6.9 **Features importance on the outcome Candidemia Yes/No** derived from LASSO coefficients matrix.

noting that the expected rate of errors during automated extraction is far lower than the one expected with manual imputation of data [100, 75]. The complete dataset will be used within the ongoing project named AUTO-CAND to assess the performance of several ML models on the task of early recognition of candidemia. The main limitation of the automatic extraction system is that its use is restricted to structured data, i.e., laboratory exams and microbiological cultures results. During the process of validation it was demonstrated that the presence of CVC implanted in the patient was not possible to deduce from the structured data and so it was decided to look for that information in the textual sections of the EMRs. However, unstructured data needed to be extracted manually from EMRs because the de-identification of this kind of data was not as easy as that of structured one. So, to preserve patients privacy, texts have been manually revised in order to remove all direct (e.g., first and last name) and quasi (e.g., name of the doctor involved) identifiers that could refer to the patient.

Therefore, in order to built a more complete picture of the patient, I developed a computational pipeline which combines ML and NLP-based methods to extract from text the information about presence and removal of CVC in a subset of critically ill patients hospitalized in ICUs. Furthermore, this constitutes a first step towards the automatic calculation of the EQUAL Candida Score from unstructured data contained into patients EMRs. The improvements of the process, that leads to the score calculation, are ultimately aimed at positively impacting patients' care. First, by providing a real time feedback (alert) to clinicians on any possible low EQUAL Candida Score in the very first days after the diagnosis of candidemia (i.e., when clinicians are still able positively impact the score), thereby improving the adherence to international guidelines and, in turn, patients' survival [80].

First, I aimed at solving two binary classification problems to obtain fundamental information to obtain a complete patient picture and for the calculation of EQUAL Candida Score. The first one is "Presence CVC? Yes/No", this is necessary to discriminate between CVC carriers and non-CVC carriers. The second one is "CVC removal? Yes/No", this is necessary because the removal of CVC, if performed in a certain time span, has an impact on the score. To do that, I devised a system which exploits the Bag of Words technique to create the numerical representation of texts necessary as input for ML models. Differently from other research projects aimed at extracting information from EMRs [27, 161, 142], the main advantage of the system is that it does not entirely rely on regular expressions or a rule-based approach. Specifically, it automatically selects the most relevant features for the outcome from the text. This means that there is no need for human intervention in the definition of keywords, that could also alter system performances if searched alone without any context. In addition to that, the system may find other links between words and characters that support the outcome. In addition to that, the use of n-grams of characters supports in addressing misspellings and abbreviations, that would cause misclassification if only the exact word was used as feature [142]. Finally, as any a-priori knowledge is involved, the system is not specifically designed for one purpose, so it could be used for both presented tasks without any need of intervention at structural level. However, the presented approach also presents some limitations that will require future investigations. First, as the CVC could be mentioned in many ways, the system is not able to identify it when less frequent synonyms are used. For that purpose, it will be necessary to investigate the impact on system performance of including in the data preprocessing phase a text normalization step. The other main limitation is that text data needs to be manually de-identified by clinicians as no automatic system is available. So, this has an impact on data availability and timeliness of collection and analysis.

As second step, I aimed at identifying the time span that occurred between the diagnosis and the eventual CVC removal to support the assignment of +3 points (CVC removal within 24 hours from diagnosis) and +2 points (CVC removal within 24 and 72 hours from diagnosis). To do that, the predictions have been stored together with patients' data into an ad-hoc database so that each prediction at note level was linked to the specific text and consequently to the date and time of the note. The main advantage of the approach is that

ML methods have been applied to predict the outcome “CVC removal? Yes/No” but then only queries on the SQL Server database have been used to determine the distance from the diagnosis. Specifically, it was not necessary to predict if the CVC was removed within 24 or 72 hours, because the information extracted from unstructured data have been linked to those obtained from the structured one (date and time). On the contrary, the main disadvantages of the presented system are that it is not able to distinguish between the mentions of CVC removal linked to a previous episode or to the present one, as highlighted by errors named “Other” or “Late”. It also happened that some notes did not contain any mention of CVC removal, but they only mentioned an insertion. The system correctly identified those cases related to a new insertion but missed those that could be interpreted as a new insertion only by reading the whole clinical history of the patient. Finally, the system was not able to distinguish between the notes that contained a request to remove the CVC, but they did not mention the actual removal, as highlighted by errors named “In advance”.

Finally, I evaluated the impact of the feature "CVC presence" on the outcome Candidemia Yes/No (which is equivalent to candidemia/bacteremia because I removed from the sample episodes labeled as bacteremia in patient also confirmed as candidemic) and it resulted to be one of the most influential. However, the use of the computational pipeline to extract an information that showed to be meaningful in both diagnosis and management is limited by the fact that only patients hospitalized in ICUs have a complete and detailed set of notes describing the procedures executed. Future investigations will be necessary to evaluate the possibility of extending this approach also to patients hospitalized in other wards.

In conclusion, this work achieved encouraging results considering both the information extraction of CVC presence and CVC removal. Further investigate methods that will support the detection of temporal references within the text are needed.

6.3 Automatic extraction of a complete microorganism’s picture from microbiological notes

During the project on the differential diagnosis of candidemia, I approached the microbiological report and its structure. I analyzed the one produced by San Martino hospital and also other hospitals to evaluate the possibility of generalizing the process of building the CDA R2 document. I noticed that often the name of the microorganism is not always contained in the analytical part of the database, while it is mentioned in the unstructured sections in various ways. For this reason I started another NLP application case study [126].

6.3.1 My contribution

The objective of my our work was to build a NLP-based pipeline for automatic information extraction from the notes of microbiological culture reports. This could represent a first step toward the development of a system able to monitor antibiotic prescriptions at a hospital and territorial level in the Abruzzo Region [27].

6.3.2 Sample Characteristics

The collected sample was derived from the LIS of the main hospital of Pescara in Abruzzo Region and was obtained from clinical notes extracted from a 1 month collection of anonymized laboratory referral. It was composed of 499 texts from culture reports, classified as follows:

- **276 (55.3%)** texts containing the name of a microorganism where an expert from the hospital confirmed its presence;
- **56 (11.2%)** texts needing to be filtered because they contained a pattern that is not useful for our analysis and was, thus, removed. An example of a sentence belonging to that pattern is the following: *"Integration of the preliminary report sent on . . . "*. Indeed, I considered the use of synonyms, e.g., "provisional" instead of "preliminary", and the presence of orthographic errors, e.g., missing letters. Therefore, I decided not to use regular expressions alone as first attempt.

After having obtained the authorization from Abruzzo region to access the entire LIS system at a regional level, the proposed system will massively test with notes produced by a wide range of persons.

6.3.3 Experimental Design

The complete schema of the developed pipeline is presented in Fig. 6.10, and it can be divided into four main sections:

Data preparation (I)

First, I prepared the inputs values, so I cleaned the clinical notes by removing punctuation and substituting patterns that could be dates with the word "data". Then, I tokenized and proceeded with stop-words removal. I considered only words longer than one character in order to exclude from the analysis strings belonging to the class of prepositions, articles, and

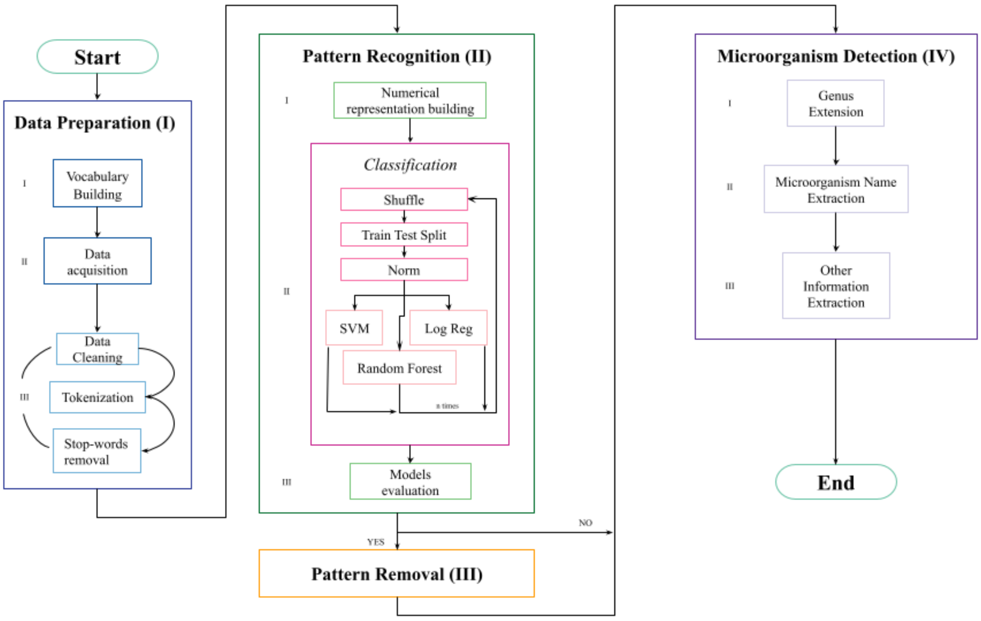


Figure 6.10 **Complete schema of the pipeline.** It can be divided into 4 main sections: data preparation, pattern recognition and removal, and microorganism detection.

adverbs, while keeping single letters that could be the abbreviation of a genus name. Then I build the vocabulary database, and loaded both vocabulary and clinical notes using pyodbc tool.

Pattern recognition (II)

Once I loaded and cleaned data, I needed to convert text into a numerical representation that could be used as input for ML algorithms. *Numerical representation building (II.I):* I adopted the bag of words technique. This choice was guided by the structure of the sentences that was fragmentary and did not respect any strict syntactic rules. Therefore, I preferred to use a context-free representation. The resulting numerical representation was composed of both n-grams of characters and n-grams of words following the proportion of 70:30. I decided to select more features composed by n-grams of characters in order to deal with misspellings, abbreviations, and limited syntactic rules. I tested the model performance considering 10 possible total numbers of selected features from 10 to 100 with a step size of 10. I obtained the best performance with a total number of features equal to 90 (Section...). *Classification (II.II):* I used the aforementioned numerical representations to learn a supervised binary classifier to predict whether the observed pattern was present in the clinical note or not. Specifically, I compared the performances of three classifiers: SLR, SVM with Gaussian kernel and Random Forest. I split the dataset into a learning and a testing set with the proportion of 80:20. On the learning set, I performed the hyperparameter search through a tenfold cross-validation, which iteratively split the learning set into a training and validation set. They were respectively used to learn the model with all the possible combinations of hyperparameters and to evaluate the performances thereafter. Then, I learned the three models with the selected set of best hyperparameters, and I evaluated the model performances on the testing set. I repeated the classification 20 times, shuffling the data each time. In order to guarantee repeatability of results, I set the random state equal to the loop index. *Models Evaluation (II.III):* I evaluated the performance of the three ML models in terms of accuracy.

Pattern Removal (III)

Once the algorithm classified the clinical notes as “containing”/“not containing” the pattern, I used regular expressions to remove the uninformative pattern from the identified notes. The schema of the regular expression was as follows:

`\b[Ii]\w.+?\bdata\b`.

The elements of the expression are defined below:
`\b` asserts the position of a word boundary. In this case, I want the pattern beginning with ‘I’ (the first letter of the word ‘Integrazione’ (integration), which can be abbreviated and/or can be uppercase or lowercase in the notes).
`\w` matches any word character and ends with ‘data’ (the word that I substituted for all dates in the data cleaning phase).
`.` matches any character (e.g., letters, numbers, and punctuation) except for line terminators.
`+?` matches the previous token between one and unlimited times, the fewest times possible, but expanding as needed.

Microorganism detection (IV)

Genus Extension (IV.I): I stored the microorganism names using the binomial nomenclature originating from the Linnaeus classification [41]. It is composed of two terms: the first is the genus name with the first letter capitalized; the second is the species name in lower case. Usually, after a microorganism’s name is written once in a text, then it can be substituted with its first capital letter, followed by a period, in subsequent mentions. However, considering the brevity of the clinical notes, a shared agreement is to always write the abbreviated form, despite the entire genus having not yet been introduced.

This binomial nomenclature does not allow the use of an abbreviation composed of two letters for the genus. Nevertheless, even though the microorganisms should be written according to this strict rule, I decided to keep words composed of only one character and not to use a regular expression, because I considered that abbreviations may be not written correctly, e.g., by using abbreviations that are not followed by a period, or where uppercase letters are followed by a period but not followed by lowercase letters. Hence, I performed the extension of the microorganism genus. Specifically, I compared the ‘‘n + 1’’ token with each species of the vocabulary. If the similarity index between the two tokens was greater than or equal to 98, then I checked the token ‘‘n’’. If the token ‘‘n’’ began with the same letter of the genus of the species in position ‘‘n + 1’’, I substituted the token ‘‘n’’ with the genus name found in the vocabulary. The schema of the treatment is presented in Fig. 6.11.

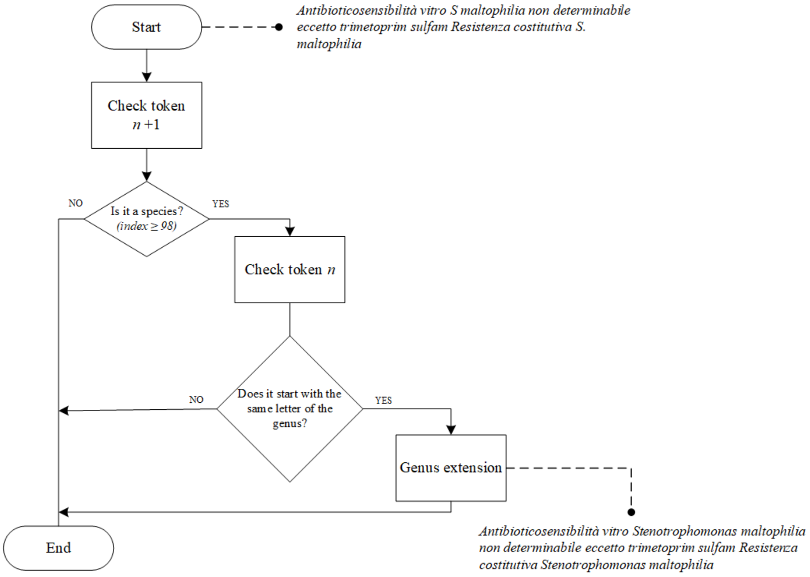

Start
Antibioticosensibilità vitro S maltophilia non determinabile eccetto trimetoprim sulfam Resistenza costitutiva S. maltophilia
Check token n+1
NO
Is it a species? *(index ≥ 98)*
YES
Check token n
NO
Does it start with the same letter of the genus?
YES
Genus extension
End
Antibioticosensibilità vitro Stenotrophomonas maltophilia non determinabile eccetto trimetoprim sulfam Resistenza costitutiva Stenotrophomonas maltophilia

Figure 6.11 **Genus extension decision flow.** The figure also includes an example. In the upper part of the figure, there is a sentence already preprocessed but before the genus extension phase, whereas, in the lower part, I can see the extended version. First, the maltophilia species is identified, while “S” as the first letter of genus Stenotrophomonas is extended.

Microorganism name extraction (IV.II): Initially, I tried to carry out a lexical and morphological analysis, but the lack of morphological structure of the clinical notes resulted in no good results. Therefore, I extracted the microorganism name by comparing each token ‘‘n’’ of the pre-processed clinical notes and the genera in the vocabulary using the

FuzzyWuzzy library. The complete workflow of the microorganism name extraction phase is shown in Fig. 6.12.

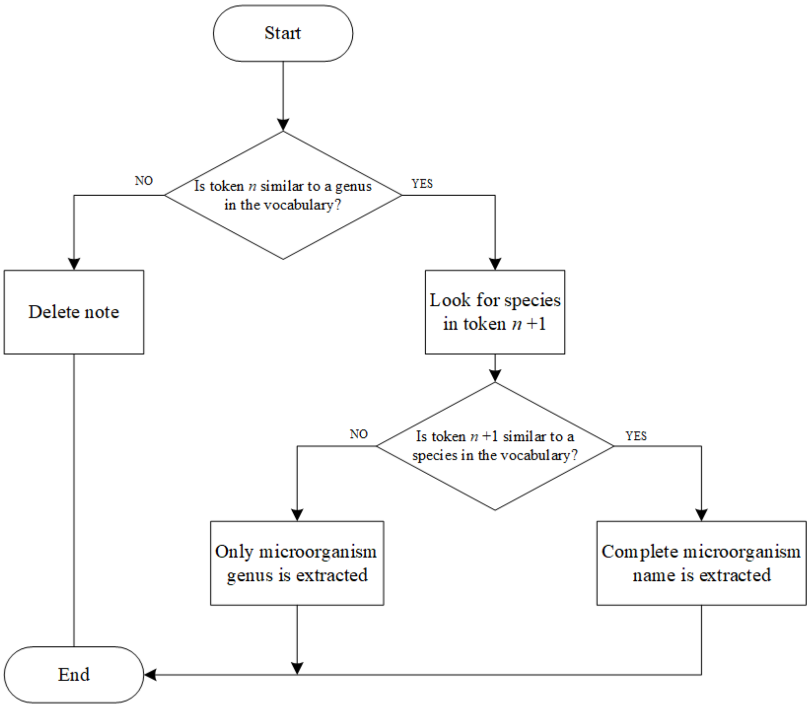

Start
Is token n similar to a genus in the vocabulary?
NO
YES
Delete note
Look for species in token n +1
Is token n +1 similar to a species in the vocabulary?
NO
YES
Only microorganism genus is extracted
Complete microorganism name is extracted
End

Figure 6.12 **Genus extension decision flow.** The figure also includes an example. In the upper part of the figure, there is a sentence already preprocessed but before the genus extension phase, whereas, in the lower part, I can see the extended version. First, the maltophilia species is identified, while “S” as the first letter of genus Stenotrophomonas is extended.

In particular, considering the genus extraction, I set the threshold of the similarity index at 75, while I set the threshold of the species index as 85 (as they were typically written correctly).

Other Information Extraction (IV.III): Together with the identification of the genus and species, in order to highlight microorganisms that could be potentially dangerous, I also searched the clinical notes for the keyword “alert”, which is an explicit indication of microbiologists regarding the danger of the identified microorganism. Similarly, but much less frequently, the “non-alert” bi-gram, with which the microbiologists indicate the harmlessness of the microorganism, may be present. To address both cases, I performed a search at the token level for the keyword “alert”. If identified at the ‘‘n’’ position, I checked if token ‘‘n - 1’’ matched the negation “non”.

6.3.4 Vocabulary building

I built a vocabulary containing the names of microorganisms (bacteria, fungi, yeasts, and viruses) from the *"National Healthcare Safety Network organism list"*, including the current taxonomic subdivision which was proposed by Carl Woase in 1990. I mapped the microorganism's genus and specie into three standard coding systems, at national and/or international level: *Italian Clinical Microbiologists Association* (AMCLI), *Systematized Nomenclature of Medicine—Clinical Terms* (SNOMED-CT), and *National Healthcare Safety Network* (NHSN). Together with the name of the microorganism, I retrieved other metadata, such as the microorganism's definitions according to the *Medical Subject Headings* (MSH) and the *National Cancer Institute* (NCI). I stored all the information in a SQL Server database, and I loaded them using the pyodbc tool.

6.3.5 Identification and Removal of a Specific Pattern

In the process of information extraction from the microbiological notes, it is useful to identify non-meaningful sentences, e.g., "integration of the provisional report of . . . ". The lack of morphological structure in the sentences led us to use a count-based method to build a numerical representation of the clinical notes. Fig. 6.13 summarizes the mean values of accuracy obtained using the three classifiers over the 20 iterations per each total number of features, shuffling the data each time.

I obtained best results in terms of mean accuracy across classifiers (99.06%) with a total number of features equal to 90. The SVM classifier with a Gaussian kernel obtained a mean accuracy of 98.99%, SLR obtained a mean accuracy of 98.99%, and random forest obtained a mean accuracy of 99.19%. This means that the pattern was correctly identified using all classifiers, and it can be securely removed from the specific clinical notes.

6.3.6 Genus Extension

Our sample of clinical notes contained a total number of 107 abbreviated genera followed by their species. After the system elaboration process of the notes, all 107 genera were extended, and they completely matched with the expert indications.

6.3.7 Microorganism Detection

Our sample of clinical notes was composed of 499 texts, and 276 (55.3%) of them actually presented the name of a microorganism. I performed two tests. First, I introduced the entire

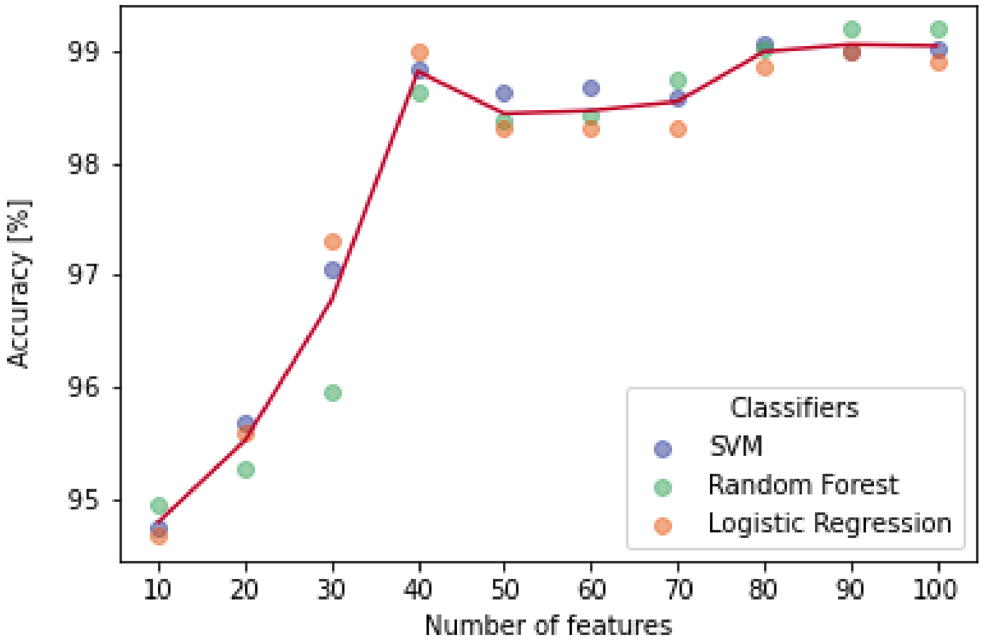


Figure 6.13 **Mean accuracy performances of the three classifiers** displayed for each value of the total number of features. Each data point is the mean value of 20 values obtained by shuffling the data.

sample into the module for microorganism extraction. Then, I introduced only those notes that actually contained the name of the microorganisms. The system correctly identified all microorganism names in both cases. In detail, it found 416 genera, and, as shown in Fig. 6.14a, the majority of them (321) had a similarity index of 100. This was also a consequence of the genus extension process.

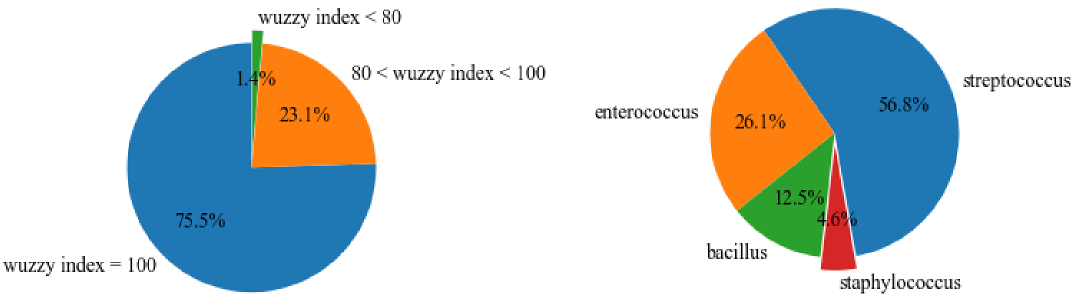


Figure 6.14 **System performance for genus extraction.** (a) Similarity index percentage distribution of genera. (b) Percentage distribution of genera found with low indices.

Fig. 6.14b shows that ‘Staphylococcus’ was the genus name with the lowest score; in particular, it usually presented a very low similarity index, between 76 and 80, if a species was not specified. Indeed, I frequently found not only the strictly scientific term, but also the Italian term in the notes, because Staphylococcus is among the most widespread bacteria and is frequently mentioned in the common discourse. This behavior affected the similarity index; in particular, Staphylococcus and ‘stafilococco/stafilococchi’ (Italian terms referring to the Staphylococcus genus) have 14 and 12 letters, respectively, within which only nine coincide, representing a Levenshtein distance of 5 (i.e., five changes are needed to transform the first word into the other). On the other hand, species never showed a Wuzzy index lower than 88. Lastly, I introduced a weight, which was a decimal parameter ranging between 0 and 1. It could be associated with the genus–species couple or only with the genus, if present alone. As the same word (genus and/or species) could be associated with more than one genus/species, this process was necessary to highlight the maximum Wuzzy indices. An example of the system output is shown in Table 1; the similarity index of the two genera Acinetobacter and Acetobacter was 92, which is quite high. Thus, in order to identify the correct genus without any doubt, I compared the following token with all species of that specific genus present in the vocabulary. If a match was found (with a Wuzzy index over 98), then I assigned to that genus–species couple a weight parameter equal to 1, while the others received a weight of 0.

Table 6.16 *Example of the system output returned when the input was “Gram negativi profilo proteomico riferibile ad A baumannii. Propensione di A baumannii alla pan-resistenza eccetto colistina (Microorganismo alert)”. The displayed columns correspond to the genus from the vocabulary, the specific word or character in text which the genus matches to, the genus Wuzzy similarity index, the species from the vocabulary, the word in text which the species matches to, the species Wuzzy similarity index, the clinical note divided into tokens, and the weight parameter.*

Genus	**Match Genus**	**Wuzzy Index Genus**	**Species**	**Match Species**	**Wuzzy Index Species**	**Clinical Notes**	**Weight**
Acinetobacter	*A*	100.0	*baumannii*	*baumannii*	100.0	[94,‘propensione‘,‘NaN‘,‘Acinetobacter‘,‘b…	1
Acetobacter	*A*	92.0	NaN	NaN	NaN	[94,‘propensione‘,‘NaN‘,‘Acinetobacter‘,‘b…	0
Aminobacter	*A*	83.0	NaN	NaN	NaN	[94,‘propensione‘,‘NaN‘,‘Acinetobacter‘,‘b…	0
Citrobacter	*A*	83.0	NaN	NaN	NaN	[94,‘propensione‘,‘NaN‘,‘Acinetobacter‘,‘b…	0

Otherwise, if the clinical note did not contain any species, and the two genera that could correspond to the same word had an identical Wuzzy index, e.g., as a consequence of a spelling error, then the algorithm would assign to both genera an equal weight of 0.5.

6.3.8 Other Information Extraction

The whole sample included 48 clinical notes that contained the keyword "alert". Our algorithm was able to correctly discriminate between the notes that contained the bi-gram "non-alert" (N = 9) and those that contained the keyword alone (N = 39).

6.3.9 Discussion

In general, the pattern recognition and the genus extension phases led to good results. The first achieved a mean accuracy value of 99.06% considering all three classifiers, while the second extracted all the names of microorganisms reported by the experts from the hospital. I should consider, however, that some ambiguities could be found during this second phase. Indeed, there are a few microorganisms with identical species and whose genera begin with the same letter. If one such case appears, then the system will duplicate the clinical note and it will extract both microorganisms, but both notes will be associated with a weight equal to 0.5. However, I should specify that, luckily, these kinds of ambiguities are quite rare. A well-known example is the intermedius species, which can belong to both the Staphylococcus genus and the Streptococcus genus. Staphylococcus intermedius is quite frequent in animals; however, it is reported as a human pathogen in very few cases, most of which are associated with exposure to animals, especially dogs. On the contrary, Streptococcus intermedius is one of the major causes of brain abscesses, but very few cases of this condition are documented annually in Italy, with an incidence that is less than 0.1% per year. Therefore, I can affirm that the probability that such ambiguity is present in the report notes of the microbiological laboratory is extremely rare. The major result of our pipeline is that I can extract a wider picture of the microorganism, because each microorganism is stored together with other metadata in the build vocabulary, such as the definition according to MeSH and its translation into national and international vocabularies. Furthermore, the pipeline also extracts the property of the microorganism under healthcare surveillance. Therefore, I can say that the system returns an object with its main characteristics. Once I accurately describe the microorganism, I can consider its identification in the clinical note as a trigger event of a series of messages and communications in accordance with the management policies of resistant microorganisms. Thus, it is possible to build a path to safeguard the patient and the community against the resistant microorganism [42]. The above-described system should be integrated in a multidisciplinary context. Correctly integrating objects from any viewpoint of the system in question requires its formal representation and management using the ISO 23903 Interoperability and Integration Reference Architecture [43]. ISO 23903 standardizes a model and framework for representing any type of system from the perspectives of the involved domains, its architectural composition/decomposition, and the related development process of implementable information and communications technology (ICT) solutions. A limitation of the presented work is the low number of samples considered due to the fact that, to be delivered to researchers outside the laboratory, all these notes were checked individually and manually in order to avoid the illicit dissemination of personal data. In the near future, the correct structuring of electronic health records (which enables in constitutive law the reuse of clinical data for the purposes of scientific research) and greater awareness of the health risk that antibiotic-resistant bacteria constitute will result in a much higher number of notes to be analyzed. The more important methodological limitations of our project and ways to overcome them are discussed in the next section.

Part IV

Discussion

The main aim of my PhD project was to first explore the usability of unstructured data in the context of differential diagnosis within those medical areas affected by a high presence of meaningful information in non-analytical sections of the EMRs. Then, I investigated the benefit of combining this kind of data with information already available in a structured format to build a more complete picture of the patient. The two use case that I worked on concern two medial areas whose research activity is affected by the high presence of text based data. The first one that I considered is Neuropsychiatry, and specifically the drug resistant epilepsy field. The second one is Infectious diseases and more in detail the field of bloodstream infections. To reach the previously mentioned objectives I combined the use of standards and NLP tools in order to create samples of high-quality data both structured and unstructured that could be used as input for ML models. Specifically, considering unstructured data, I devised an NLP-based pipeline to deal with texts written in Italian language. It is able to read them from a SQL Server database, pre-process them and create the most suitable numerical representation.

However, while creating the necessary infrastructures, I faced some of the major bottlenecks in the application of ML-based methods on clinical RWD and in general in the clinical data management for research purposes.

Among all, accurate data collection is one of the most important aspects [158, 10]. The growing interest for machine learning has led to its use in many scenarios that may not have enough labeled data. In addition to that, deep learning techniques, when involved in the process of automatic features extraction, require even a larger amount of learning data. So, the increasing usage of ML in the clinical scenario should be accompanied with a change in the way data are collected. It is a shared knowledge that a tool based on artificial intelligence is as good as the database it is trained on [145], therefore data management plays a fundamental role.

The manual collection by medical staff of the necessary data used within research projects could no more be considered the optimal solution. Thus, at least for the data already collected in a structured format into the original data source, e.g., hospital LISs, the data reuse paradigm should be the answer. As mentioned previously, it has several advantages. First of all the larger amount of samples that makes available. In effect the automatic transfer reduces the time necessary to complete the operation of data collection and the need for human intervention. Then it ensures a higher quality of the data. In fact the manual copy of the data from the original data source to another database or template sheet inevitably introduces errors [110]. While it is possible to improve the accuracy of a computerized system and minimize errors, it is more difficult to work on the level of attention of a human being.

The automatic data transfer system that I built allowed the collection of data related to thousands of episodes of patients hospitalised with bloodstream infections. This operation would have been difficult or even impossible to perform manually, as we can see by previous studies on this topic performed within the same center which involved a much smaller sample, i.e., hundreds of episodes [67, 122].

Another important aspect linked to the use ML models is that often data from multiple centers are grouped together to reach a more consistent and reliable amount of samples. However, EMRs are heterogeneous across hospitals and for this reason it is difficult to run AI systems without a deep integration process and the support offered by standards. So, the criteria that I used to build the system able to organize, transfer and store the data, are guided by the aim of making it suitable also for future collaborations. Therefore, I created a system which relies on the mandatory standard format at national level for clinical document exchange, i.e., the CDA R2. The implementation guideline "Referto di Medicina di Laboratorio", produced by HL7 Italy, identifies as mandatory the translation of the code of the laboratory exam tests in an international vocabulary. Among all the available ones, I chose LOINC according to national regulations. This will constitute the basis to allow the participation in multi-center clinical trials because it completely describes a test result and therefore makes it comparable with others.

Finally, the lack of a standardized format in patient information collected day by day, is one of the major problems in the successful application of predictive modeling to routinely collected data [19]. The presented system exploits the CDA R2 which is the same standard document format used within the National Electronic Health Records (NEHRs). So, this will enable a future re-use of the architecture to directly read data contained in the NEHRs by performing only minimal modifications.

The second main problem in information retrieval is that about the 80% of available data are in an unstructured form [39]. Dealing with this kind of data is not straightforward because a manual inspection requires a lot of time and human effort and so usually they are not used for research. However, unstructured data contain rich information that could not be find anywhere else. Therefore many research projects further investigated automatic techniques to extract meaningful information from natural language text. This is especially true for those branches of medicine where EMRs does not contain sufficient and/or suitable analytical sections to store patients data [203, 126].

In recent years, many research projects focused on the use of NLP techniques to identify and discriminate patients based on textual documents contained in their EMRs [85, 38, 42, 148].

Considering the epilepsy medical field, to the best of my knowledge, no other research project addressed the problem of localizing the EZ by automatically analyzing the semiological description of seizures, but [203] summarizes results obtained in several other text classification tasks in this medical area.

For example, considering the problem of identifying patients candidates to surgery, [32] addressed a binary text classification problem in identifying patients candidates to surgery. The study involved a sample of 200 patients, balanced between the two classes, and several text models built with different combinations of features uniquely extracted from text (unigrams, bigrams and information about drugs). The best performance, presented in terms of F1 score, is obtained with SVM in cross-validation process and it is 0.82. [198] on the contrary combined structured information, selected according to literature and experts' opinion (it was not specified if they were extracted automatically of manually from EMRs), and unigrams of the most frequent words, represented as Boolean value, extracted from different unstructured data sources. The study involved two datasets, one for pediatric patients and one for adults, both highly unbalanced. Best performance was obtained with random forest and was AUROC = 0.92. These results suggest that the combination of both structured and unstructured data may have a positive impact on the outcome, but it is highly dependent on the source of data. There are medical institutions where data are only stored in the form of free text, so it is not easy to extract the specific information used in [198]. As mentioned above, manual collection of data, especially when dealing with a great number of patients can not be considered an option.

Considering the HAIs field, [161] performed a sentence classification task to automatically identify Catheter-Related Events from clinical notes in Norwegian language. The final sample was composed by 730 notes and five possible classes: *None*, no CVC use is mentioned (predominant class); *Plan*, an insertion is planned within the note; *Ins*, the note mentions the insertion of a CVC; *Use*, the note mentions for example the use of CVC for care or the planned removal, which implies in turn its presence; *Rem*, the note mentions the CVC removal. Texts were converted into numerical representations using TfidfVectorizer of Sklearn. The SVM with linear kernel and different regularization methods were considered. Performances were evaluated in terms of F1 score (multiplied by 100). Best performances for each class were: 99.9 for *None*, 51.2 for *Plan*, 27.5 for *Ins*, 79.5 for *Use* and 39.1 for *Rem*.

The computational pipeline that I devised combines the use of ML and NLP-based methods to analyze text written in Italian language, automatically extract the most relevant features and assign a binary label.

Considering the first use case, i.e., the localization of the EZ in DRE patients, its performances, in terms of weighted F1 score, are up to 0.853 in a cross-validation on the training set and 0.781 on a blind testing set. These results show that NLP-based numerical representation combined with ML-based classification models may help in localizing the origin of the seizures relying only on seizures-related semeiological text data. Accurate early recognition of EZ could enable a more appropriate patient management and a faster access to epilepsy surgery to potential candidates. That is a viable option with focal onset drug-resistant patient but a delayed or incorrect diagnosis of EZ location severely limits its efficacy.

Considering the second use case, i.e., the identification of presence and removal of the CVC in critically ill patients hospitalized in ICUs, the best performance, in a cross-validation on the training set, in terms of weighted F1 score is 0.98 in the identification of CVC presence while is 0.92 in CVC removal (classification at patient level). On the blind test set, the best performance on the first task is 0.916 and on the second one is 0.837. These results constitute a step towards the automatic calculation of a very important score for HAIs management, *i.e.,* the EQUAL Candida Score, which can be used to provide a real time feedback on clinical practice and patient management. It could also improve the adherence to international clinical guidelines and, as a consequence, patients' survival [80].

Within the pipeline, two methods to produce the numerical representation of text were considered, and one or both of them are used according to text characteristics. Specifically, the Bag of Words technique, which is based on a measure of pattern frequency, was used with short texts missing a complete grammar scheme. Tokens based on combinations of characters were used in order to matched sub-words and therefore reduce the impact of misspellings on the extraction of most frequent patterns. Then, combination of exact words were also included to keep some context, that otherwise using the Bag of Words technique would be lost. On the contrary, Word Embedding model, tuned on the available training sample, was used only with complete and longer texts.

Then, the pipeline included two steps aimed at manipulating text in order to remove uninformative patterns and on the contrary highlight important ones. Specifically, in the use case involving texts with a complete grammar structure, *i.e.,* the epilepsy use case, the pipeline included a lemmatization step. It aimed at reducing the variability of the tokens by converting each word in its base form and therefore support the generation of text representation with both techniques. First because the frequency of the specific lemma increased, highlighting its eventual importance in features extraction with count-based methods, and then, considering the model build with Word2Vec, multiple contexts containing the same word supported the insertion of the word vector in the multidimensional space. I think that it is especially useful for languages such as Italian, because, for example, adjectives can appear in four forms, i.e., feminine or masculine and singular or plural, and they should not be considered as different features. However, considering the analysis performed in the second use case and the errors that occurred because of different way of mentioning CVC, further investigations on the impact of a normalization process suitable for short text not following grammars will be necessary [167]. While considering both use cases the step of stop-words removal was included as the count-based model was generated and it automatically selects the words with higher frequency.

Finally, the problem was addressed as a text classification task because a simple keywords extraction was considered not effective, *e.g.,* some keywords together may have a meaning, but if other words are present the meaning may change. An example belonging to the Infectious Diseases use case is: "Sostituita medicazione CVC" (which sounds like *“replaced the CVC dressing"*) and "Sostituzione CVC" (which sounds like *“CVC replacement”*), only the second one means that CVC has been removed. So, the label should be assigned on a combination of elements that are present in the text.

One of the main advantages of this approach is that any a-priori knowledge was used in the features engineering process, so the pipeline is suitable for different use cases.

While, the main disadvantages relate to the building phase of the word embedding model and the features selection process. First of all, they highly depend on the data that they are trained on. This means, that they also depend on the specific way of writing of medical staff present in the center, which could contain hidden pattern related to shared knowledge and rules within the center. In addition to that, the limited amount of available data has an impact on the performance as well, because it implies a limited vocabulary of known words. This problem is partially addressed by the model built with the Bag of Words technique because it combines together with exact words also patterns of characters that could at least partly match unknown words. On the contrary the word embedding model assigns a vector of zeros to the words not in the vocabulary. So, those words have no impact on the final numerical representation of text because it performed as mean of words vectors or applying the tf-idf formula to them. Another issue is that both NLP approaches included in the pipeline do not take into account the problem of word sense disambiguation, which means that one word may have different meanings based on the context. A possible solution could be to include additional information while building the numerical representation of words, such as distinguishing words based on their role in the sentence and therefore using Part of Speech tagging techniques.

While, to address the problem of limited number of samples which implies limited vocabulary, a larger amount of text should be collected. However, in this context is still not possible to design an automatic transfer system as, differently from structured data, it is not easy to remove sensitive information. First because data are not organized in columns, that could be easily included or excluded, and second because a manual pseudonymization or de-identification would require many resources. To overcome this problem some research projects focused on the building of an automatic de-identification system, but this, in turn, is linked with language and also privacy issues because no stand alone open sources tools specific for Italian language is available.

A possible solution could be to move towards deep learning based architectures such as the *Bidirectional Encoder Representations from Transformers* (BERT) [141, 91]. It belongs to the family of masked-language models and it first appeared in 2018 when it was published by researches at Google [45]. The main advantages are that models pre-trained on Italian language are available, unknown words are automatically addressed by splitting them into tokens of known sub-words, and each word can be represented by more than one vector depending on the context. In addition to that, BERT-based models could also address the problem of word sense disambiguation as each word is mapped with more than one vector within the model and the most suitable one is used based on the word context in the text. However, there are also some issues linked to the use of this kind of architectures. Among all, the size of the vectors is fixed to 512 tokens and this means that their usage is not straightforward with long texts. Then, as they are trained on texts that are not specific, they require a fine-tuning phase on data belonging to the research context. However, this could raise again the problem mentioned above of dependence to specific hospitals and clinicians. For this reason, further investigations on the benefit of involving texts from more than one center will be required.

Finally, another advantage of the presented approach, that became even more evident while going forward in the project, is the use of a database. It has been used not only to deal with structured data but also to support the management of the unstructured ones. A considerable effort was necessary at the beginning of the work as unstructured data could not be automatically collected in a ready to use way and they required an additional preprocessing phase. However, once that unstructured data were contained into a fixed database structure, their usage was decidedly simplified. For example, in the epilepsy use case, the storing of text belonging to the different sections of the EMRs (such as anamnestic information, seizures descriptions and conclusions) into different datatables, supported the training phase of word embedding model. In effect, it was easy to include or exclude sections, while it could be a heavy operation to perform manually if data where only stored in text files.

Another important aspect is that the storing of unstructured data into a database allowed to link them together with the structured ones. Therefore, a more complete picture of the patient was build and this allowed further investigations.

In conclusion, the presented approach constitutes a solid basis in creating a more complete picture of the patient useful to conduct other further analysis and to improve the differential diagnosis process. The results highlight the potential of using the unstructured data. First of all, as it involves clinical RWD already stored into the EMRs, it is not expensive in terms of resources necessary for data collection and it is reproducible with new patients. Then, the devised computational pipeline is easily applicable on different kind of texts and it showed good results on both use cases.

Finally, I suggest to further investigate automatic de-identification techniques to allow the automatic retrieval of unstructured data [15, 25]. This will further enhance research because it enables the retrieval of larger samples of data combined from different data sources. Then, since I analyzed the importance of using standards both for collecting and mapping structured data, another future work that I consider worth investigating is develop a system able to address the same problem but processing unstructured data. For example, to guarantee the semantic interoperability in the classification process, e.g. for future studies with more than one centre involved, it could be important to develop a system able to automatically map each seizure description and clinical note with the following main standards used in literature [86, 31, 154]: *International Classification of Diseases 9th and 10th edition* (ICD9 and ICD10); ILAE codes [168]. At last, it should also be noted that terminology standards evolve over time, so in order to guarantee the correct management of the terminology mappings as new versions will be release, the approach described in [64] could be considered.

References

[1] Steven P Abney. “Parsing by chunks”. In: *Principle-based parsing: Computation and Psycholinguistics* (1992), pp. 257–278.

[2] JY Adamu et al. “Antimicrobial susceptibility testing of Staphylococcus aureus isolated from apparently healthy humans and animals in Maiduguri, Nigeria”. In: *International Journal of Biomedical and Health Sciences* 6.4 (2021).

[3] David Ahmedt-Aristizabal et al. “Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey”. In: *Epilepsia* 58.11 (2017), pp. 1817–1831.

[4] Mehreen Ali and Tero Aittokallio. “Machine learning and feature selection for drug response prediction in precision oncology applications”. In: *Biophysical reviews* 11.1 (2019), pp. 31–39.

[5] Ali Alim-Marvasti et al. “Machine learning for localizing epileptogenic-zone in the temporal lobe: Quantifying the value of multimodal clinical-semiology and imaging concordance”. In: *Frontiers in Digital Health* 3 (2021), p. 559103.

[6] Raymond D Aller. “Software standards and the laboratory information system.” In: *American journal of clinical pathology* 105.4 Suppl 1 (1996), S48–53.

[7] Eman Yahia Alqaissi, Fahd Saleh Alotaibi, and Muhammad Sher Ramzan. “Modern machine-learning predictive models for diagnosing infectious diseases”. In: *Computational and Mathematical Methods in Medicine* 2022 (2022).

[8] Ashwin N Ananthakrishnan et al. “Improving case definition of Crohn’s disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach”. In: *Inflammatory bowel diseases* 19.7 (2013), pp. 1411–1420.

[9] World Medical Association. “World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects”. In: *JAMA* 310.20 (Nov. 2013), pp. 2191–2194. ISSN: 0098-7484. DOI: 10.1001/jama.2013.281053. eprint: https://jamanetwork.com/journals/jama/articlepdf/1760318/jsc130006.pdf. URL: https://doi.org/10.1001/jama.2013.281053.

[10] Stephen H Bach et al. “Learning the structure of generative models without labeled data”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 273–282.

[11] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. “Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 238–247.

[12] Silpi Basak, Priyanka Singh, and Monali Rajurkar. “Multidrug resistant and extensively drug resistant bacteria: a study”. In: *Journal of pathogens* 2016 (2016).

[13] Sally L Baxter et al. “Text processing for detection of fungal ocular involvement in critical care patients: cross-sectional study”. In: *Journal of Medical Internet Research* 22.8 (2020), e18855.

[14] Sándor Beniczky et al. “A web-based algorithm to rapidly classify seizures for the purpose of drug selection”. In: *Epilepsia* 62.10 (2021), pp. 2474–2484.

[15] Hanna Berg et al. “De-identification of Clinical Text for Secondary Use: Research Issues.” In: *HEALTHINF* (2021), pp. 592–599.

[16] Eta S Berner and Mark L Graber. “Overconfidence as a cause of diagnostic error in medicine”. In: *The American journal of medicine* 121.5 (2008), S2–S23.

[17] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O’Reilly Media, Inc.", 2009.

[18] J Martin Bland and Douglas G Altman. “Multiple significance tests: the Bonferroni method”. In: *Bmj* 310.6973 (1995), p. 170.

[19] Rogério Blitz et al. “Design and implementation of an informatics infrastructure for standardized data acquisition, transfer, storage, and export in psychiatric clinical routine: Feasibility study”. In: *JMIR Mental Health* 8.6 (2021), e26681.

[20] Emily S Brouwer et al. “Leveraging unstructured data to identify hereditary angioedema patients in electronic medical records”. In: *Allergy, Asthma & Clinical Immunology* 17.1 (2021), pp. 1–10.

[21] Michael Buckland and Fredric Gey. “The relationship between recall and precision”. In: *Journal of the American society for information science* 45.1 (1994), pp. 12–19.

[22] Rich Caruana and Dayne Freitag. “Greedy attribute selection”. In: *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 28–36.

[23] Sara Casciato et al. “Knowledge and attitudes of neurologists toward epilepsy surgery: an Italian survey”. In: *Neurological Sciences* 43.7 (2022), pp. 4453–4461.

[24] Gloria Castellazzi et al. “A machine learning approach for the differential diagnosis of alzheimer and vascular dementia fed by MRI selected features”. In: *Frontiers in neuroinformatics* 14 (2020), p. 25.

[25] Rosario Catelli et al. “Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set”. In: *Applied soft computing* 97 (2020), p. 106779.

[26] Chris Cheadle et al. “Application of z-score transformation to Affymetrix data.” In: *Applied bioinformatics* 2.4 (2003), pp. 209–217.

[27] Liang Chen et al. “Using natural language processing to extract clinically useful information from Chinese electronic medical records”. In: *International journal of medical informatics* 124 (2019), pp. 6–12.

[28] Davide Chicco and Giuseppe Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC genomics* 21 (2020), pp. 1–13.

[29] Billy Chiu et al. “How to train good word embeddings for biomedical NLP”. In: *Proceedings of the 15th workshop on biomedical natural language processing*. 2016, pp. 166–174.

[30] Noam Chomsky. “7. The Logical Basis of Linguistic Theory”. In: *Eight decades of general linguistics*. Brill, 2013, pp. 123–236.

[31] Taridzo Chomutare, Andrius Budrionis, and Hercules Dalianis. “Combining deep learning and fuzzy logic to predict rare ICD-10 codes from clinical notes”. In: *2022 IEEE International Conference on Digital Health (ICDH)*. IEEE. 2022, pp. 163–168.

[32] Kevin Bretonnel Cohen et al. “Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning”. In: *Biomedical informatics insights* 8 (2016), BII–S38308.

[33] Jill Collis and Roger Hussey. “Business research: A practical guide for undergraduate and postgraduate students”. In: (2003).

[34] Ramona Cordani et al. “Sleep disturbances in craniopharyngioma: a challenging diagnosis”. In: *Journal of neurology* 268 (2021), pp. 4362–4369.

[35] Luca Corradi et al. “A repository based on a dynamically extensible data model supporting multidisciplinary research in neuroscience”. In: *BMC medical informatics and decision making* 12.1 (2012), pp. 1–14.

[36] Martin R Cowie et al. “Electronic health records to facilitate clinical research”. In: *Clinical Research in Cardiology* 106 (2017), pp. 1–9.

[37] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge university press, 2000.

[38] Licong Cui et al. “EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification”. In: *AMIA Annual Symposium Proceedings*. Vol. 2012. American Medical Informatics Association. 2012, p. 1191.

[39] Hercules Dalianis. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature, 2018.

[40] Hercules Dalianis and Sumithra Velupillai. “De-identifying Swedish clinical text-refinement of a gold standard and experiments with Conditional random fields”. In: *Journal of biomedical semantics* 1 (2010), pp. 1–10.

[41] Marc Damashek. “Gauging similarity with n-grams: Language-independent categorization of text”. In: *Science* 267.5199 (1995), pp. 843–848.

[42] Surabhi Datta, Elmer V Bernstam, and Kirk Roberts. “A frame semantic overview of NLP-based information extraction for cancer-related EHR notes”. In: *Journal of biomedical informatics* 100 (2019), p. 103301.

[43] Bonnie B Dean et al. “Use of electronic medical records for health outcomes research: a literature review”. In: *Medical Care Research and Review* 66.6 (2009), pp. 611–638.

[44] Danielle D DeSouza et al. “Natural language processing as an emerging tool to detect late-life depression”. In: *Frontiers in Psychiatry* 12 (2021), p. 719125.

[45] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[46] Thomas G Dietterich. “Approximate statistical tests for comparing supervised classification learning algorithms”. In: *Neural computation* 10.7 (1998), pp. 1895–1923.

[47] Thomas G Dietterich. “Ensemble methods in machine learning”. In: *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*. Springer. 2000, pp. 1–15.

[48] Wouter B van Dijk et al. “Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study”. In: *Journal of Clinical Epidemiology* 132 (2021), pp. 97–105.

[49] Wilfrid J Dixon and Frank J Massey Jr. *Introduction to statistical analysis.* McGraw-Hill, 1951.

[50] Son Doan et al. “Building a natural language processing tool to identify patients with high clinical suspicion for Kawasaki disease from emergency department notes”. In: *Academic Emergency Medicine* 23.5 (2016), pp. 628–636.

[51] Sandrine Dudoit et al. “Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments”. In: *Statistica sinica* (2002), pp. 111–139.

[52] Khaled El Emam et al. “The use of electronic data capture tools in clinical trials: Web-survey of 259 Canadian trials”. In: *Journal of medical Internet research* 11.1 (2009), e8.

[53] Dominique Estival, Chris Nowak, and Andrew Zschorn. “Towards ontology-based natural language processing”. In: *Proceeedings of the Workshop on NLP and XML (NLPXML-2004): RDF/RDFS and OWL in Language Technology*. 2004, pp. 59–66.

[54] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. “Regularization networks and support vector machines”. In: *Advances in computational mathematics* 13 (2000), pp. 1–50.

[55] T Fahey, S Griffiths, and TJ Peters. “Evidence based purchasing: understanding results of clinical trials and systematic reviews”. In: *Bmj* 311.7012 (1995), pp. 1056–1059.

[56] Ruth Farmer et al. “Promises and pitfalls of electronic health record analysis”. In: *Diabetologia* 61 (2018), pp. 1241–1248.

[57] Beata Fonferko-Shadrach et al. “Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system”. In: *BMJ open* 9.4 (2019), e023232.

[58] Valeria Fonti and Eduard Belitser. “Feature selection using lasso”. In: *VU Amsterdam research paper in business analytics* 30 (2017), pp. 1–25.

[59] Elizabeth Ford et al. “Extracting information from the text of electronic medical records to improve case detection: a systematic review”. In: *Journal of the American Medical Informatics Association* 23.5 (2016), pp. 1007–1015.

[60] Jeffrey EF Friedl. *Mastering regular expressions*. O’Reilly Media, Inc., 2006.

[61] Milton Friedman. “The use of ranks to avoid the assumption of normality implicit in the analysis of variance”. In: *Journal of the american statistical association* 32.200 (1937), pp. 675–701.

[62] Sunyang Fu et al. “Automated detection of periprosthetic joint infections and data elements using natural language processing”. In: *The Journal of arthroplasty* 36.2 (2021), pp. 688–692.

[63] William D Gaillard et al. “Establishing criteria for pediatric epilepsy surgery center levels of care: report from the ILAE Pediatric Epilepsy Surgery Task Force”. In: *Epilepsia* 61.12 (2020), pp. 2629–2642.

[64] Roberta Gazzarata et al. “A terminology service compliant to CTS2 to manage semantics within the regional HIE”. In: *European Journal of Biomedical Informatics* 13.1 (2017).

[65] Robin E Gearing et al. “A methodology for conducting retrospective chart review research in child and adolescent psychiatry”. In: *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 15.3 (2006), p. 126.

[66] P Genton et al. “Do carbamazepine and phenytoin aggravate juvenile myoclonic epilepsy?” In: *Neurology* 55.8 (2000), pp. 1106–1109.

[67] Daniele Roberto Giacobbe et al. “Combined use of serum (1, 3)-β-D-glucan and procalcitonin for the early differential diagnosis between candidaemia and bacteraemia in intensive care units”. In: *Critical Care* 21 (2017), pp. 1–9.

[68] Daniele Roberto Giacobbe et al. “Desirability of outcome ranking (DOOR) for comparing diagnostic tools and early therapeutic choices in patients with suspected candidemia”. In: *European Journal of Clinical Microbiology & Infectious Diseases* 38 (2019), pp. 413–417.

[69] Daniele Roberto Giacobbe et al. “Early detection of sepsis with machine learning techniques: a brief clinical perspective”. In: *Frontiers in medicine* 8 (2021), p. 617486.

[70] Daniele Roberto Giacobbe et al. “Validation of an Automated System for the Extraction of a Wide Dataset for Clinical Studies Aimed at Improving the Early Diagnosis of Candidemia”. In: *Diagnostics* 13.5 (2023), p. 961.

[71] Mauro Giacomini et al. “Data modeling for tools and technologies for the analysis and synthesis of NANOstructures (TASNANO) project”. In: *Journal of Information Technology Research (JITR)* 2.3 (2009), pp. 49–70.

[72] Barbara Giannini et al. “From Liguria HIV Web to Liguria infectious diseases network: how a digital platform improved doctors’ work and patients’ care”. In: *AIDS research and human retroviruses* 34.3 (2018), pp. 239–240.

[73] William Grimson et al. “Specifying an open clinical laboratory information system”. In: *Computer methods and programs in biomedicine* 50.2 (1996), pp. 95–109.

[74] Thomas R Gruber. “Toward principles for the design of ontologies used for knowledge sharing?” In: *International journal of human-computer studies* 43.5-6 (1995), pp. 907–928.

[75] Kenric W Hammond et al. “Are electronic medical records trustworthy? Observations on copying, pasting and duplication”. In: *AMIA Annual Symposium Proceedings*. Vol. 2003. American Medical Informatics Association. 2003, p. 269.

[76] Zellig S Harris. “Distributional structure”. In: *Word* 10.2-3 (1954), pp. 146–162.

[77] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[78] Werner Heisenberg et al. *Physics and beyond*. Allen & Unwin London, 1971.

[79] Julia Hirschberg and Christopher D Manning. "Advances in natural language processing". In: *Science* 349.6245 (2015), pp. 261–266.

[80] Martin Hoenigl et al. "Guideline adherence and survival of patients with candidaemia in Europe: results from the ECMM Candida III multinational European observational cohort study". In: *The Lancet Infectious Diseases* (2023).

[81] George Hripcsak and Adam S Rothschild. "Agreement, the f-measure, and reliability in information retrieval". In: *Journal of the American medical informatics association* 12.3 (2005), pp. 296–298.

[82] Lu-Chou Huang et al. "Privacy preservation and information security protection for patients' portable electronic health records". In: *Computers in Biology and Medicine* 39.9 (2009), pp. 743–750.

[83] Geert Huys et al. "Intra-and interlaboratory performance of antibiotic disk-diffusion-susceptibility testing of bacterial control strains of relevance for monitoring aquaculture environments". In: *Diseases of aquatic organisms* 66.3 (2005), pp. 197–204.

[84] Massimiliano Izzo et al. "A digital repository with an extensible data model for biobanking and genomic analysis management". In: *BMC genomics* 15 (2014), pp. 1–15.

[85] Olof Jacobson and Hercules Dalianis. "Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections". In: *Proceedings of the 15th workshop on biomedical natural language processing*. 2016, pp. 191–195.

[86] Nathalie Jette et al. "ICD coding for epilepsy: past, present, and future—a report by the International League Against Epilepsy Task Force on ICD codes in epilepsy". In: *Epilepsia* 56.3 (2015), pp. 348–355.

[87] Li Jiyun, Wang Junping, and Pei Hongxing. "Data cleaning of medical data for knowledge mining". In: *Journal of Networks* 8.11 (2013), p. 2663.

[88] Thorsten Joachims. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* Tech. rep. Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[89] Paul Johannesson and Erik Perjons. *An introduction to design science*. Vol. 10. Springer, 2014.

[90] Melissa D Johnson et al. "Core recommendations for antifungal stewardship: a statement of the mycoses study group education and research consortium". In: *The Journal of infectious diseases* 222.Supplement_3 (2020), S175–S198.

[91] Katikapalli Subramanyam Kalyan and S. Sangeetha. "SECNLP: A survey of embeddings in clinical natural language processing". In: *Journal of Biomedical Informatics* 101 (2020), p. 103323. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2019.103323. URL: https://www.sciencedirect.com/science/article/pii/S1532046419302436.

[92] Kerstin A Kessel and Stephanie E Combs. "Review of developments in electronic, clinical data collection, and documentation systems over the last decade–are we ready for big data in routine health care?" In: *Frontiers in oncology* 6 (2016), p. 75.

[93] Apeksha Khabia and MB Chandak. “A cluster based approach with n-grams at word level for document classification”. In: *International Journal of Computer Applications* 117.23 (2015).

[94] Divya Khyani et al. “An Interpretation of Lemmatization and Stemming in Natural Language Processing”. In: *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology* 22 (2020), pp. 350–357.

[95] Dong Wook Kim et al. “Localization value of seizure semiology analyzed by the conditional inference tree method”. In: *Epilepsy Research* 115 (2015), pp. 81–87.

[96] David G Kleinbaum et al. *Logistic regression*. New York: Springer, 2002.

[97] David Michael Kleinberg-Levin. *The opening of vision: Nihilism and the postmodern situation*. Taylor & Francis, 1988.

[98] Teia Kobulashvili et al. “Diagnostic and prognostic value of noninvasive long-term video-electroencephalographic monitoring in epilepsy surgery: A systematic review and meta-analysis from the E-PILEPSY consortium”. In: *Epilepsia* 59.12 (2018), pp. 2272–2283.

[99] Kamran Kowsari et al. “Text classification algorithms: A survey”. In: *Information* 10.4 (2019), p. 150.

[100] Marcin Kozak et al. “The effects of data input errors on subsequent statistical inference”. In: *Journal of Applied Statistics* 42.9 (2015), pp. 2030–2037.

[101] William H Kruskal and W Allen Wallis. “Use of ranks in one-criterion variance analysis”. In: *Journal of the American statistical Association* 47.260 (1952), pp. 583–621.

[102] Rebecca Daniels Kush et al. “FAIR data sharing: the roles of common data elements and harmonization”. In: *Journal of Biomedical Informatics* 107 (2020), p. 103421.

[103] Clete A Kushida et al. “Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies”. In: *Medical care* 50.Suppl (2012), S82.

[104] Kenneth H Lai et al. “Automated misspelling detection and correction in clinical free-text records”. In: *Journal of biomedical informatics* 55 (2015), pp. 188–195.

[105] PM Lavanya and E Sasikala. “Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey”. In: *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*. IEEE. 2021, pp. 603–609.

[106] Breiman Leo. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[107] Vladimir I Levenshtein et al. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. Soviet Union. 1966, pp. 707–710.

[108] Irene Li et al. “Neural Natural Language Processing for unstructured data in electronic health records: A review”. In: *Computer Science Review* 46 (2022), p. 100511.

[109] Xiaozheng Li et al. “Intelligent diagnosis with Chinese electronic medical records based on convolutional neural networks”. In: *BMC bioinformatics* 20 (2019), pp. 1–12.

[110] Jiabin Liu et al. “Automatic data acquisition for deep learning”. In: *Proceedings of the VLDB Endowment* 14.12 (2021), pp. 2739–2742.

[111] Tiqing Liu et al. “BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities”. In: *Nucleic acids research* 35.suppl_1 (2007), pp. D198–D201.

[112] Tommaso Lo Barco et al. “Improving early diagnosis of rare diseases using Natural Language Processing in unstructured medical records: an illustration from Dravet syndrome”. In: *Orphanet Journal of Rare Diseases* 16.1 (2021), pp. 1–12.

[113] Thorvardur Jon Love, Tianxi Cai, and Elizabeth W Karlson. “Validation of psoriatic arthritis diagnoses in electronic medical records using natural language processing”. In: *Seminars in arthritis and rheumatism*. Vol. 40. 5. Elsevier. 2011, pp. 413–420.

[114] A-P Magiorakos et al. “Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance”. In: *Clinical microbiology and infection* 18.3 (2012), pp. 268–281.

[115] R Malarvizhi and Antony Selvadoss Thanamani. “K-nearest neighbor in missing data imputation”. In: *Int. J. Eng. Res. Dev* 5.1 (2012), pp. 5–7.

[116] Devin M Mann et al. “Predictors of adherence to diabetes medications: the role of disease and medication beliefs”. In: *Journal of behavioral medicine* 32 (2009), pp. 278–284.

[117] Henry B Mann and Donald R Whitney. “On a test of whether one of two random variables is stochastically larger than the other”. In: *The annals of mathematical statistics* (1947), pp. 50–60.

[118] Christopher D Manning. *An introduction to information retrieval*. Cambridge university press, 2009.

[119] Frank J Massey Jr. “The Kolmogorov-Smirnov test for goodness of fit”. In: *Journal of the American statistical Association* 46.253 (1951), pp. 68–78.

[120] Sibylle C Mellinghoff et al. “EQUAL Candida Score: An ECMM score derived from current guidelines to measure QUAlity of Clinical Candidaemia Management”. In: *Mycoses* 61.5 (2018), pp. 326–330.

[121] Stéphane M Meystre et al. “Extracting information from textual documents in the electronic health record: a review of recent research”. In: *Yearbook of medical informatics* 17.01 (2008), pp. 128–144.

[122] Malgorzata Mikulska et al. “Sensitivity of Serum Beta-D-Glucan in Candidemia According to Candida Species Epidemiology in Critically Ill Patients Admitted to the Intensive Care Unit”. In: *Journal of Fungi* 8.9 (2022), p. 921.

[123] Malgorzata Mikulska et al. “Tocilizumab and steroid treatment in patients with COVID-19 pneumonia”. In: *Plos one* 15.8 (2020), e0237831.

[124] Nor Hamizah Miswan, Chee Seng Chan, and Chong Guan Ng. “Hospital readmission prediction based on improved feature selection using grey relational analysis and LASSO”. In: *Grey Systems: Theory and Application* (2021).

[125] Marie-Francine Moens. *Information extraction: algorithms and prospects in a retrieval context*. Vol. 1. Springer, 2006.

[126] Sara Mora et al. “A NLP Pipeline for the Automatic Extraction of a Complete Microorganism’s Picture from Microbiological Notes”. In: *Journal of Personalized Medicine* 12.9 (2022), p. 1424.

[127] Sara Mora et al. “A wide database for future studies aimed at improving early recognition of Candidemia”. In: *Public Health and Informatics*. IOS Press, 2021, pp. 1081–1082.

[128] Sara Mora et al. “Ten Years of Medical Informatics and Standards Support for Clinical Research in an Infectious Diseases Network”. In: *Applied Clinical Informatics* 14.01 (2023), pp. 16–27.

[129] George J Mouly. “Science of educational research”. In: (1970).

[130] Lynne Murray et al. “Controlled trial of the short-and long-term effect of psychological treatment of post-partum depression: 2. Impact on the mother-child relationship and child outcome”. In: *The British Journal of Psychiatry* 182.5 (2003), pp. 420–427.

[131] Maureen A Murtaugh et al. “Regular expression-based learning to extract bodyweight values from clinical notes”. In: *Journal of biomedical informatics* 54 (2015), pp. 186–190.

[132] Monika A Myszczynska et al. “Applications of machine learning to diagnosis and treatment of neurodegenerative diseases”. In: *Nature Reviews Neurology* 16.8 (2020), pp. 440–456.

[133] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. “Natural language processing: an introduction”. In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 544–551.

[134] Organisation for Economic Co-operation OECD, European Centre for Disease Prevention Development, and Control. *Antimicrobial Resistance—Tackling the Burden in the European Union—Briefing Note for EU/ EEA Countries*. 2019. URL: https://www.oecd.org/health/health-systems/AMR-Tackling-the-Burden-in-the-EU-OECD-ECDC-Briefing-Note-2019.Pdf (visited on 08/25/2022).

[135] Angel Ois et al. “Misdiagnosis worsens prognosis in subarachnoid hemorrhage with good Hunt and Hess score”. In: *Stroke* 50.11 (2019), pp. 3072–3076.

[136] Erin M Okazaki et al. “Usage of EpiFinder clinical decision support in the assessment of epilepsy”. In: *Epilepsy & Behavior* 82 (2018), pp. 140–143.

[137] World Health Organization. *Epilepsy: a public health imperative*. Geneva: World Health Organization, 2019, 146 p.The Chinese version is published by China Association Against Epilepsy.

[138] Satish Patel and James Lyons-Weiler. “caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer”. In: *Applied bioinformatics* 3 (2004), pp. 49–62.

[139] Ivan Pavlovic and I Lazarevic. “Reshaping Clinical Trial Data Collection Process to Use the Advantages of the Web-Based Electronic Data Collection”. In: *11th Mediterranean Conference on Medical and Biomedical Engineering and Computing 2007: MEDICON 2007, 26-30 June 2007, Ljubljana, Slovenia*. Springer. 2007, pp. 741–744.

[140] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[141] Yifan Peng, Shankai Yan, and Zhiyong Lu. “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets”. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 58–65. DOI: 10.18653/v1/W19-5006. URL: https://aclanthology.org/W19-5006.

[142] Janet FE Penz, Adam B Wilcox, and John F Hurdle. “Automated identification of adverse events related to central venous catheters”. In: *Journal of biomedical informatics* 40.2 (2007), pp. 174–182.

[143] Bethany Percha. “Modern clinical text mining: a guide and review”. In: *Annual review of biomedical data science* 4 (2021), pp. 165–187.

[144] Luıs Pereira et al. “Using text mining to diagnose and classify epilepsy in children”. In: *2013 IEEE 15th international conference on E-health networking, applications and services (Healthcom 2013)*. IEEE. 2013, pp. 345–349.

[145] Neoklis Polyzotis et al. “Data lifecycle challenges in production machine learning: a survey”. In: *ACM SIGMOD Record* 47.2 (2018), pp. 17–28.

[146] Katrina K Poppe et al. “Developing and validating a cardiovascular risk score for patients in the community with prior cardiovascular disease”. In: *Heart* 103.12 (2017), pp. 891–892.

[147] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O’Reilly Media, Inc.", 2012.

[148] Sara Rabhi, Jérémie Jakubowicz, and Marie-Helene Metzger. “Deep learning versus conventional machine learning for detection of healthcare-associated infections in French clinical narratives”. In: *Methods of information in medicine* 58.01 (2019), pp. 031–041.

[149] Prabhakar Raghavan, S Amer-Yahia, and L Gravano. “Structure in text: Extraction and exploitation”. In: *Proceeding of the 7th international Workshop on the Web and Databases (WebDB), ACM SIGMOD/PODS*. Vol. 1. 2004.

[150] Priya Ranganathan and Rakesh Aggarwal. “Study designs: Part 1–An overview and classification”. In: *Perspectives in clinical research* 9.4 (2018), p. 184.

[151] Riina Rautemaa-Richardson et al. “Impact of a diagnostics-driven antifungal stewardship programme in a UK tertiary referral teaching hospital”. In: *Journal of Antimicrobial Chemotherapy* 73.12 (2018), pp. 3488–3495.

[152] Protection Regulation. “DECRETO LEGISLATIVO 10 agosto 2018, n. 101”. In: *Regulation (ITA)* 101 (2018). URL: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2018;101.

[153] Protection Regulation. “Regulation (EU) 2016/679 of the European Parliament and of the Council”. In: *Regulation (EU)* 679 (2016).

[154] Sonja Remmer, Anastasios Lamproudis, and Hercules Dalianis. “Multi-label diagnosis classification of Swedish discharge summaries–ICD-10 code assignment using KB-BERT”. In: *International Conference Recent Advances in Natural Language Processing (RANLP’21), online, September 1-3, 2021*. INCOMA Ltd. 2021, pp. 1158–1166.

[155] Angela Revelas. “Healthcare–associated infections: A public health problem”. In: *Nigerian medical journal: journal of the Nigeria Medical Association* 53.2 (2012), p. 59.

[156] Andrea Ripoli et al. “Personalized machine learning approach to predict candidemia in medical wards”. In: *Infection* 48 (2020), pp. 749–759.

[157] Erik Roelofs et al. “Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial”. In: *Radiotherapy and Oncology* 108.1 (2013), pp. 174–179.

[158] Yuji Roh, Geon Heo, and Steven Euijong Whang. “A survey on data collection for machine learning: a big data-ai integration perspective”. In: *IEEE Transactions on Knowledge and Data Engineering* 33.4 (2019), pp. 1328–1347.

[159] Bernd Röhrig et al. “Types of study in medical research: part 3 of a series on evaluation of scientific publications”. In: *Deutsches Arzteblatt International* 106.15 (2009), p. 262.

[160] David A Rorie et al. “Electronic case report forms and electronic data capture within clinical trials and pharmacoepidemiology”. In: *British journal of clinical pharmacology* 83.9 (2017), pp. 1880–1895.

[161] Thomas Brox Røst et al. “Identifying catheter-related events through sentence classification”. In: *International Journal of Data Mining and Bioinformatics* 23.3 (2020), pp. 213–233.

[162] Patrick Ruch, Robert Baud, and Antoine Geissbühler. “Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record”. In: *Artificial intelligence in medicine* 29.1-2 (2003), pp. 169–184.

[163] Stefano Rusconi et al. “Maraviroc as intensification strategy in HIV-1 positive patients with deficient immunological response: an Italian randomized clinical trial”. In: *PLoS One* 8.11 (2013), e80157.

[164] Philippe Ryvlin, J Helen Cross, and Sylvain Rheims. “Epilepsy surgery in children and adults”. In: *The Lancet Neurology* 13.11 (2014), pp. 1114–1126.

[165] David L Sackett. “Evidence-based medicine”. In: *Seminars in perinatology*. Vol. 21. 1. Elsevier. 1997, pp. 3–5.

[166] Debopam Samanta, Megan Leigh Hoyt, and Michael Scott Perry. “Healthcare professionals’ knowledge, attitude, and perception of epilepsy surgery: A systematic review”. In: *Epilepsy & Behavior* 122 (2021), p. 108199.

[167] Guergana K Savova et al. “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications”. In: *Journal of the American Medical Informatics Association* 17.5 (2010), pp. 507–513.

[168] Ingrid E Scheffer et al. “ILAE classification of the epilepsies: position paper of the ILAE Commission for Classification and Terminology”. In: *Epilepsia* 58.4 (2017), pp. 512–521.

[169] Tobias Schnabel et al. “Evaluation methods for unsupervised word embeddings”. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 298–307.

[170] William Shakespeare. *Hamlet*. Feed books, 1948.

[171] William Shakespeare. *Romeo and juliet*. Vol. 1. Classic Books Company, 2000.

[172] Soo-Yong Shin et al. “A de-identification method for bilingual clinical texts of various note types”. In: *Journal of Korean medical science* 30.1 (2015), pp. 7–15.

[173] Zbyněk Šidák. “Rectangular confidence regions for the means of multivariate normal distributions”. In: *Journal of the American Statistical Association* 62.318 (1967), pp. 626–633.

[174] Gonzalo Sirgo et al. “Validation of the ICU-DaMa tool for automatically extracting variables for minimum dataset and quality indicators: The importance of data quality assessment”. In: *International Journal of Medical Informatics* 112 (2018), pp. 166–172.

[175] L Karen Soiferman. “Compare and Contrast Inductive and Deductive Research Approaches.” In: *Online Submission* (2010).

[176] Elena Solli et al. “Deciphering the surgical treatment gap for drug-resistant epilepsy (DRE): a literature review”. In: *Epilepsia* 61.7 (2020), pp. 1352–1364.

[177] Wencheng Sun et al. “Data processing and text mining technologies on electronic medical records: a review”. In: *Journal of healthcare engineering* 2018 (2018).

[178] Abbas Tashakkori and John W Creswell. “Exploring the nature of research questions in mixed methods research”. In: *Journal of mixed methods research* 1.3 (2007), pp. 207–211.

[179] William O Tatum et al. “Minimum standards for inpatient long-term video-EEG monitoring: A clinical practice guideline of the international league against epilepsy and international federation of clinical neurophysiology”. In: *Clinical Neurophysiology* 134 (2022), pp. 111–128.

[180] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[181] Eric WK Tsang. *The philosophy of management research*. Taylor & Francis, 2016.

[182] Alexander Turchin et al. “Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes”. In: *Journal of the American Medical Informatics Association* 13.6 (2006), pp. 691–695.

[183] Nizamuddin Uddin and Hercules Dalianis. “Detection of spelling errors in Swedish clinical text”. In: *1st Nordic workshop on evaluation of spellchecking and proofing tools (NorWES T2014)*. 2014.

[184] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[185] C Van Rijsbergen. “Information retrieval: theory and practice”. In: *Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems*. Vol. 79. 1979.

[186] Vladimir Vapnik. “The support vector method of function estimation”. In: *Nonlinear modeling*. Boston: Springer, 1998, pp. 55–85.

[187] Antonio Vena et al. “Clinical characteristics, management and in-hospital mortality of patients with coronavirus disease 2019 in Genoa, Italy”. In: *Clinical Microbiology and Infection* 26.11 (2020), pp. 1537–1544.

[188] Grace Wahba et al. “Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV”. In: *Advances in Kernel Methods-Support Vector Learning* 6 (1999), pp. 69–87.

[189] Bin Wang et al. “Evaluating word embedding models: Methods and experimental results”. In: *APSIPA transactions on signal and information processing* 8 (2019).

[190] Yanshan Wang et al. “Clinical information extraction applications: a literature review”. In: *Journal of biomedical informatics* 77 (2018), pp. 34–49.

[191] Richard C Wasserman. “Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research”. In: *Academic pediatrics* 11.4 (2011), pp. 280–287.

[192] Jonathan J Webster and Chunyu Kit. “Tokenization as the initial phase in NLP”. In: *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*. 1992.

[193] Nicole G Weiskopf et al. “Defining and measuring completeness of electronic health records for secondary use”. In: *Journal of biomedical informatics* 46.5 (2013), pp. 830–836.

[194] Frank Wilcoxon. *Individual comparisons by ranking methods*. Springer, 1992.

[195] Maarten A Wildeman et al. “Can an online clinical data management service help in improving data collection and data quality in a developing country setting?” In: *Trials* 12.1 (2011), pp. 1–7.

[196] Daya C Wimalasuriya and Dejing Dou. “Ontology-based information extraction: An introduction and a survey of current approaches”. In: *Journal of Information Science* 36.3 (2010), pp. 306–323.

[197] Benjamin D Wissel et al. “Early identification of epilepsy surgery candidates: A multicenter, machine learning study”. In: *Acta Neurologica Scandinavica* 144.1 (2021), pp. 41–50.

[198] Benjamin D Wissel et al. “Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery”. In: *Epilepsia* 61.1 (2020), pp. 39–48.

[199] Raymond E Wright. *Logistic regression.* American Psychological Association, 1995, pp. 217–244.

[200] Stephen Wu et al. “Deep learning in clinical natural language processing: a methodical review”. In: *Journal of the American Medical Informatics Association* 27.3 (2020), pp. 457–470.

[201] Cao Xiao, Edward Choi, and Jimeng Sun. “Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review”. In: *Journal of the American Medical Informatics Association* 25.10 (2018), pp. 1419–1428.

[202] Yan Yan et al. “Comparison of standard and penalized logistic regression in risk model development”. In: *JTCVS Open* 9 (2022), pp. 303–316. ISSN: 2666-2736. DOI: https://doi.org/10.1016/j.xjon.2022.01.016. URL: https://www.sciencedirect.com/science/article/pii/S2666273622000286.

[203] Arister NJ Yew et al. “Transforming epilepsy research: a systematic review on natural language processing applications”. In: *Epilepsia* (2022).

[204] Lotfi A Zadeh. “Fuzzy logic”. In: *Granular, Fuzzy, and Soft Computing*. Springer, 2023, pp. 19–49.

[205] Lotfi A Zadeh. “Fuzzy sets”. In: *Information and control* 8.3 (1965), pp. 338–353.

[206] Andrea Zanichelli et al. “Hereditary angioedema with C1 inhibitor deficiency: delay in diagnosis in Europe”. In: *Allergy, Asthma & Clinical Immunology* 9 (2013), pp. 1–4.

List of Publications

Journal Papers

Giacobbe, Daniele Roberto, Emilio Di Maria, Alberto Stefano Tagliafico, Martina Bavastro, Carlo Simone Trombetta, Cristina Marelli, Gabriele Di Meco, Greta Cattardico, Sara Mora, Alessio Signori, et al. (2023). "External validation of unsupervised COVID-19 clinical phenotypes and their prognostic impact". In: *Annals of Medicine* 55.1, p. 2195204.

Dettori, Silvia, Chiara Russo, Sara Mora, Mauro Giacomini, Lucia Taramasso, Chiara Dentone, Antonio Vena, Matteo Bassetti, and Antonio Di Biagio (2022). "Prevalence of viral hepatitis in unselected, consecutively enrolled patients hospitalised for SARS-CoV-2". In: *Journal of Community Health* 47.5, pp. 800–805.

Lazarova, Elena, Sara Mora, Norbert Maggi, Carmelina Ruggiero, Alessandro Cosolito Vitale, Paolo Rubartelli, and Mauro Giacomini (2022). "An interoperable electronic health record system for clinical cardiology". In: *Informatics* 9.2, p. 47.

Taramasso, Lucia, Federica Bozzano, Anna Casabianca, Chiara Orlandi, Francesca Bovis, Sara Mora, Mauro Giacomini, Lorenzo Moretta, Mauro Magnani, Antonio Di Biagio, et al. (2022a). "Persistence of Unintegrated HIV DNA Associates With Ongoing NK Cell Activation and CD34+ DNAM-1brightCXCR4+ Precursor Turnover in Vertically Infected Patients Despite Successful Antiretroviral Treatment". In: *Frontiers in Immunology* 13.

Taramasso, Lucia, Sergio Lo Caputo, Laura Magnasco, Federica Briano, Mariacristina Poliseno, Serena Rita Bruno, Sergio Ferrara, Rachele Pincino, Giovanni Sarteschi, Sabrina Beltramini, et al. (2022b). "Long-Term Effectiveness of Rilpivirine-Based Single-Tablet Regimens in a Seven-Year, Two-Center Observational Cohort of People Living with HIV". In: *AIDS Research and Human Retroviruses* 38.6, pp. 472–479.

De Marzo, Vincenzo, Antonio Di Biagio, Roberta Della Bona, Antonio Vena, Eleonora Arboscello, Harusha Emirjona, Sara Mora, Mauro Giacomini, Giorgio Da Rin, Paolo Pelosi, et al. (2021). "Prevalence and prognostic value of cardiac troponin in elderly patients hospitalized for COVID-19". In: *Journal of Geriatric Cardiology: JGC* 18.5, p. 338.

Dentone, Chiara, Federica Portunato, Antonio Vena, Silvia Dettori, Sara Mora, Filippo Ansaldi, and Matteo Bassetti (2021a). "A comparative analysis of the first and second COVID-19 wave in Italy: evaluation of mortality in the Infectious Disease Unit of Genoa University Hospital". In: *New Microbiologica* 44.4, pp. 245–247.

Dentone, Chiara, Antonio Vena, Maurizio Loconte, Federica Grillo, Iole Brunetti, Emanuela Barisione, Elisabetta Tedone, Sara Mora, Antonio Di Biagio, Andrea Orsi, et al. (2021b). "Bronchoalveolar lavage fluid characteristics and outcomes of invasively mechanically ventilated patients with COVID-19 pneumonia in Genoa, Italy". In: *BMC Infectious Diseases* 21, pp. 1–9.

Giacobbe, Daniele Roberto, Chiara Russo, Veronica Martini, Silvia Dettori, Federica Briano, Michele Mirabella, Federica Portunato, Chiara Dentone, Sara Mora, Mauro Giacomini, et al. (2021). “Use of ceftaroline in hospitalized patients with and without COVID-19: a descriptive cross-sectional study”. In: *Antibiotics* 10.7, p. 763.

Russo, Elisa, Pasquale Esposito, Lucia Taramasso, Laura Magnasco, Michela Saio, Federica Briano, Chiara Russo, Silvia Dettori, Antonio Vena, Antonio Di Biagio, et al. (2021). “Kidney disease and all-cause mortality in patients with COVID-19 hospitalized in Genoa, Northern Italy”. In: *Journal of nephrology* 34, pp. 173–183.

Sarteschi, Giovanni, Antonio Di Biagio, Emanuele Focà, Lucia Taramasso, Francesca Bovis, Anna Celotti, Michele Mirabella, Laura Magnasco, Sara Mora, Mauro Giacomini, et al. (2020). “Viremia copy-years and risk of estimated glomerular filtration rate reduction in adults living with perinatal HIV infection”. In: *Plos one* 15.10, e0240550.

Taramasso, Lucia, Laura Magnasco, Bianca Bruzzone, Patrizia Caligiuri, Giorgio Bozzi, Sara Mora, Elisa Balletto, Paola Tatarelli, Mauro Giacomini, and Antonio Di Biagio (2020a). “How relevant is the HIV low level viremia and how is its management changing in the era of modern ART? A large cohort analysis”. In: *Journal of Clinical Virology* 123, p. 104255.

Taramasso, Lucia, Antonio Vena, Francesca Bovis, Federica Portunato, Sara Mora, Chiara Dentone, Emanuele Delfino, Malgorzata Mikulska, Daniele Roberto Giacobbe, Andrea De Maria, et al. (2020b). “Higher mortality and intensive care unit admissions in COVID-19 patients with liver enzyme elevations”. In: *Microorganisms* 8.12, p. 2010.

Taramasso, Lucia, Antonio Di Biagio, Niccolò Riccardi, Federica Briano, Elisa Di Filippo, Laura Comi, Sara Mora, Mauro Giacomini, Andrea Gori, and Franco Maggiolo (2019). “Lipid profile changings after switching from rilpivirine/tenofovir disoproxil fumarate/emtricitabine to rilpivirine/tenofovir alafenamide/emtricitabine: Different effects in patients with or without baseline hypercholesterolemia”. In: *PLoS One* 14.10, e0223181.

Conference papers

Di Meco, Gabriele, Sara Mora, Daniele Roberto Giacobbe, Silvia Dettori, Ilias Karaiskos, Matteo Bassetti, Mauro Giacomini, et al. (2022). “A Wide Database for a Multicenter Study on Pneumocystis jirovecii Pneumonia in Intensive Care Units.” In: vol. 294, pp. 557–558.

Mora, S, B Blobel, R Gazzarata, and M Giacomini (2022). “CTS2 OWL: Mapping OWL Ontologies to CTS2 Terminology Resources.” In: vol. 299, pp. 44–52.

Lazarova, Elena, Sara Mora, Davide Armanino, Alessandro Poire, Fabio Furlani, and Mauro Giacomini (2021a). “A Web Based Tool to Support a Personalized Therapeutic Path Through the Use of Psychological Tests.” In: *pHealth*, pp. 211–216.

Lazarova, Elena, Sara Mora, Paolo Rubartelli, Alessandro Cosolito Vitale, Luisa Pareto, Norbert Maggi, Carmelina Ruggiero, and Mauro Giacomini (2021b). "Integrating an electronic health record system into a regional health information system: An HL7 FHIR architecture". In: *Public Health and Informatics.* IOS Press, pp. 1087–1088.

Mora, Sara, Jacopo Attene, Roberta Gazzarata, Giustino Parruti, and Mauro Giacomini (2021). "A NLP Pipeline for the Automatic Extraction of Microorganisms Names from Microbiological Notes." In: *pHealth*, pp. 153–158.

Lazarova, Elena, Sara Mora, Antonio DI BIAGIO, Antonio Vena, and Mauro Giacomini (2020). "Reuse of Clinical COVID-19 Patient Data: Pre-Processing for Future Classification". In: p. 117.

Mora, Sara, Sumit Madan, Stephan Gebel, and Mauro Giacomini (2020). "Proposal of an Architecture for Terminology Management in a Research Project". In: *Digital Personalized Health and Medicine.* IOS Press, pp. 1371–1372.

Bonetto, M, S Mora, N Maggi, C Ruggiero, and M Giacomini (2019). "Standards for the Reuse of Clinical Data in Research: the Connection of the Liguria HIV Network and the CISAI Cohort". In: *2019 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE).* IEEE, pp. 1–3.

Giannini, Barbara, Sara Mora, Roberta Gazzarata, Antonio Di Biagio, Giovanni Cenderello, Chiara Dentone, Maurizio Setti, Daniela Fenoglio, Giovanni Cassola, Claudio Viscoli, et al. (2019). "The Ligurian HIV Network: How Medical Informatics Standards Can Help Clinical Research". In: *MEDINFO 2019: Health and Wellbeing e-Networks for All.* IOS Press, pp. 1666–1667.

Mora, S, A Venturini, G Cenderello, D Fiorellino, A Pilotto, and M Giacomini (2019). "A Web-Based Tool for a Complete Evaluation of Fragility in Senior Hiv+ Patients". In: *pHealth 2019: Proceedings of the 16th International Conference on Wearable Micro and Nano Technologies for Personalized Health.* Studies in health technology and informatics, pp. 299–302.