PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

HENRIQUE DIAS PEREIRA DOS SANTOS

APPLYING MACHINE LEARNING TO ELECTRONIC HEALTH RECORDS: A STUDY ON TWO ADVERSE EVENTS

Porto Alegre

2021

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
SCHOOL OF TECHNOLOGY
COMPUTER SCIENCE GRADUATE PROGRAM

APPLYING MACHINE LEARNING TO ELECTRONIC HEALTH RECORDS: A STUDY ON TWO ADVERSE EVENTS

HENRIQUE D.P. DOS SANTOS

Doctoral Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Ph. D. in Computer Science.

Advisor: Prof. PhD. Renata Vieira

Porto Alegre
2021

Ficha Catalográfica

S237a Santos, Henrique Dias Pereira dos Santos

Applying machine learning to electronic health records : a study on two adverse events / Henrique Dias Pereira dos Santos Santos. – 2021.

88.

Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Renata Vieira.

1. electronic health records. 2. adverse events. 3. machine learning. 4. supervised learning. 5. unsupervised learning. I. Vieira, Renata. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

Henrique Dias Pereira dos Santos

APPLYING MACHINE LEARNING TO ELECTRONIC HEALTH RECORDS: A STUDY ON TWO ADVERSE EVENTS

This Doctoral Thesis has been submitted in partial fulfillment of the requirements for the degree of Doctor of Computer Science, of the Graduate Program in Computer Science, School of Technology of the Pontifícia Universidade Católica do Rio Grande do Sul.

Sanctioned on 26th March, 2021.

COMMITTEE MEMBERS:

Prof. Dr. Rafael Heitor Bordini (PUCRS)

Prof. Dr. Cristiano André da Costa (UNISINOS)

Prof. Dr. Silvio César Cazella (UFCSPA)

Prof. Dr. Renata Vieira (PUCRS - Advisor)

“When I look at any of my achievements, I find that it is there because of certain opportunities I had, as well as because of my personal effort. I cannot claim to have created or commanded the opportunities; they were given to me. I happened to find myself in the right circumstances, so I could grow and learn what I needed to learn. I met with the right person; I happened to read the right book; I enjoyed the right company; someone came forward with the right guidance at the right time. There are so many factors behind an achievement. I cannot really say I created any of them. When I look at the facts, I must see that any achievement that I claim as mine is not due to my will or skill alone but is due to certain things that were provided to me and certain opportunities I had. And for whatever abilities I seem to have, I should be grateful.”

(The Value of Values by Swami Dayananda)

ACKNOWLEDGMENTS

The development of this project follows a series of fortunate events that I call "Alignment of the Stars." The first alignment was the sending of two stars, incredibly, into the same orbit. Ana Helena Ulbrich is a Ph.D. pharmacist with eight years of experience in clinical pharmacy, working in a tertiary hospital. She is also my sibling and she dove into this project and helped write the collaborative project with Hospital Nossa Senhora da Conceição and co-developed the DDC-Outlier. I thank her for all her work!

These two stars belong to a stellar system of passion for providing society with wellness and fairness. My parents, Celso and Zelinha, have always guided us towards a life of humbleness and servitude. Together with my twin brother, Augusto, and older sister, Juliana, we all serve one purpose: to use our work to benefit our community. I thank my family for that!

Accidentally, nearby, in the same galaxy, two other stars decided to join this constellation. Vinicius, my Ph.D. colleague, showed me the methods for experimenting with machine learning models, discussing our findings and rapidly validating hypotheses. Meanwhile, Prof. Renata accepted the challenge of becoming the advisor of this bold project. She welcomed the idea of investigating artificial intelligence in healthcare and created a research group to encourage hospitals to collaborate. I thank them both for their support!

During this Ph.D. training, I conducted research with fine colleagues: Leandro, Karin, Joaquim, Sandra, Bolivar, Bernardo, Thaila, Thiago, Marlo, Silvia, Evandro, Greice, Jackson, Rafael, João, Soraia, Isabel, Aline, Jhonatan, Juliana and Sara. Every collaboration resulted in published articles and promoted improvements in the development of the thesis. I thank them for all of that!

Another bright star in this constellation is Prof. Janete, from Nursing School. She bridged the gap between our group and health sciences and introduced us to brilliant students: Amanda, Maria, and Haline. This collaboration deepens our knowledge of patient safety and risk management. I thank them for their efforts!

The last star in this alignment is the Innovation Center of Hospital Santa Casa and its pharmacy staff. They accepted the challenge of deploying an A.I. system that improved but drastically changed their workflow. Fran, Raquel, Luana, Karol, Tati, Diego, Rogério, Rodrigo, Fábio, Marcelo. I thank you all for your work and patience!

APLICANDO APRENDIZADO DE MÁQUINA À PRONTUÁRIOS ELETRÔNICOS DO PACIENTE: UM ESTUDO EM DOIS EVENTOS ADVERSOS

RESUMO

No ambiente hospitalar, a incidência de eventos adversos (EA) (incidentes imprevistos que causam danos aos pacientes) é a principal preocupação das equipes de gerenciamento de risco. Esta tese desenvolve experimentos para avaliar abordagens de aprendizado de máquina para identificar dois grandes eventos adversos em prontruários eletrônicos do paciente (PEP). O primeiro algoritmo foi criado para identificar eventos de queda em evoluções clínicas usando modelos de linguagem e redes neurais. Anotamos 1.402 sentenças em evoluções clínicas com eventos de queda para treinar um Classificador de Token (TkC) para detectar palavras dentro do contexto de quedas. O TkC foi capaz de identificar corretamente 85% das sentenças com eventos de queda. Para a avaliação de prescrições, construímos um algoritmo não-supervisionado com base em estrutura de grafos para classificar as prescrições fora-do-padrão. Em nossos experimentos, o algoritmo proposto, o DDC-Outlier, classificou corretamente 68% (Medida-F) dos medicamentos prescritos como subdoses e overdoses. Finalmente, para entender melhor o desempenho de nossa abordagem em um cenário do mundo real, implantamos um sistema de suporte à decisão para farmácia clínica em um hospital de 1.200 leitos. Todos os experimentos, códigos-fonte e conjuntos de dados anônimos estão disponíveis publicamente na página GitHub de nosso grupo de pesquisa.

Palavras-Chave: prontuário elettrônico do paciente, eventos adversos, aprendizado de máquina, aprendizado supervisionado, aprendizado não-supervisionado.

APPLYING MACHINE LEARNING TO ELECTRONIC HEALTH RECORDS: A STUDY ON TWO ADVERSE EVENTS

ABSTRACT

In the hospital environment, the incidence of adverse events (AE) (unforeseen incidents that cause harm to patients) is the primary concern of risk management teams. The use of machine learning techniques could help healthcare professional to identify and mitigate adverse events.This thesis develops experiments to evaluate machine learning approaches to identify two major adverse events in electronic health records (EHR). The first algorithm was created to identify fall events in clinical notes using language models and neural networks. We annotated 1,402 clinical sentences with fall events to train a Token Classifier (TkC) to detect words within the context of falls. The TkC was able to correctly identify 85% of the sentences with fall events. For medication review, we built an unsupervised algorithm based on graph structure to rank outlier prescriptions. In our experiments, the proposed algorithm, the DDC-Outlier, correctly classified 68% (F-measure) of prescribed medications as underdoses and overdoses. Finally, to better understand the performance of our approach in a real-world scenario, we deployed a decision support system for clinical pharmacy in a 1,200-bed hospital. All experiments, source-codes, and the anonymized datasets are publicly available on the GitHub page of our research group.

Keywords: electronic health records, adverse events, machine learning, supervised learning, unsupervised learning.

CONTENTS

1. INTRODUCTION

In 2020, humanity faced one of the most significant health crises in history. Besides a dangerous virus, people were isolated in their homes for months. Health sciences offered people hope that a vaccine would be found, and technology made it possible for people to connect with one another through the Internet.

Health Information Technology continues producing positive effects on medical outcomes, which certainly supports efforts that prepare them to be used meaningfully (Kruse and Beane, 2018). Aside from deploying health information systems, artificial intelligence techniques have produced valuable healthcare improvements in recent years. Algorithms can predict disease, cluster likely outcomes, and detect cancer cells (Wang and Preininger, 2019). Moreover, computer science can bring benefits to other fields of healthcare.

The primary research subareas applying machine learning to health informatics are imaging (39%), diagnosis (37%), and public health (26%), followed by sensing (16%) and bioinformatics (14%). Besides being an important topic related to patient safety, adverse events (3%) are not a frequent subject: there have been few published papers in recent years[1].

Aside from being essential subfields, imaging and diagnosis are tasks related to physicians' activities that are critical (Char et al., 2018). According to (Mateen et al., 2020), in general, while theoretical and technical contributions using clinical data to illustrate applicability are fundamental to the progress of the field, they are by nature different from attempts to create a prediction model for clinical practice (Mateen et al., 2020). Computer science could be useful in other healthcare administrative departments, such as risk management and risk assessment. Adverse events are a hospital issue related to the non-critical task of risk management, thus favoring the use of machine learning techniques as a clinical decision support system. Adverse events should be understood as opportunities to improve the quality of care and may serve as a basis for the development of patient safety management strategies.

Adverse events could be identified in electronic health records (EHRs) by using machine learning algorithms. The most common adverse events relate to surgical procedures, medical procedures, diagnosis, obstetrics, medications, and fractures (Da Saúde (BR), 2014, Mendes et al., 2013). It may be possible to find these events by using unsupervised learning (finding patterns in the input data) or supervised learning (learning how to map the

[1]The distribution of published papers that use machine learning in subareas of health informatics has been obtained from Google Scholar (similar to (Ravì et al., 2016)); the search phrase is defined as the subfield name ("imaging," "diagnosis," "public health," "sensing," "bioinformatics") with the exact phrase "machine learning" and at least one of the following terms: "medical" or "health," e.g., "imaging" "machine learning" medical OR health. The total number is the phrase: "machine learning" medical OR health

input to correct values provided by a supervisor). Each adverse event detection task may use a specific machine learning solution.

The main objective of this thesis is to evaluate the use of machine learning techniques on electronic health records (EHRs) for decision support systems for clinical risks. Thus, this thesis seeks to use electronic health records as the source of information for such systems. Clinical features such as prescriptions and clinical notes could be used to identify and mitigate adverse events. To evaluate the use of machine learning in healthcare problems, we selected two non-critical, but important, tasks in the hospital environment: fall detection for risk management and prescription prioritization for clinical pharmacy.

We first approached the risk management problem by using supervised learning. We used a Bidirectional Long Short-Term Memory (BiLSTM) neural network to identify fall events in clinical notes. A fall event is classified as the event in which “the person inadvertently falls to the ground or lower levels." In the past ten years, other studies used machine learning to detect falls in clinical notes (Tremblay et al., 2009, McCart et al., 2013, Luther et al., 2015, Bates et al., 2016, Shiner et al., 2016, Topaz et al., 2019). Nevertheless, none of them used neural networks and language models. They also could not be used in Portuguese. Our experiments show that the BiLSTM, together with the Conditional Random Field (CRF), improved the results in this task (using distinct datasets). The latest studies (Luther et al., 2015, Topaz et al., 2019) reached up to 90% of F-measure using machine learning on clinical notes. We annotated 1,402 clinical sentences with fall events to train a Token Classifier (TkC) to detect words within the context of falls. The TkC was able to correctly identify 85% of the sentences with fall events and achieve an F-measure of 96% in Cross-Validation.

The next problem we tackled was medication errors. We built an unsupervised algorithm based on graph structure to rank outlier prescriptions, speeding up the medication review process performed by clinical pharmacists. A medication error with clinical significance is defined as an unintentional decision error that can reduce the likelihood of the treatment being effective or increase the risk of injury to the patient. Previous works, such as (Park et al., 2017) and (Nangle et al., 2017), developed machine learning models on prescription data to extract information or show divergent patterns in prescriptions for the same diseases, but they had no application to clinical pharmacy. Another study tried to detect medication errors using machine learning (Schiff et al., 2017), but it only used the patients' dosage history to alert physicians about dosage errors. In our experiments, we used historical hospital data to detect outlier prescriptions. The outlier can be used to assist pharmacists in the medication review process. The proposed unsupervised algorithm, the DDC-Outlier, is best used to detect medication overdosing and underdosing. The DDC-Outlier correctly classified 68% (F-measure) of the medications prescribed in our experiments. Both machine learning techniques can help develop a decision support system and each learning approach (supervised or unsupervised) fits the proper task.

Moreover, this thesis goes beyond a mere evaluation of the outlier algorithm using historical data. As pointed out by Mateen et al. (2020), it is quite challenging to translate laboratory results into realistic settings. The knowledge of machine learning researchers needs to be integrated with the knowledge of healthcare experts. This type of endeavor takes time as there is a need to build trust between all parties (Mateen et al., 2020). In an effort to better understand the performance of the DDC-Outlier, we developed a real-world case study in partnership with one hospital. The selected hospital has 1,200 beds and diverse patient profiles, which are important features to evaluate the algorithm's generalization capability. We deployed the proposed algorithm embedded in a decision support system for clinical pharmacy, also developed as part of this Ph.D. candidature. We gathered data from 24,702 prescriptions reviewed by the hospital's pharmacists during six months, showing promising results. The DCC-Outlier correctly classified the dose and frequency (posology) of several medications in the real-scenario application, improving the work of the hospital's pharmacists.

1.1 Thesis Organization

The rest of this thesis is organized as described in Figure 1.1: in the next Chapter, we state the concepts around inpatient risk, medication review, machine learning, and its uses in the healthcare industry. Chapter 3 describes the experiments that use a neural network to detect fall events in a supervised task. In Chapter 4, we report the experiments that use unsupervised learning in the task of prescription prioritization.

This thesis presents the proposed algorithms and experiments in two separate chapters, as described above. For that reason, each of these two chapters (3 and 4) have their own related-work section, presenting previous work about each subject. In Chapter 5, we detail the results of the case study using the DDC-Outlier embedded in a decision support system for the task of prescription prioritization in clinical pharmacy. Finally, in Chapter 6, we summarize our conclusions, present further research directions, and enumerate the publications related to this thesis.

Besides the main aforementioned topics, during this Ph.D. study, we developed other experiments using data from electronic health records. These studies cover areas such as information extraction, de-identification of clinical notes, patient complexity detection, quality evaluation of word embeddings, and a systematic review of fall prediction. We list the papers that describe these other related studies in Appendix A.

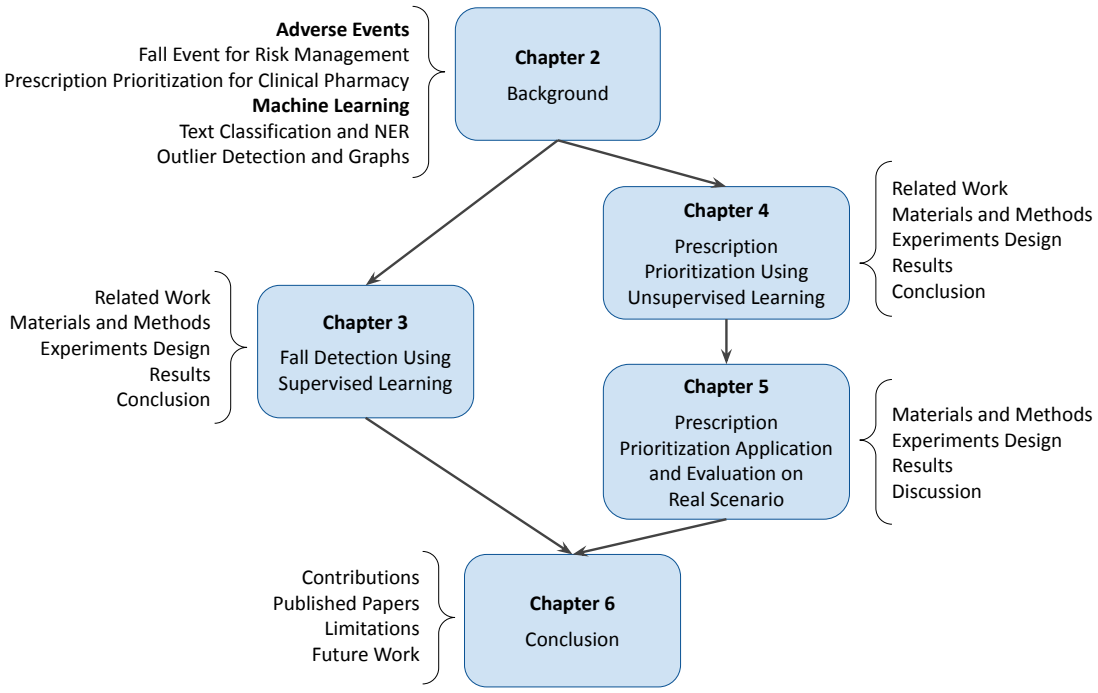

Figure 1.1 – Schema - Thesis Organization

1.2 Ethical Aspects

All data used in the experiments conducted for this thesis came from a project developed with Hospital Nossa Senhora da Conceição (HNSC) and Hospital Santa Casa (HSC). Ethical approval to use the hospital datasets in this research was granted by the Research Ethics Committee of the Hospital Group under the number 71571717.7.0000.5530.

2. BACKGROUND

This chapter introduces the relevant concepts necessary to understand this work. Therefore, the chapter establishes the foundation from which the techniques and research propositions are derived.

2.1 Hospital Adverse Events

Adverse Events (AEs) are defined as unwanted complications resulting from the care provided to patients. These complications are not attributed to the natural evolution of the underlying disease. AEs are currently one of the biggest challenges to improve quality in healthcare: their presence reflects the marked distance between ideal care and real care. It should be noted that 50% to 60% of AEs are preventable (Gallotti, 2004).

Adverse events should be understood as opportunities to improve the quality of care and may serve as a basis for the development of patient safety management strategies. Being aware of these problems is essential to plan improvement actions and guide the development of policies with a focus on safety and quality care.

In Brazil, in order to improve the quality of health products and of the management and monitoring of adverse events in hospitals, the Brazilian Health Regulatory Agency (ANVISA) created the Brazilian Sentinel Hospital Network in 2002. This network enforces that hospitals must notify any severe adverse events to ANVISA (De Oliveira et al., 2016). Besides, the Institute for Healthcare Improvements has developed the IHI Global Trigger Tool for the detection of adverse events. The tool uses "triggers," or clues, to identify adverse events (AEs), which is an effective method to measure the overall level of harm in a healthcare organization (Griffin and Resar, 2009).

These initiatives aim to improve patient safety in healthcare and monitor adverse events. The following section focuses on two services that are responsible for two adverse events related to patient safety.

2.1.1 Fall Detection for Risk Management

In the hospital environment, the risk management team develops actions within three scopes: pharmacovigilance, which is responsible for the control and surveillance of drugs; hemovigilance, which receives reports on side effects, blood transfusions, and blood products; and technovigilance, which controls the quality of hospital inputs and equipment

and oversees patient care, being responsible for the reporting of adverse events related to nursing care (De Oliveira et al., 2016).

One of the largest categories of adverse event reports is "falls." The World Health Organization (WHO) defines a fall as the event in which “the person inadvertently falls to the ground or lower levels” (Ageing and Unit, 2008). Regarding patient care, falls comprise the largest category of adverse event reports within hospitals and nursing homes. Approximately 30% of in-patient falls result in injury and 4% to 6% result in serious injury (Hitcho et al., 2004). There are two distinct tasks related to falls in risk management departments: fall detection and fall risk assessment. First, fall detection refers to identifying the falls that occurred in the hospital and mapping the risk factors associated with each event. Second, it aims to promote educational interventions, determine barriers, and perform fall risk assessment during patient admission (Fortinsky et al., 2004).

Besides risk management (which relates to patient care), the hospital environment requires other risk assessments, such as evaluations regarding finances, innovation, legal protection, elderly patients, professional staff, and information protection (Etges et al., 2018). Furthermore, medication errors are also considered a crucial risk factor for patients. This leads to a more specific hospital service that handles medication risks: the clinical pharmacy, explained in the following section.

2.1.2 Prescription Prioritization for Clinical Pharmacy

In a hospital facility, the clinical pharmacy performs a key activity to improve the appropriateness, effectiveness, safety, adherence, and affordability of drug therapies. Clinical pharmacists provide care to patients as members of multidisciplinary patient care teams, assuming responsibility and ensuring accountability for optimizing medication-related outcomes. Pharmacists provide fundamental services that are the core components of the pharmacy practice (e.g., drug order fulfillment, patient education, information on drugs, public health-related services) (Saseen et al., 2017).

One of the responsibilities of clinical pharmacists is medication review. Medication review is a structured evaluation of a patient’s medication to optimize medication use and improve health outcomes. The medication review entails detecting drug-related problems and recommending interventions (changes in the patients’ prescription) (Griese-Mammen et al., 2018). Besides improving patient outcomes (Graabæk and Kjeldsen, 2013), this evaluation contributes to the economic efficiency of hospitals (Touchette et al., 2014) and has an impact on the decrease of mortality rates (Bond and Raehl, 2007).

To improve the medication review process, pharmacists create tools to measure patient acuity and prioritize pharmaceutical care (Alshakrah et al., 2019). These tools may be computer-based systems or manual protocols that they follow step-by-step. Most tools

are designed to identify patients at a greater risk of adverse drug reactions, adverse drug events, or medication errors, guiding appropriate pharmaceutical care. The prioritization task uses several risk factors for the early detection and prompt management of high-risk patients in clinical settings. Risk factors include the drug-related and patient-related risks presented in Table 2.1 (Alshakrah et al., 2019).

Drug Related	Patient Related
•high-risk medication, •drugs requiring monitoring, •polypharmacy, •use of total parenteral nutrition/nasogastric tube, •high-cost medication, •number of intravenous medications, •and number of unlicensed medication	•age, •renal impairment, •comorbidity, •hepatic impairment, •reason/time/type of admission, •readmission, •allergies, •and length of stay

Table 2.1 – Risk Factors for Prescription Prioritization (Alshakrah et al., 2019)

Several aspects may be improved in both hospital services — risk management and clinical pharmacy. The advent of electronic health records enables risk assessment to be automated by using a computer-based system together with artificial intelligence (AI) (Goldstein et al., 2017). We detail the existing data in electronic health records in the following section.

2.2 Electronic Health Records

Electronic health records (EHRs) have produced valuable improvements in hospital practices by integrating patient information. In fact, the understanding of this data can mitigate mistakes that may put patients' lives at risk. EHRs contain unstructured data (e.g., text and images) and structured data (e.g., prescriptions, laboratory results, and patient profiles). Figure 2.1 shows the variety of data that compose EHRs and its possible uses.

According to Jensen et al. (2012), EHRs can be seen as a repository of information regarding patients' health status in a computer-readable format. An encounter with the healthcare system generates several types of patient-linked data. In the example shown in Figure 2.1, medication, laboratory, imaging, and narrative data are acquired. Each data type is ideally captured according to standards or classifications, such as RxNorm (Fung et al., 2008) for prescription data, Logical Observation Identifiers Names and Codes (LOINC) for laboratory data, and Digital Imaging and Communication in Medicine (DICOM) for imaging files. Clinical narratives are inherently free text, but often feature clinical terms that are coded according to the International Classification of Disease-9 (ICD-9), the ICD-10, or the Systematized Nomenclature of Medicine — Clinical Terms (SNOMED CT) (De Silva et al., 2011).

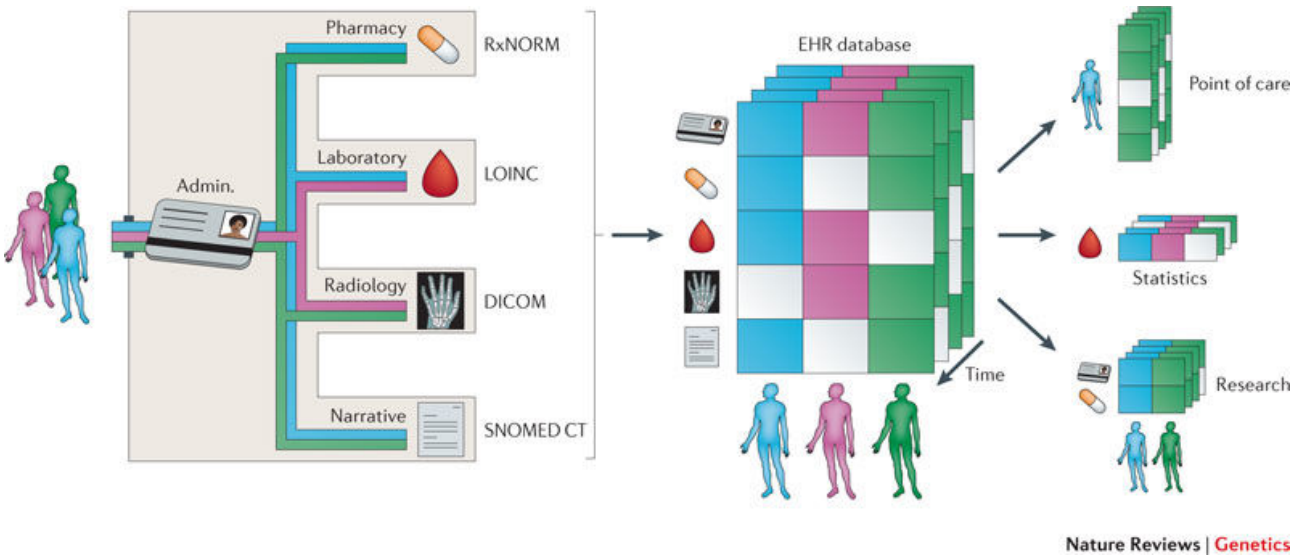

Figure 2.1 – Electronic health records and its potential uses (Jensen et al., 2012).

Integrated auto-coding systems may in some cases render free text into clinical terms. Patient data is stored in a database and can be viewed in formats that match the needs and authorities of specific user groups.

For example, a clinician might request EHR data for a particular patient, a statistical summary of all laboratory procedures, and a specific cohort extraction for drug research (Jensen et al., 2012). In the next section, we cover a subarea of A.I. that could improve patient safety using EHR data as a source of information: machine learning.

2.3 Machine Learning

Machine Learning (ML) has received a lot of attention from computer science researchers and other research fields in recent years (Jordan and Mitchell, 2015). The growth of processing power and the amount of data now available enabled the rising of ML in several applications in several domains.

According to Apaydin (2020), machine learning consists of programming computers to optimize a performance criterion using example data or past experiences. The model is defined by some parameters, and the learning comprises the execution of a computer program to optimize the parameters of the model using the training data. The model may be predictive (to make predictions in the future), descriptive (to gain knowledge from data), or both (Alpaydin, 2020).

The data-related approaches of ML can be divided into supervised and unsupervised learning (other techniques, such as reinforcement learning, are not related to real data but simulations). Both learning strategies attempt to understand patterns from real data. Learning a rule from data also allows knowledge extraction and pattern recognition. The

rule is the extraction of a simple model that explains the data; by looking at this model, an explanation of the process underlying the data is proposed. For each approach, there are several algorithms and applications for different domains, each working better depending on the characteristics of the data and the size of the dataset (Han et al., 2011).

In the subsequent sections, we discuss in more detail the tasks of ML used in this work: text classification and outlier detection.

2.3.1 Text Classification and Named Entity Recognition

In supervised learning (also known as predictive learning), the aim is to learn a mapping from the input to an output whose correct values are provided by a supervisor. The approach in machine learning is that we assume a model defined by a set of parameters learned by using data labeled by an expert. The algorithms can learn patterns for binary or multiple classes (targets/labels) as classification problems and can learn continuous outputs as regression problems (Alpaydin, 2020).

Sequence(or text) classification is where an entire text or document is assigned to a category, using supervised learning (Jurafsky and Martin, 2014). One common text categorization task is sentiment analysis, that is, the extraction of sentiments — the positive or negative orientation that a writer expresses toward an object. Traditional algorithms for text classification, such as Naive Bayes and Logistic Regression, have been outperformed by neural networks.

Named entity recognition (NER) is the task to find spans of text that constitute proper names and tag the type of the entity. The most common entity tags: PER (person), LOC (location), ORG (organization), or GPE (geopolitical entity). However, the term “named entity” commonly also refers to things that are not entities per se, including dates, times, and other kinds of temporal expressions and even numerical expressions, such as prices. Here is an example of the output of a NER tagger (Jurafsky and Martin, 2014). NER could also be used for text classification when an entire text has a tag or not. Algorithms commonly used for NER, such as Hidden Markov Model (HMM) and Conditional Random Fields (CRF), have also been outperformed in this task by neural networks.

In the last decade, after the implementation of neural networks in graphics processing units (GPU) (Oh and Jung, 2004) and the rise of open-source frameworks (TensorFlow (Abadi et al., 2016) by Google, PyTorch (Paszke et al., 2019) by Facebook), neural network algorithms resurfaced prominently. An artificial neural network structure is a non-parametric estimator that can be used for classification, regression, and other tasks. Two network topologies have been especially useful: Recurrent Neural Networks (RNNs) for natural processing language and Convolutional Neural Networks (CNNs) for computer vision. RNNs, such as Long Short-Term Memory (LSTM) or Bidirectional-LSTM, have memory ca-

pabilities that are able to extract stream features from text processing to voice recognition and achieve better results in classification tasks than classic machine learning methods such as Naive Bayes and CRF. Besides, CNNs can extract image patterns from pictures to movies and are better for object recognition (Goodfellow et al., 2016).

Recurrent neural networks have proven to be an effective approach to language modeling, both in sequence labeling tasks such as part-of-speech tagging, as well as in sequence classification tasks such as sentiment analysis and topic classification (Jurafsky and Martin, 2014). Many studies have been using neural networks as self-supervised learning to train language models. Models that assign probabilities to sequences of words are called language models (LMs) (Jurafsky and Martin, 2014). More sophisticated language models than probabilistic LMs use recurrent neural networks to generate the models (word embeddings). RNN-based language models are designed to process sequences in segments, attempting to predict the next word in a sequence by using the current word and the previous hidden state as inputs (Mikolov et al., 2013).

One of the first models to use such a strategy was word2vec (Mikolov et al., 2013), which was built with a feed-forward neural network. Subsequently, other studies improved these neural networks by developing complex topologies for robust language models. Some examples are Glove (Pennington et al., 2014), FastText (Bojanowski et al., 2017), ElMo (Peters et al., 2018a), Flair (Akbik et al., 2018), and Transformers-based language models (contextual embeddings), such as BERT (Devlin et al., 2018)[1], OpenAi GPT (Radford et al., 2019), and XLNet (Yang et al., 2019).

2.3.2 Outlier Detection and Graphs

In unsupervised learning (also known as descriptive learning), there is no supervisor; there is only input data and no known output. Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns (Barlow, 1989). There is a structure to the input space such that certain patterns occur more often than others, and we want to see what generally happens and what does not (Alpaydin, 2020).

Outlier detection (also known as anomaly detection) is the process of finding data objects with behaviors that are very different from expectation, using, in most cases, unsupervised learning (Han et al., 2011). Outlier detection and clustering analysis are two highly related tasks. Clustering finds the majority patterns in a data set and organizes the data

[1]Originally, BERT is not a traditional language model. It is a model trained on a masked language model loss, and it cannot be used to compute the probability of a sentence like a regular LM. A regular LM takes an autoregressive factorization of the probability of the sentence

accordingly, whereas outlier detection tries to capture those exceptional cases that deviate substantially from the majority patterns (Han et al., 2011).

According to the assumptions made, we can categorize outlier detection methods into three types: statistical methods, proximity-based methods, and clustering-based methods. Proximity-based methods assume that an object is an outlier if the nearest neighbors of the object are far away in feature space, that is, the proximity of the object to its neighbors significantly deviates from the proximity of most of the other objects to their neighbors in the same dataset (Han et al., 2011).

There are two types of proximity-based outlier detection methods: distance-based and density-based methods. A distance-based outlier detection method consults the neighborhood of an object, which is defined by a given radius. An object is then considered an outlier if its neighborhood does not have enough other points. A density-based outlier detection method investigates the density of an object and that of its neighbors (Han et al., 2011).

Outlier detection usually is used in vector-space data, but graphs could be used for this type of task. For instance, (Muller et al., 2013) employed centrality algorithms to rank nodes based on their centrality index to distinguish them between inliers and outliers. The same idea was successfully employed, for example, to rank textual information of documents and automatically create textual summaries that contain the most common words based on their centrality on graph (Woloszyn et al., 2017a, Woloszyn et al., 2017b). Another approach in the graph centrality field used a random walk on a graph to perform outlier detection (Moonesinghe and Tan, 2008). It relied on computing the node similarity and the number of shared neighbors between nodes. Afterward, they used a Markov chain model to compute the score for each node of the graph. Nodes with lower scores were considered outliers.

2.4 Machine Learning for In-Hospital Adverse Events

According to Goldstein et al. (2017), several in-hospital adverse events (AE) could take advantage of machine learning algorithms. Hospital-acquired infections, virological failures, acute lung injuries, fractures, readmissions, pressure ulcers, and sepsis are some examples of AE problems that researchers have been trying to solve by using prediction algorithms (Goldstein et al., 2017). These predictive models typically used techniques such as generalized linear models, bayesian methods, random forests, and regularized regression. Most studies that use regression have incorporated some form of variable selection, most often via stepwise approaches (Goldstein et al., 2017).

The following studies have used neural networks to predict adverse events in text information concerning health. Some studies regarding adverse drug events (ADE) used re-

current neural network architecture to detect events in electronic health records (EHRs) (Jagannatha and Yu, 2016, Wunnava et al., 2018) and attention neural networks to highlight important words related to these events (Huynh et al., 2016), to identify harm events in patient care (Cohan et al., 2017), and to indicate the possible occurrence of adverse cardiac events (Chu et al., 2018).

This thesis focuses on two tasks: token classification in clinical notes and outlier detection in medication data. We use a Recurrent Neural Network with Conditional Random Fields to classify tokens related to fall events for the first task. In the second task, outlier prescription detection, we develop new unsupervised algorithms that consider the distance, density, and centrality of prescribed medications to evaluate their outlierness. The next chapters present two different machine learning approaches used to identify two different adverse events: detection of fall events using supervised learning and prescription prioritization using unsupervised learning.

3. FALL DETECTION USING SUPERVISED LEARNING

Falls are critical adverse events that occur in the hospital environment. Within hospitals and nursing homes, falls constitute the largest category of adverse event reports. Approximately 30% of inpatient falls result in injury, with 4% to 6% resulting in serious injury (Hitcho et al., 2004). Therefore, a starting point for fall prevention programs should always be a critical review of evidence (Oliver, 2007). An automated system to detect falls could assist in the smart screening of adverse events.

Traditional fall risk protocols (Morse et al., 1989) and fall detection protocols (Resar et al., 2006) were developed for hospital environments without EHR systems. These protocols are useful but time-consuming and do not consider cultural changes for various hospitals and countries (De Souza Urbanetto et al., 2013). The adoption of electronic health records in hospital environments brings many benefits for patients (Buntin et al., 2011). For example, the data extracted from EHRs is commonly used in clinical decision support systems to improve patient safety and healthcare quality.

This chapter proposes a new approach in terms of fall detection in clinical notes. We built an annotated dataset and used a state-of-the-art natural language processing neural network to detect fall events from text information present in EHRs (clinical notes). In the following section, we present a systematic review of fall detection using machine learning techniques.

3.1 Related Work

In this section, we detail a systematic literature review performed to understand previous work related to fall detection. Walsh et al. (2016) also proposed a systematic review of the detection of fall events, but focused on non-automated models, listing articles about fall risk prediction models, predicting falls among inpatients and recording falls in the community setting (Walsh et al., 2016). Another fall-related review focused on sensor information using machine learning algorithms in wearable, ambiance, and vision-based devices (Mubashir et al., 2013), not electronic health records.

Thus, we focused on understanding how EHR data has been used to develop and validate automatically-built models to identify in-hospital fall events. We focused on EHR data because a large amount of useful information is generated during patients' stay.

We selected five relevant digital libraries in Computing Science and Health: ACM Digital Library[1], ScienceDirect[2], IEEExplore[3], Scopus[4], and Pubmed[5]. Scopus is a general database that indexes several other databases, covering approximately 19,500 titles from more than 5,000 international publishers, including coverage of 16,500 peer-reviewed journals in the scientific, technical, medical, and social sciences. Afterward, keywords related to the research topic were identified, such as “fall”, “electronic health records”, and “artificial intelligence techniques”. These terms were combined to create the search expressions. The search expressions were adapted according to the mechanism of each digital library, so as not to alter their logical sense. The searches were performed in the abstract, title, and keywords fields.

For the “fall” concept, we used the search expressions *(fall detection OR falls detection) AND*; for “electronic health records”, we used *(electronic health records OR EHR OR electronic medical records OR EMR OR narratives OR free-text records OR clinical notes)* AND; for “artificial intelligence techniques”, we used *(machine learning OR data mining OR text mining OR neural networks OR natural language processing OR information extraction OR decision trees OR prediction)*.

Table 3.1 presents the studies selected for this systematic review. In the following sections, we summarize the studies considering research design, data types, outcomes, and evaluation to list what is relevant and useful about these models.

Table 3.1 – Characteristics of Studies regarding Fall Incident Detection

Author	Data Points	Source	Algorithms	Eval	Ss	F1
(Tremblay et al., 2009)	2,157	notes	K-Means, LR	T/T	0.83	
(Toyabe, 2012)	4,821	notes	Syntactic Rules	DV	0.87	0.84
(McCart et al., 2013)	26,010	notes	LR, SVM	T/T	0.93	0.85
(Rochefort et al., 2015)	NR	inc-rep	NR	NR	0.83	
(Luther et al., 2015)	26,010	notes	SVM	T/T	0.94	0.90
(Bates et al., 2016)	8,288	rad-rep	SVM	CV	0.94	0.93
(Shiner et al., 2016)	2,730	notes	MaxEnt, CRF	T/T	0.97	
(Topaz et al., 2019)	750	notes	Random Forest	T/T	0.90	0.89

NR = Not Reported; notes = Clinical Notes; inc-rep = Incident Reports; rad-rep = Radiology Reports; LR = Logistic Regression; Eval = Evaluation Method; T/T = Dataset split into training and test sets once; DV = Direct Validation; CV = Cross Validation; Ss = Sensitivity; F1 = F-Measure;

[1] https://dl.acm.org/
[2] https://www.sciencedirect.com/
[3] https://ieeexplore.ieee.org/Xplore/home.jsp
[4] https://www.scopus.com/
[5] https://www.ncbi.nlm.nih.gov/pubmed/

3.1.1 Design of Studies on Fall Detection in EHRs

All fall detection studies analyzed employed textual information extracted from EHRs as input for their proposed models. Fall events reported in EHRs are usually presented as unstructured data (text). Clinical notes were the most commonly used sources (n = 6) (Tremblay et al., 2009, Toyabe, 2012, McCart et al., 2013, Luther et al., 2015, Shiner et al., 2016, Topaz et al., 2019). Other sources used include radiology reports (n = 1) (Bates et al., 2016) and incident reports (n = 2) (Toyabe, 2012, Rochefort et al., 2015).

Regarding the strategy to build fall detection models, most papers used machine learning algorithms (n = 6) (Tremblay et al., 2009, McCart et al., 2013, Luther et al., 2015, Bates et al., 2016, Shiner et al., 2016, Topaz et al., 2019). Three of them selected Support Vector Machines as the most commonly used algorithm (n = 3). Only one study used syntactic rules (Toyabe, 2012) to develop fall detection models.

(Tremblay et al., 2009) developed a logistic regression model using unsupervised term importance weighting. In another study, (Toyabe, 2012) made syntactic category decision rules to detect inpatient falls from texts. (McCart et al., 2013) trained a support vector machine and logistic regression model with several parameter searches on records of the Veterans Health Administration (VHA). (Luther et al., 2015) improved McCart's experiments with VHA data using linear SVM and normalizing terms using the lexical tool of the National Library of Medicine. (Bates et al., 2016) also used SVM on texts but extracted features using the Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010).

(Shiner et al., 2016) selected admission charts of Veterans Affairs (VA) hospitals and trained the model using Automated Retrieval Console (ARC) (D'Avolio et al., 2010) to detect fall incidents. ARC is able to map text for various part-of-speech tags (e.g., nouns, verbs, noun phrases, etc.) and find unique medical concepts in the UMLS (Unified Medical Language System of the National Library of Medicine). Lastly, (Topaz et al., 2019) developed NimbleMiner, an NLP tool that combines several machine learning approaches to extract terms and train a random forest model. (Rochefort et al., 2015) did not explain the methods they used in their research.

3.1.2 Limitations

Besides the contribution of the aforementioned papers, we also analyzed what some authors listed as the limitations of their studies. The most-reported limitations were the data selection (n = 4) (Tremblay et al., 2009, McCart et al., 2013, Shiner et al., 2016, Topaz et al., 2019), where the author has to select the input data before training the model. The

next most common issue was the generalization problem of the models (n = 2) (Shiner et al., 2016, Topaz et al., 2019).

(Tremblay et al., 2009) showed examples where the trained model misclassified fall-related adverse events. The examples featured words such as hip, pain, and knee, which are commonly found in fall incidents; however, that is not always the case. (McCart et al., 2013) discussed the problems in the gathered dataset. An inadequacy in the reported incidents added bias to the trained model.

(Shiner et al., 2016) warned about their small and random sample to identify falls. Their study design reduced possible variations in the way falls are described. They stated that further work should test multiple methods for fall identification, including incident reports, manual record reviews, and patient self-reports.

Other than the vast use of deep learning in several areas of medicine (Topol, 2019), no studies used neural networks to predict fall incidents. To the best of our knowledge, there are no previous studies addressing fall event detection from text using word embeddings or deep learning. In the next section, we cover the dataset, the neural network, and the language models (word embeddings) used in the experiments described in this thesis.

3.2 Fall Event Detection in Clinical Notes

This section details the experiments performed in this study to evaluate the fall detection models in EHRs. First, we cover the dataset used and the annotation process. Then we explain the language models and neural network topologies used to build a fall detection model.

3.2.1 Materials and Methods

A retrospective cohort study was developed in a dataset from a large public tertiary hospital in the city of Porto Alegre, in Southern Brazil. The dataset contains 2,698 clinical notes from 1,694 patients who had a fall between the years 2012 and 2017; during this period, nurses voluntarily reported 1,971 fall incidents in patients' charts. Although all of these patients suffered a fall, as included in the incident reports, 342 (32.97%) of the patients' records did not contain the clinical notes of the incident.

An incident report is a form attached to the patients' records describing the fall event in detail. In the hospital under study, these reports are not directly related to each clinical note; they were associated based on the date of the incident reported and the patients'

identification. These voluntary reports were used to flag each clinical note as positive or negative regarding fall incidents.

Because of limited time and resources, we only used a sample of this dataset. The sample considered a percentage of 22% of fall incident reports, given the sampling error of 2.5% and the statistical significance of 5%. Therefore, 1,078 clinical notes, 367 patients, and 441 incident reports with records of possible fall incidents were included in this study. Each incident report had on average 2.4 clinical notes referring to the same patient and date.

The following steps were performed to prepare the dataset to train the machine learning models:

- Selection: identifying all inpatients with at least one reported fall incident and their clinical notes;
- De-identification: de-identifying the data to ensure patient anonymity;
- Annotation: creating a "gold standard" with the charts reviewed by nursing students.

Table 3.2 – Example of a Corpus Used for the Clinical Note Classification

Original Note	**Translated Note**
Evolução: Transtornos mentais devidos ao uso de álcool síndrome, estado de abstinência. Plantão de intercorrências: **queda da cama** com TCE frontal fechado, sem alteração de consciência, sem cervicalgia.	Clinical note: Mental disorders due to the use of alcohol syndrome, withdrawal status. Complications on call: **falling out of bed** with closed frontal TBI, no alteration of consciousness, no neck pain.
Evolução: T08 Fratura da coluna. Restrição a atividades físicas rigorosas: capaz de realizar trabalhos leves e de natureza sedentária. Dor abdominal e pélvica. R53 Mal estar fadiga. Paciente bem ativo no leito sem queixas. Tosse seca não aceitou sentar.	Clinical note: T08 spinal fracture. Restriction to vigorous physical activities: able to perform light and sedentary work. Abdominal and pelvic pain. R53 Illness fatigue. PHYSICAL THERAPY: Patient very active in bed with no complaints. Dry cough refused to sit.

Examples of annotated clinical notes are shown in Table 3.2.

3.2.2 Falls Annotation Process

The data collection of the incident reports and data annotation of clinical notes used the WebAnno system (Yimam et al., 2013) and lasted four months, consisting of careful reading by three different nursing students, with double checks. In cases of incongruities

or doubts, notes were taken in a spreadsheet and later discussed during meetings of the research group.

Each word or phrase was annotated with several definitions related to the fall, according to the WHO Technical Report about Patient Safety (Organization, 2009). Some of the annotated concepts are: procedure after fall; medical assessment; damage level (none, low, medium, high, death); damage type (physical, psychological, social).

The annotated dataset totaled 1,078 clinical notes, 723 (68%) of which did not have any fall incidents, while 355 (32%) notes have fall-related incidents annotated by the nursing students. In our experiments, we designed the task as a classification problem and used the notes with and without fall-related incidents. Table 3.3 shows the distribution of in-hospital fall incidents among the patients.

Table 3.3 – Fall per Patient in the Annotated Dataset

# of Patients	% of Total	# of Falls
316	87.0%	1fall
36	10.1%	2 falls
11	3.0%	3 falls
1	0.3%	4 falls
2	0.5%	5 falls
1	0.3%	6 falls

3.2.3 Language Models

Word vector representations (word embeddings) bring a new perspective for Natural Language Processing. This approach outperforms traditional rule-based or machine learning methods (Li and Yang, 2017). To evaluate word embeddings, we developed three language models using three data sources. This approach focuses on evaluating biomedical-domain and general-domain language models in the task of fall detection in health records.

Both Word2Vec and FastText are context-free representations of the words. The following list presents the data sources used to build each language model:

- WIKI: A simple language model built with Portuguese articles from the May 2019 dump of Wikipedia-PT. This corpus has a total of 250 million tokens. The model was trained with 300 dimensions per word and a minimum word count of 10 (ten).
- NILC: They are pre-computed language models that feature vectors generated from a large corpus of Brazilian Portuguese and European Portuguese, from varied sources and genres. Seventeen different corpora were used, totaling 1.3 billion tokens (Hartmann et al., 2017).

- EHRs: We used 24 million sentences with 603 million tokens from the hospital clinical notes extracted from electronic health records. The generated model has 300 dimensions per word and contains words with a minimum of 100 occurrences. This model resulted in 79,000 biomedical word vectors used as a semantic model in the neural network below.

3.2.4 Neural Networks

In this section, we present the neural networks used for the two supervised-learning tasks that utilize a pre-trained model (downstream tasks) evaluated in this study. In both neural networks, we used the FLAIR framework (Akbik et al., 2019), developed in PyTorch[6]; it has all the features of parameter tuning for model regulation. All our experiments were performed on Google Colab[7]. Deep learning algorithms are extensively used in biomedical language processing tasks (Jiang et al., 2015).

Both neural networks in our experiments use Recurrent neural network (RNN) topologies for the classification task. RNNs are demonstrably an effective approach to language modeling, both in sequence labeling tasks such as part-of-speech tagging, as well as in sequence classification tasks such as sentiment analysis and topic classification. Moreover, bidirectional RNNs have also proven to be quite effective for sequence classification. In a simple recurrent network, the hidden state at a given time t represents everything the network knows about the sequence up to that point in the sequence (Jurafsky and Martin, 2014).

A Bi-RNN consists of two independent RNNs, one where the input is processed from the beginning to the end, and the other from the end to the beginning. We then combined the outputs of the two networks into a single representation that captures both the left and right contexts of an input at each point in time. Bidirectional RNNs have also proven to be quite effective for sequence classification. As a result, during the backward pass of training, the hidden layers are subject to repeated multiplications, as determined by the length of the sequence (Jurafsky and Martin, 2014).

To address these issues, more complex network architectures have been designed to explicitly manage the task of maintaining relevant contexts over time. Long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) divide the context management problem into two sub-problems: removing information no longer needed from the context and adding information likely to be needed for later decision-making. When receiving a sentence, this network topology concatenates information in both directions of the sentence. This makes the network have a larger context window, providing greater disambiguation

[6]https://pytorch.org
[7]https://colab.research.google.com

of meanings and more accurate automatic feature extraction engineering (Hochreiter and Schmidhuber, 1997).

Clinical Note Classification (CnC): Neural network algorithms are often associated with word vector representation. In our experiments, we used a deep learning algorithm to classify the notes: word embedding representations with a recurrent neural network (RNN) called LSTM (Long Short-Term Memory Network). RNNs are modifications of feed-forward neural networks with recurrent connections. In our experiments, we used the FLAIR implementation: an open-source framework for state-of-the-art NLP (Akbik et al., 2019). The CnC is trained to classify the entire clinical note with non-fall or fall events.

Token Classification (TkC): As (Akbik et al., 2018), we used a traditional *sequence labeling* to learn how to detect falls at the token level. Our *sequence labeling* is the product of training a BiLSTM neural network with a final CRF layer for token labeling. Bidirectional Long Short-Term Memory (BiLSTM) networks have achieved state-of-the-art results in NLP downstream tasks, mainly for sequential classifications (Jiang et al., 2019, Straková et al., 2019, Peters et al., 2018b). The TkC is trained to classify each word (token) with non-fall or fall events.

3.2.5 Experiment Datasets

We divided our data into two parts: *Development-Dataset* and *Evaluation-Dataset*. The *Development-Dataset* is the dataset that we used to train the two neural networks (CnC and TkC). Once these models were trained, we evaluated them a second time with the *Evaluation-Dataset*. The *Evaluation-Dataset* was used to investigate whether the resulting models are capable of making quality classifications in texts that were not part of the model training process.

Development-Dataset: First, a corpus to identify fall events was manually annotated. The corpus is formed by 1,078 clinical notes with 1,402 sentences, where 441 fall events were identified. The data came from a large public tertiary hospital in Southern Brazil.

We adapted the structure of the dataset to train the two approaches proposed in this study: TkC and CnC. We first converted the original corpus to the CoNLL format (Sang and Erik, 2002), where each token receives a tag (see Table 3.4). The tag *B-FALL* indicates the first token of the fall event. *I-FALL* indicates tokens subsequent to the indicated event. The letters *B* and *I* are part of the BIO notation, common in name entity recognition corpora. They indicate the (B)eginning of the entity, the tokens that are (I)nside the entity, and those that are (O)utside the entity.

Then, we had to adapt the data in the CoNLL format to the *Note-Label* format in order to train the CnC. The *Note-Label* format is structured with one sentence per line and its respective binary label. Thus, we have the same pieces of data in two formats: CoNLL for the TkC and *Note-Label* for the CnC.

We split the corpus into Train, Development, and Test, as is usually done, corresponding to a proportion of 80%, 5%, and 15%. This resulted in 1,122 training sentences, 70 for development and 210 for testing.

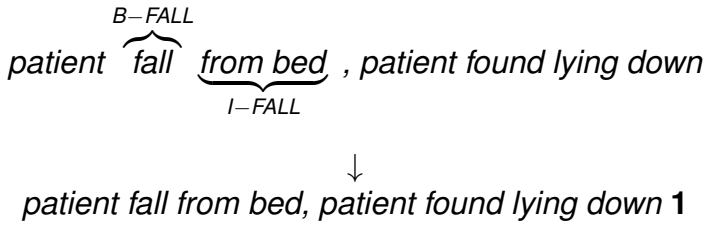

Figure 3.1 – BIO Annotation and Format of the Note Classification

Table 3.4 – Example of a Corpus Used to Train the *Token Classification*

Original Token	**Token Translation**	**Labels**
Paciente	Patient	O
refere	reports	O
tosse	cough	O
.	.	O
Relata	(he/she)Reports	O
ter	having	O
escorregado	slipped	B-FALL
em	on	I-FALL
a	the	I-FALL
escadinha	stairs	I-FALL
e	and	I-FALL
caiu	fell	I-FALL
.	.	O

Evaluation-Dataset: this is an additional dataset, annotated by nursing students, to evaluate the predictive models resulting from training the neural networks. That is, the data did not participate in the model training process and was annotated regarding the whole clinical note, not the fall event tokens. The evaluation process consists of classifying the clinical notes using the model trained with the *Development-Dataset* and comparing them with the annotations. This dataset contains 2,390 clinical notes with a binary annotation: if the note features a fall event, it receives the label **1**, if not, it receives **0**.

Table 3.5 – Dataset Statistics

Development-Dataset *number of sentences*	Train	1,122
	Development	70
	Test	210
	Total	1,402
Evaluation-Dataset *number of clinical notes*	Class 0	1,913
	Class 1	477
	Total	2,390

3.2.6 Data Sharing

The Development-Dataset with 1,078 clinical notes and the algorithms are provided for replicability purposes on the GitHub Page[8] of the project. This dataset contains 1,402 sentences from Hospital Nossa Senhora da Conceição.

3.2.7 Evaluation of the Classification Task

Figure 3.2 illustrates the flow of the experiments. The upper rectangle features the Token Classification (TkC), with the training corpus in the CoNLL format as the input for the BiLSTM-CRF neural network. At the end of the training, a predictive model is generated. In the lower rectangle, we present the flow of the Clinical Note Classification (CnC), starting with the input of the corpus to the LSTM network, followed by the generation of the predictive model (blue box). After that, we move on to our evaluation algorithm, which receives two predictive models (TkC and CnC) to predict whether or not there is a fall event in the clinical notes of our evaluation corpus.

Our goal with the assessment was to identify which is the best fall classifier. We evaluated the models as binary classifiers — if a model predicts 1, the notes contain a fall event; otherwise, it returns 0. The Clinical Note Classification model (CnC) is a binary classifier. However, the sequence labeling model Token Classification is not, since it treats notes as a classification of tokens using the BIO notation. Thus, our algorithm identifies if there are any labels (*B-FALL* or *I-FALL*) in the notes (predicted by the TkC model) indicating a fall, labeling them 1 if the note contains a fall event or 0 if it does not.

At the end of the classification task, we have three labels for each clinical note: one containing the classification label (predicted by the CnC model); another with the token classifier label (TkC labels) converted to the binary classification; and the correct label (gold

[8] https://github.com/nlp-pucrs/fall-token-classifier

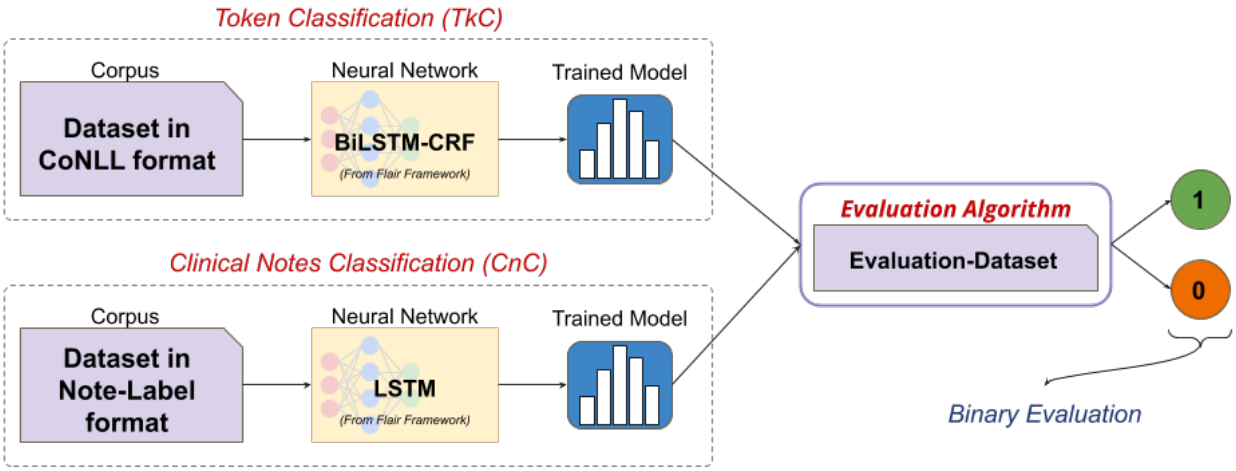

Figure 3.2 – Evaluation pipeline

standard). We used the Sklearn (Pedregosa et al., 2011) package to extract the Precision, Recall, and F_1-Measure metrics.

3.3 Results

Altogether, we performed six experiments identified in the 'Models' column in Table 3.6. All models were trained with the *Development-Dataset*. The results regarding this corpus refer to 210 separate notes for testing. Once these models were trained, we evaluated them a second time with the *Evaluation-Dataset*.

Regarding the results of the TkC and CnC tasks using the *EHR* language model for the target Class **1**, we noticed that in both corpora the TkC obtained better results, indicating its predictive capacity. Moreover, the CnC is unable to generalize its learning to other pieces of data that did not participate in the training using *EHRs*. The small difference between results with the *Development-Dataset* and the *Evaluation-Dataset*, using the *EHR* language model, shows that the TkC did not suffer from *overfitting* and *underfitting*.

When analyzing the results using the *NILC* and *WIKI* language models, both classifiers performed similarly (considering F-measure). The CnC only performs better for the target Class **1** in the *Evaluation-Dataset* using the WIKI language model.

To further analyze the results, we calculated the frequency of notes in a token range. That is, given a note s of size m and a list of intervals $\mathcal{L}$, we found $I \in \mathcal{L}$, such that $I := [I_{start}, I_{end})$. Moreover, we counted how many times the model made a correct prediction. Tables 3.7 and 3.8 show the intervals ($\mathcal{L}$), the number of notes in the class, and the number of correct predictions for the CnC and TkC models in their respective binary classes. For example, in Table 3.8, class $I = [492, 587)$ has 23 notes that have a minimum of 492 tokens

Table 3.6 – Experiment Results

Models	Classes	Development-Dataset			Evaluation-Dataset		
		Prec.	**Rec.**	**F-Measure**	**Prec.**	**Rec.**	**F-Measure**
CnC-EHR	Class 0	0.75	0.85	0.80	0.92	0.67	0.78
	Class 1	0.59	0.42	0.49	0.37	0.76	0.49
CnC-NILC	Class 0	0.93	0.88	0.90	0.93	0.93	0.93
	Class 1	0.78	0.86	0.81	0.72	0.73	0.73
CnC-WIKI	Class 0	0.96	0.91	0.93	0.95	0.95	0.95
	Class 1	0.83	0.91	0.87	0.78	0.79	0.79
TkC-EHR	Class 0	0.97	0.98	0.98	0.96	0.96	0.96
	Class 1	0.96	**0.94**	**0.95**	0.85	**0.84**	**0.85**
TkC-NILC	Class 0	0.90	0.99	0.95	0.91	0.98	0.94
	Class 1	**0.98**	0.78	0.87	**0.89**	0.61	0.73
TkC-WIKI	Class 0	0.93	0.99	0.96	0.91	0.97	0.94
	Class 1	0.97	0.86	0.91	0.84	0.61	0.71

Table 3.7 – Frequency of Clinical Notes by Token Range (*Development-Dataset*)

Classes ($\mathcal{L}$)	Number of CNs		Correctly Predicted Labels			
			TkC		CnC	
	0	**1**	**0**	**1**	**0**	**1**
$[27 - 112)$	55	38	55	31	41	24
$[112 - 197)$	41	22	41	16	30	13
$[197 - 282)$	20	5	19	4	16	2
$[282 - 367)$	9	1	8	1	8	0
$[367 - 452)$	4	2	4	2	3	1
$[452 - 537)$	6	0	6	0	6	0
$[537 - 622)$	3	0	3	0	3	0
$[622 - 707)$	1	1	1	1	1	0
$[707 - 792]$	2	0	2	0	2	0

and a maximum of 586 tokens. It is important to note that $min(I) = I_{start}$ and $max(I) = I_{end} - 1$, denoted here by the symbol '$\vdash$'. In contrast, in the final class $I = [1157, 1252]$, there is only one note that has a minimum of 1,157 tokens and a maximum of 1,252 tokens, because, in this case, $max(I) = I_{end}$, denoted by the symbol '$\dashv$'.

Based on Tables 3.7 and 3.8, we noticed that the CnC loses the ability to predict falls as the size of clinical notes increases in the case of the *Development-Dataset*. But in the case of the *Evaluation-Dataset*, the model rated almost every note as **0**, resulting in the

Table 3.8 – Frequency of Clinical Notes by Token Range (*Evaluation-Dataset*)

Classes ($\mathcal{L}$)	**Number of CNs**		**Correctly Predicted Labels**			
			TkC		**CnC**	
	0	**1**	**0**	**1**	**0**	**1**
$[17 - 112)$	930	283	912	190	921	1
$[112 - 207)$	591	128	573	87	583	0
$[207 - 302)$	281	41	264	29	280	0
$[302 - 397)$	56	11	54	8	56	0
$[397 - 492)$	22	6	21	3	22	0
$[492 - 587)$	19	4	18	1	19	0
$[587 - 682)$	8	3	8	3	8	0
$[682 - 777)$	4	1	4	1	4	0
$[777 - 872)$	1	0	1	0	1	0
$[872 - 1157)$	0	0	-	-	-	-
$[1157 - 1252]$	1	0	1	0	1	0

identification of only one fall event. The data in Tables 3.7 and 3.8 also shows that the TkC maintains its ability to identify falls as the clinical notes grow in size.

In addition to the resulting metrics, we also analyzed some of the classifications made by the TkC and CnC. We selected some fragments of clinical notes, as can be seen in Table 3.9. The first column presents the original fragments in Portuguese and the second column features the fragments translated into English. In both columns, the bold text means which tokens the TkC has classified as a fall event. The CnC has no token classification. Underlined texts are tokens annotated manually by nursing students.

In the sequence, the columns 'TkC', 'CnC', and 'Class' represent, respectively, the predictions made by the TkC and CnC and the correct label. In the first example, the TkC model can identify exactly the tokens that represent a fall event, but the CnC, in this case, says there is no fall event. The same is true in the second example. In both examples, there is no direct mention of the word *Queda* (in English, *Fall*). This makes it more difficult to correctly classify the event; nonetheless, the TkC model was able to identify the fall. The third example shows a clinical note that was not correctly identified by the TkC. Although there was an explicit word — *Queda* (in English, *Fall*) — in the note, the CnC was able to correctly identify and classify it. In the last example, both models were wrong about the predictions.

Table 3.9 – Fragments of Clinical Notes and Predictions Made by the TkC and CnC Models

Original Notes in Portuguese	Translation into English	TkC	CnC	Class
[...] Paciente informa que **escorregou no chão** do banheiro e acabou batendo a região lateral esquerda da cabeça [...]	Patient reports that he **slipped on the bathroom floor** and ended up hitting the left side of his head	1	0	1
[...] Paciente relata que quando estava no banheiro as muletas **escorregaram caindo sentado** no piso [...]	Patient reports that when he was in the bathroom, his crutches **slipped and he fell sitting** on the floor	1	0	1
[...] Refere dor no pé direito. Teve queda com trauma [...]	He reports pain in his right foot. He had a fall with trauma.	0	1	1
[...] Teve queda no banheiro, encontramos paciente no chão. Refere ter batido a cabeça na parede. [...]	There was a fall in the bathroom, we found the patient on the floor. He said he hit his head against the wall.	0	0	1

3.4 Chapter Conclusion

Results of this study point to the validity and feasibility of the classification method to detect fall events in clinical notes. We were able to discern fall incidents with minimal error using natural language processing (NLP) features, without the need for specialized software to process the texts in this dataset.

Biomedical-domain word embeddings (EHR-Notes) prove to be the best language model for fall detection. Despite this result, general-domain NILC could also be a proper alternative in datasets with a lower density of clinical notes (not enough text to train word vectors).

In the next chapter, we evaluate the potential of unsupervised learning in hospital environments. The algorithm uses historical data to assist pharmacists in the prioritization of medication reviews. This approach enables machine learning to find outlier prescriptions.

4. PRESCRIPTION PRIORITIZATION USING UNSUPERVISED LEARNING

Hospital pharmacy tasks vary from medication dispensing, administrative work, discussion of clinical cases, and clinical pharmacy, among others (Doloresco and Vermeulen, 2009). An important activity performed by the clinical pharmacy department is medication review, which aims to improve patient outcomes and reduce adverse events. One way of achieving these goals during this process is reducing the prescription errors, commonly present in the hospital environment (Bond and Raehl, 2007).

Due to the large number of prescribed medications in a hospital, clinical pharmacists must prioritize prescription reviews for patients with potentially more prescription errors or critical conditions. The prioritization task uses several risk factors for the early detection and prompt management of high-risk patients in clinical settings. Risk factors include drug-related risks (e.g., drug-to-drug interactions) and patient-related risks (e.g., acute kidney injury) (Alshakrah et al., 2019).

A study performed by Ashcroft (Ashcroft et al., 2015) revealed that errors in prescribing are a common issue in the healthcare process, affecting around 9% of all medication orders. Although not all of these errors put patients' lives at risk, a harmless error could lead to undesirable side effects and affect patients' confidence in their medical treatment. Along these lines, a global campaign to prevent medication errors was launched by the *World Health Organization* to highlight the importance of this subject for the quality of healthcare (Sheikh et al., 2017). This particular campaign aims to significantly reduce the indices of severe and harmful medication errors in the next five years.

In this study, we aim to reduce medication errors by ranking the prescriptions based on their rareness. We used past data from electronic records to map the most common way physicians prescribe each medication. We propose an unsupervised algorithm based on graph models to automatically learn the threshold between normal and abnormal doses for each medication in electronic medication orders, highlighting potential misuses. The node centrality score is one of the features used to solve this task (Akoglu et al., 2014). This context-aware characteristic is crucial since there are many different prescribing practices in the world (Baldwin et al., 2012). To the best of our knowledge, there is no previous study addressing the automatic detection of wrong dosages and frequencies (posology) for medications in electronic prescriptions.

4.1 Related Work

Electronic and manual prescriptions have been compared by researchers, and it has been found that electronic records alone may not reduce prescription errors. More advanced systems with posology checks are needed to mitigate potentially harmful errors (Gandhi et al., 2005). However, to the best of our knowledge, electronic prescriptions have not been exploited to detect medication errors concerning dose and daily frequency. Studies mostly focus on mitigating prescription errors by creating better systems that, for example, calculate the right unit for the medication (Okanda and Kanyaru, 2014) or that reduce medication errors by using patient histories (Agrawal, 2009, Kopp et al., 2006).

(Park et al., 2017) created a probabilistic graphical model to extract patterns from large prescription data and then showed divergent patterns in prescriptions for the same diseases. (Nangle et al., 2017) employed electronic prescription messages to extract the quantities, units, and frequencies of drug doses from freely-typed texts, with the aid of natural language processing techniques. Alternatively, drug information from several sources is used to avoid drug side effects in prescriptions. Reps et al. developed a system that combines different methods and sources to provide side effect alerts when prescribing medications (Reps et al., 2014).

In terms of detecting prescription errors, three previous studies mitigated this issue. (Hauskrecht et al., 2013) used historical EHR data (e.g., laboratory tests, medication orders, and procedures) to develop a system that is able to detect outlier actions for a given patient. One of the actions the alert system identified is a possible mistake regarding the medication order. Besides, (Rash-Foanio et al., 2017) used the patients' historical medical orders and diagnostic claims as data to detect look-alike/sound-alike medication errors. Both studies focus on possible medication mistakes considering only the name of the medication itself, not taking dosage and frequency errors into account.

The only work addressing dosage outliers used a commercial piece of software that tackles the problem to detect prescription errors (Schiff et al., 2017). This approach evaluated medication dosage outliers using a machine-learned dosage distribution of the medication in the population and/or the patient's history. The paper, however, does not detail the techniques. In our work, we propose an unsupervised method that uses graph centrality to detect potential prescription errors regarding posology (prescribed dose and frequency) based on data on prescription history.

Graph models have been employed for outlier detection in other domains. For instance, (Muller et al., 2013) employed centrality algorithms to rank nodes based on their centrality index to distinguish them between inliers and outliers. The same idea was successfully adopted, for instance, to rank textual information of documents and automatically create textual summaries that contain the most common words based on their centrality on

a graph (Woloszyn et al., 2017a, Woloszyn et al., 2017b). Another approach in the graph centrality field used a random walk on a graph to perform outlier detection (Moonesinghe and Tan, 2008). It relied on computing the node similarity and the number of shared neighbors between nodes. Afterward, they used a Markov chain model to compute the score for each node of the graph. Nodes with lower scores were considered outliers.

The DDC-Outlier is an unsupervised method to detect potential prescription errors regarding posology based on data on prescription history. Our approach is context-aware as it uses hospital historical data to perform outlier prediction. In the next section, we introduce the Density-Distance-Centrality (DDC) outlier algorithm.

4.2 Outlier Detection in Prescriptions

This section details the experiments performed in this work to evaluate the outlier detection in prescriptions. First, we describe the proposed algorithm and the intuition behind it. Then we describe the dataset used in the experiments and the pre-processing of the data.

4.2.1 Materials and Methods

The intuition behind the Density-Distance-Centrality (DDC) is that the detection of medication outliers can be regarded as the problem of finding groups of low-density and low-similarity prescriptions among other prescriptions. Low density, in this sense, is an uncommon prescription, rarely prescribed historically; and low similarity is a prescription like any other regarding posology (prescribed dose and frequency). To solve this problem, our approach relies on the concept of graph centrality to rank prescriptions according to their centrality index. Overdoses or underdoses are probably prescriptions whose centrality score lies below a mean centrality index for each medication.

The main step to compute prescription outliers is how to represent each prescription in the vector space. Here we fit each prescribed medication into a bi-dimensional vector with the daily posology. Since each medication/presentation is used in its own way, regarding dosage and frequency, the graph is built considering each medication/presentation (e.g., Omeprazole 20 mg and Omeprazole 40 mg dispersible tablet belong to different graphs).

We represent the relationship between prescriptions as a graph, in which the vertices are the prescriptions (e.g., Omeprazole 20 mg: 40 mg twice a day) and the edges are defined in terms of the similarity between a pair of prescriptions. We define the similarity function as the pairwise similarity between the bi-dimensional vectors (dose and frequency). The pairwise similarity accepts any pairwise metric, discussed in Section 4.2.3. We hypoth-

esize that a normal prescription has a high centrality index since it is similar to many other prescriptions.

Let P be a set of prescriptions for a specific medication and $p \in P$ a tuple $\langle d, f \rangle$, where $p.d$ represents the dose of the medication and $p.f$ the daily frequency that the physician prescribed for their patient. First, the DDC builds a distribution list for this medication D counting the frequency of the tuple $\langle d, f \rangle$. Then, the DDC builds a graph representation $G = (V, E)$, where V is the unique posology (dose and frequency) and E is the set of edges that connects pairs $\langle u, v \rangle$ where $v, u \in V$. In the next step, it uses Weighted PageRank to calculate the centrality scores for each vertex. Finally, considering the mean centrality score as the outlier threshold, the DCC generates the outlier list, where each prescription above or equal to the threshold is assigned as an inlier and all those below are considered outlier prescriptions.

Algorithm 4.1 - DDC-Outlier Algorithm (P, α): O

- Input: a set of prescribed medications P, the frontier threshold α.
- Output: list O containing the computed outlier value for each prescription $\in P$, 1 for inlier and -1 for outlier prescriptions.

```
1:  D ← 0
2:  for each p ∈ P do
3:      D[p.d, p.f] ← +1
4:  end for
5:  for each u, v ∈ D do
6:      W[u, v] ← similarity(u, v)
7:  end for
8:  C ← WeightedPageRank(W, D)
9:  Ē ← mean(C)
10: for each p ∈ P do
11:     if C[p.d, p.f] ≥ Ē * α then
12:         O[p] ← 1
13:     else
14:         O[p] ← −1
15:     end if
16: end for
17: Return O
```

The pseudo-code of the DDC is displayed in Algorithm 4.1, where G is represented by an adjacency matrix W. In the remainder of this section, we detail the process to obtain the centrality index for each posology.

4.2.2 Weighted PageRank Centrality

To compute the centrality of each prescription, the DDC relies on PageRank (Page et al., 1999), which considers each edge as a vote to determine the overall centrality score of each node in a graph. However, as in many types of networks, not all relationships are considered of equal importance. The premise underlying PageRank is that the importance of a node is measured in terms of both the number and the importance of the vertices it relates to.

In one extension of PageRank, the algorithm takes into account the importance of both the inlinks and the outlinks of the nodes (Xing and Ghorbani, 2004). In another extension, the authors of PageRank adopted a more realistic and less democratic stance by using a better (and more flexible) perturbation matrix, where the "personalization" vector $v^T > 0$ is a probability vector that allows non-uniform probabilities of teleporting to particular pages (Langville and Meyer, 2005).

Our approach uses both extensions — weighted links and weighted nodes — to compute the centrality score of prescriptions. The Weighted PageRank function is given by:

$$WPR(u) = \sum_{v \in B_u} W(v,u) \frac{WPR(v)}{N_v} \tag{4.1}$$

where B_u is the set containing all neighborhoods of u and N_v represents the number of neighborhoods of v. Besides, $W(v,u)$ is the weight of the outlink from v to u.

The intuition behind using PageRank in the DCC is that the more a prescribed medication is connected to prescriptions that are highly similar to other prescriptions, the more representative it is in the distribution of prescriptions.

4.2.3 Pairwise Metric

The DDC-Outlier algorithm could be used with any pairwise metric. The intuitive metric to gather similar prescribed medications is the cosine similarity. In our experiments, we also analyzed other metrics to evaluate our algorithm to detect prescription outliers.

- DDC: cosine similarity between instances;
- DDC-C, Cosine: cosine distance between instances;
- DDC-J, Jaccard: a statistic used to compare the similarity and diversity of sample sets;

4.2.4 Source of Prescription Data

The dataset was obtained from Hospital Nossa Senhora da Conceição (HNSC). The database contains 240,000 Computerized Physician Order Entries (CPOE) entered between January and September 2017. All records concern prescriptions, with 2 million medications prescribed to 16,000 patients. Most patients were born between the 1950s and the 1990s and were treated through the Brazilian public healthcare system.

HNSC belongs to the public healthcare system in Brazil, and, as a standard procedure, the hospitals in the public hospital environment always use generic names for medications (regardless of the brand acquired, the name registered in the system is the name of the active principle).

Each prescription record has the following information: the patient's register number, date of prescription, name and presentation of the medication, dose, route, frequency, and a free-text comment field. In our experiments, we only used the posology (dose and frequency) of each medication/presentation. In the next section, we detail the pre-processing tasks performed over this data.

4.2.5 Pre-Processing of Prescriptions

We noticed that there was noise in the raw data provided by the CPOE, requiring data cleansing to ensure good data quality in the experiments. Therefore, we performed the four pre-processing tasks described below.

Unit Table

Some prescribed medications are presented in milligrams and others, in grams. To avoid errors in the quantity, we set a standard unit for all medications. Then we listed all dose units used by the physicians in the CPOE and defined a factor to multiply the non-standard units.

Frequency Table

In the CPOE, each physician describes the frequency in different ways. We standardized it to a daily frequency: 3 times a day (3), once a day (1), 6 times a day (6). This included terms such as “twice a day" (2), “6h/6h" (4), and “3x/day with a meal" (3).

Dosage Table

For validation purposes, we included the daily maximum and minimum doses for 345 prescribed drugs in the dataset, following two evidence-based references: Micromedex® (Solutions, 2017) and UpToDate® (LLC, 2017). Medications with no daily maximum and minimum doses were not included in the experiments.

Medication Pruning

Following the methodology proposed by (Emmott et al., 2013), medications with less than 1,000 records were discarded. Besides, when the number of candidate anomalous data points was small, we excluded the medication. We chose to discard medications with less than ten outlier observations due to insufficient support.

4.2.6 Prescriptions Stats

After performing all the described pre-processing tasks, 51 medications remained in our dataset, with a total of 563,000 records. Each medication has a particular distribution of prescriptions: some of them are used only for disease treatments; others, only to reduce symptoms; while others are intended for prophylactic use. These characteristics allow a wide evaluation of the proposed algorithm regarding a variety of prescription scenarios. In Table 4.1, we show the overall statistics concerning the prescriptions issued at HNSC.

Table 4.1 – Prescription Dataset

Total Prescribed Medications	563,171
Overdosed Prescribed Medications	6,666
Underdosed Prescribed Medications	4,868

The large number of prescriptions (1,000 per day) hinders the screening process performed by the department of pharmacy services. Despite their efforts, there were more than 6,000 overdosed and almost 5,000 underdosed prescriptions in 2017. This scenario shows the importance of an automated system that is able to identify prescription outliers.

Figure 4.1 shows an example of the resulting data on the drug Acyclovir tablet. This reveals a typical distribution, featuring some of the most common uses of this medication in the hospital, such as 200 mg three times a day and 400 mg three times a day. Besides, there are unusual prescriptions, such as 400 mg Acyclovir six times a day. The only points that are considered outliers for Acyclovir are the red ones, with prescriptions for 800 mg six times a day as an overdose and 200mg once a day as an underdose.

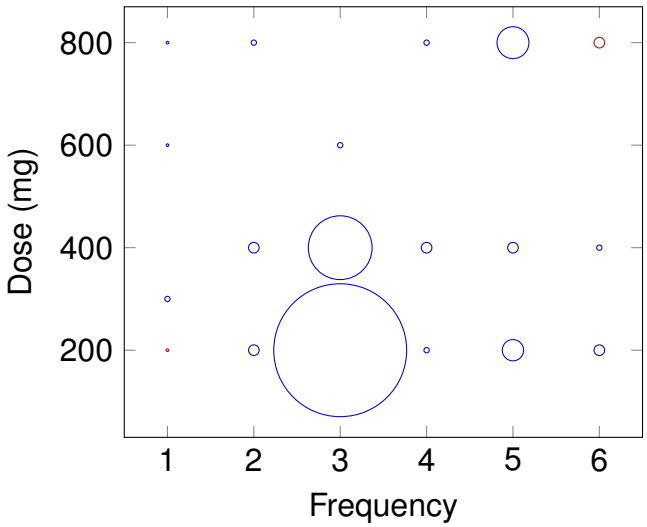

Figure 4.1 – Prescription of Acyclovir 200 mg tablet, where diameter size means how many times this pair (dose, frequency) is prescribed. Blue points mean normal prescriptions and red points indicate outliers.

All infrequent dosages of the prescribed medication, represented by small circles, could be potential medication mistakes and could be highlighted by the department of pharmacy services for additional verification.

4.2.7 Data Sharing

All the content of the work (algorithm, sample dataset, and experiments) is available on the GitHub Page[1] of the project in order to be easily replicated. The sample dataset has no patient data; it contains only a subset of the medication dataset (150,000 real prescriptions), with information on the dose, frequency, overdose, and underdose.

In the next section, we detail the experiments using the information on the maximum and minimum doses concerning the medications.

4.2.8 Experiments

In order to evaluate the performance of our proposed approach, we designed a task for each medication in the prescription dataset. The task consisted of identifying overdoses, that is, medications prescribed above the maximum daily dose specified in the literature, and underdoses, when the dose is below the lowest daily dose in the literature (regardless of indication).

[1] https://github.com/nlp-pucrs/prescription-outliers

Our experiments followed the methodology proposed by (Emmott et al., 2013):

- Top-3 rankings: it shows the number of medications in which each algorithm appeared in the top-3 algorithms when ranked by F-measure;
- Parameter search: all analyzed algorithms require a threshold parameter. We employed parameter search to find the best parameters to maximize the hits (as described below, in Section 4.2.11). In all cases, we made a good faith effort to maximize the performance of all methods.

4.2.9 Baseline

To evaluate our approach, we selected several state-of-the-art unsupervised methods as the baselines used to detect outliers (Han et al., 2011), as explained below:

- **One-Class SVM**: it was introduced by Schölkopf et al. It requires the election of an SVM kernel and a scalar parameter to define the frontier of outlier instances. Here we chose the RBF kernel, which better fits our experiments (Schölkopf et al., 2001).
- **Local Outlier Factor**: it computes a score reflecting the degree of anomalous instances. It measures the local density deviation of a given data point with respect to its neighbors. The idea is to detect the samples that have substantially lower densities than their neighbors (Breunig et al., 2000).
- **Gaussian Mixture**: it is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters (Agarwal, 2006).
- **Robust Covariance**: assuming that the inlier data is Gaussian distributed, it will estimate the inlier location and covariance in a robust way (i.e., without being influenced by outliers). The Mahalanobis distances obtained from this estimate are used to derive a measure of outlyingness (Rousseeuw and Driessen, 1999).
- **Isolation Forest**: it isolates the observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produces shorter path lengths for particular samples, they are highly likely to be anomalies (Liu et al., 2008).

4.2.10 Performance Metrics

The task of detecting outliers is a classification with a class imbalance problem, where the main class of interest is rare. That is, the dataset distribution reflects a significant majority of the negative class (non-outlier prescriptions) and a minority of the positive class (overdosed and underdosed prescriptions). Therefore, choosing the right performance metric is essential to correctly evaluate all methods concerning this problem. In this regard, we briefly describe the most common metrics to evaluate binary classifiers (Han et al., 2011):

- **Accuracy**: the percentage of instances labeled as the correct class (positive or negative);
- **Recall**: a measure of completeness (i.e., what percentage of positive instances is labeled as such);
- **Precision**: it can be thought of as a measure of exactness (i.e., what percentage of instances labeled as positive is actually positive);
- **F-Measure**: it corresponds to the harmonic mean between precision and recall.
- **Top-3 Rankings**: whenever the algorithm ranked among the top 3 regarding F-measure.

An algorithm that predicts that most instances belong to the negative class (not medication errors) in an imbalance problem could have a high accuracy score, but it is useless to predict the aimed positive class (medication errors). Conversely, an algorithm that predicts that all instances relate to the positive class will have a high recall, but it is unable to distinguish between positive and negative instances. For this reason, we selected F-measure as the main metric to evaluate the performance of outlier algorithms. The F-measure gives equal weight to precision and recall, both important to evaluate the task of detecting overdosed and underdosed prescriptions.

4.2.11 Parameter Tuning

All outlier algorithms analyzed were sensitive to the parameter that defines the frontier between normal and abnormal observations (Xie, 2006). Besides, each type of medication has its own distribution, making the frontier particular for each drug. To ensure the best F-measure, we performed a simple parameter search, varying it for each algorithm and each medication. In all cases, we made a good faith effort to maximize the performance of all methods.

All algorithms allow only a specific range for the frontier parameter (some call it a contamination parameter). Below we list the interval searched for each set of algorithms:

- For Local Outlier Factor, Isolation Forest, and Robust Covariance, the search ranges from 0.01 to 0.5;
- For One-Class SVM, Gaussian Mixture, and DDC, the search ranges from 0.01 to 1.0.
- For every algorithm, we applied a 0.01 step between intervals.

4.3 Results

In this section, we discuss the evaluation of the DDC with regards to the adopted baselines in terms of detecting overdoses and underdoses for all 51 medications. We also address the results of the stability of the algorithm regarding its parameter search and its run-time performance.

In Table 4.2, we show the overall results of all algorithm-detecting overdosed and underdosed outliers for the 51 medications. The DDC-J, density-distance-centrality using the Jaccard similarity, achieved the best mean F-measure and ranked among the top 3. The Isolation Forest and the DDC using the cosine similarity also achieved good results, both remaining in the top-3 ranking for more than 20 medications.

Table 4.2 – Mean Performance of Outlier Detection.

Algorithm	**Recall**	**Precision**	**F-Measure**	**Top 3** ↑
DDC-J	0.90	0.61	0.68	**31**
Iso. Forest	0.91	0.52	0.61	**26**
DDC	0.86	0.50	0.58	**21**
SVM	0.94	0.39	0.48	20
DDC-C	0.72	0.54	0.51	19
Covariance	0.60	0.37	0.39	16
Gau	0.95	0.29	0.37	13
LOF	0.87	0.38	0.44	12

When counting the best methods, the top-3 ranking discarded the F-measure when it stood at less than 0.4. Besides, all methods are counted in the top 3 when there is a tie in the third position.

Covariance, Gaussian Mixture, and Local Outlier Factor had the worst performance regarding F-measure. Furthermore, these three algorithms achieved lower results in the top-3 ranking. Since each medication has its own outlier distribution, some algorithms work better for some medications, but the DDC-J had the best overall result.

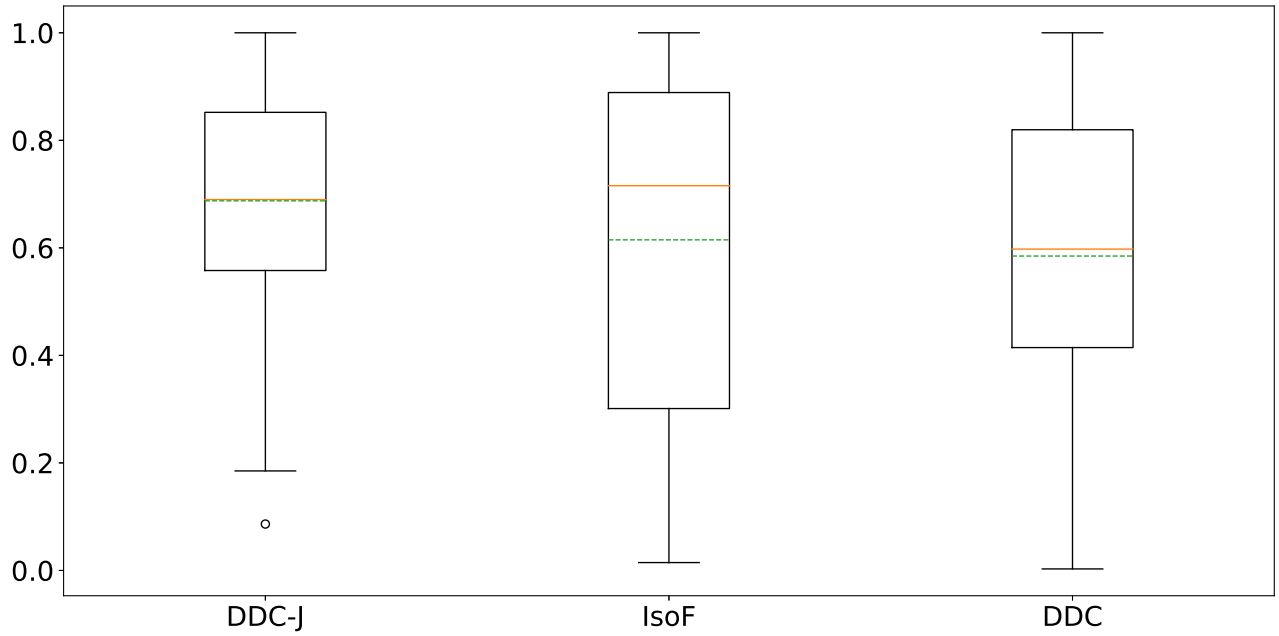

Figure 4.2 – Distribution of F-measures obtained in the top-3 algorithms for all 51 medications. Solid lines are the medians and dash lines are the means.

In Figure 4.2, we show the distribution of the F-measures for all medications in each of the top-3 algorithms in Table 4.2. The DDC-J also had lower variation in comparison with the Isolation Forest. In the following section, we cover some insights about the run-time of the algorithms.

4.3.1 Evaluation of the Run-Time

Some anomaly detection algorithms are very time-consuming when the dataset is greater than 10,000 instances. We developed an experiment to evaluate the time performance of the algorithms. We selected 2 (two) medications with 30,000 instances, ran every algorithm from 3,000 to 30,000 instances with 3,000 steps, and computed the time spent on these medications.

For the task of detecting outliers in a large historical dataset, the DDC algorithm has an important scalability property to perform this analysis. Figure 4.3 shows that the time consumption of the One-class SVM and Local Outlier Factor is exponential when the data size grows. The time in seconds is represented in a log scale on the y-axis. This experiment shows that these algorithms are not fit for big data analysis.

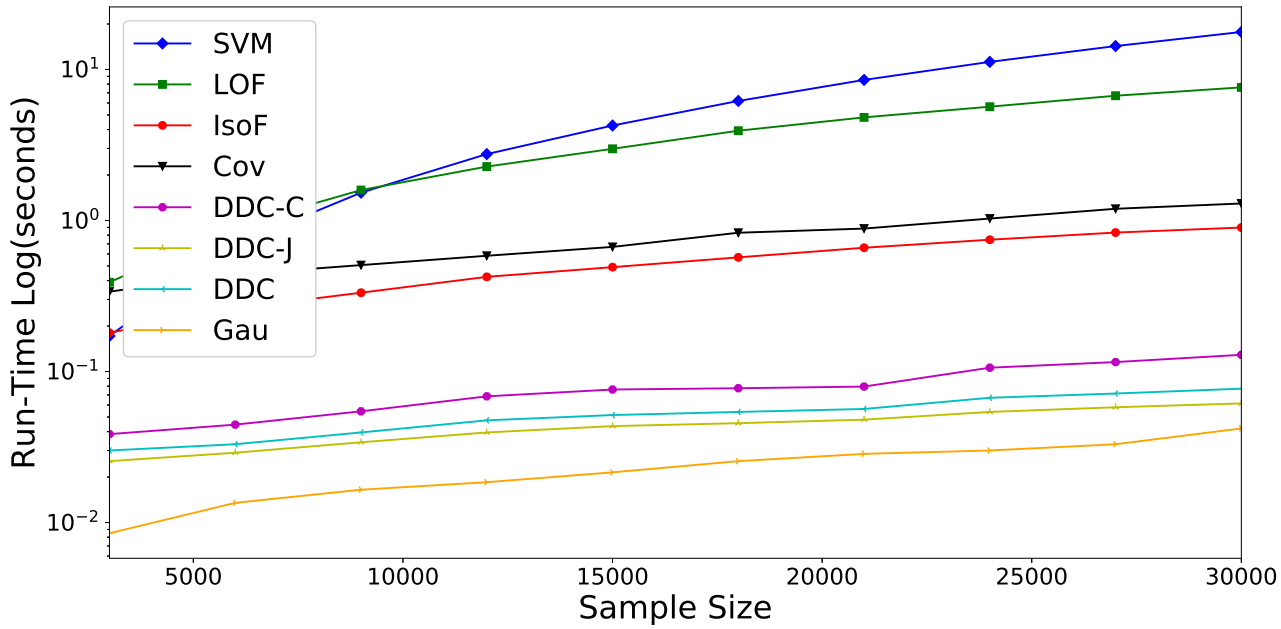

Figure 4.3 – Run-time comparison between the DDC and baselines for two medications with 30,000 instances.

4.3.2 Stability of the Algorithm

The frontier is a sensible parameter that needs to be set in outlier algorithms. With this aspect in mind, we performed an experiment to evaluate the stability of the frontier parameter for each algorithm.

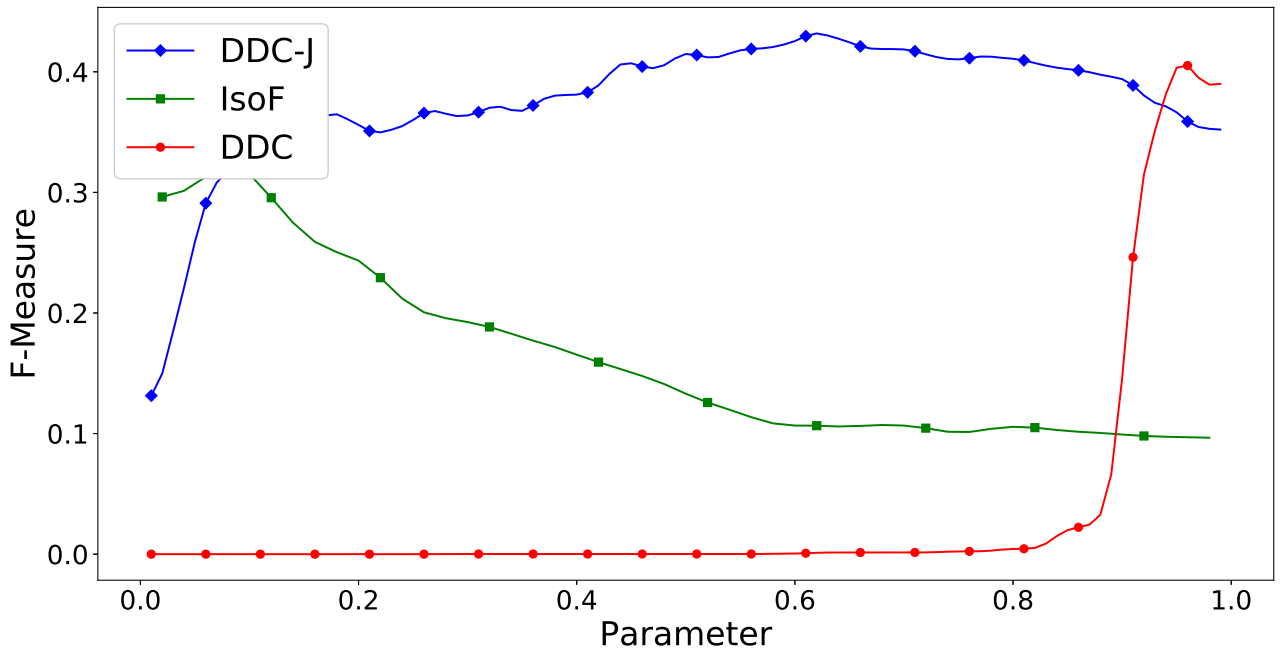

Figure 4.4 – Influence of the threshold parameter on the F-measure results for the top-3 algorithms

In Figure 4.4, we show the mean F-measure considering all parameters for each top-3 algorithm in Table 4.2. Regarding the DDC-J, it presents the lowest variation using all possible values for the α parameter. Nevertheless, when $\alpha > 0.1$, it achieves the best results in comparison to the baselines. In regard to the DDC, it achieves better results when $\alpha > 0.9$.

4.3.3 Parameter Regression Estimation

All outlier algorithms are very sensitive to their parameters for each kind of dataset. To tackle this problem, we performed a regression analysis considering several statistical information on the medication distribution to estimate the best parameter.

The following statistics were used in the regression analysis: mean, standard deviation, median, and percentile at 20, 50, and 75. All statistics were computed across dose, frequency, and both (dose and frequency). The best parameter found for each algorithm was used as the regression target.

Despite all regression algorithms evaluated and the efforts to combine specific sets of medications, the parameter obtained in the regression drastically decreased all algorithm performances regarding the F-measure in overdose/underdose experiments.

4.3.4 Quality Evaluation

Qualitative assessments were performed by a pharmacist, and the different reasons why a prescription might be outside the pattern were investigated. The F-measure considered only overdoses/underdoses as true positives. Therefore, a more thorough analysis was required to assess the other cases of outliers. The algorithms could detect other prescriptions to be improved by the department of pharmacy services. In the list below, we included a few examples of prescriptions detected by the algorithms; they were not overdoses/underdoses, but their singularity indicates other problems that should also be avoided and could be reviewed by double-checking.

- Prescriptions with more suitable presentations for the prescribed dose. It was not necessary to split tablets or even dispose of medications unnecessarily (e.g., Amlodipine 10 mg, 5 mg prescribed — Amlodipine 5 mg was available at the hospital);
- Prescriptions whose right dose is difficult to administer (e.g., Levothyroxine 100mcg, a dose of 88 mcg was prescribed);
- Prescriptions with an unusual frequency (e.g., Meropenem 2g once a day)

- Unusual frequencies, making compliance difficult (e.g., Hydralazine 50 mg 6x/day);
- Unusual doses (e.g., Allopurinol 600 mg once a day);
- Half-dose prescriptions of medications that should not be split (e.g., Hydralazine 50 mg, with a 25 mg dose prescribed).

The performance of the algorithms, in general, regarding the outlier detection of drug prescriptions with a homogeneous distribution of prescriptions (e.g., Doxazosin tablet, Enalapril 20 mg tablet — Figure 4.5), was better than for drugs with a more sparse distribution of prescriptions — dose x frequency (e.g., Potassium chloride oral solution, Carbamazepine oral suspension — Figure 4.6).

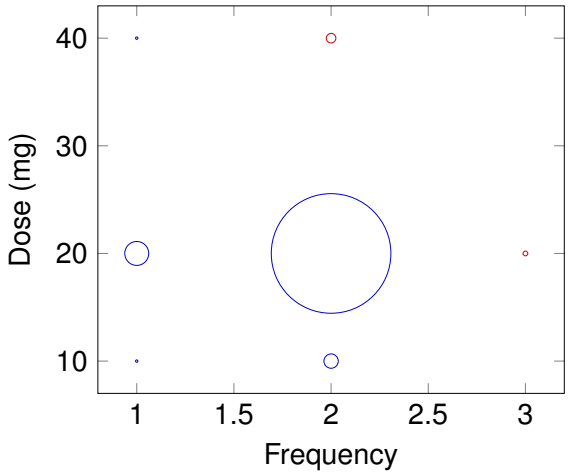

Figure 4.5 – Prescription of Enalapril 20mg cp, where blue points mean normal prescriptions and red points indicate outliers.

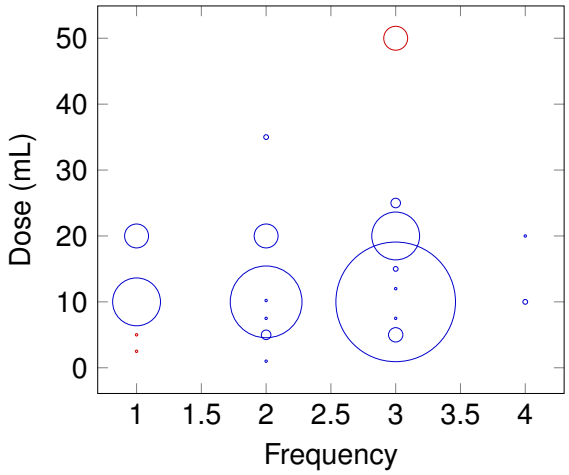

Figure 4.6 – Prescription of Carbamazepine 20mg/mL oral suspension, where blue points mean normal prescriptions and red points indicate outliers.

In addition to being better in the outlier detection of prescriptions by assessing the F-measure (with overdoses/underdoses as true positives), the DDC-J algorithm was also

able to detect more cases like those described above. Therefore, the algorithm proved to be suitable to generate warnings for double-checking, improving prescriptions, and providing greater safety for patients. The DDC-J obtained a good recall but poor precision (for cases of overdoses/underdoses); however, this characteristic allows it to select more prescriptions that are potentially incorrect or likely to be improved.

4.3.5 Limitations

Our experiments focused on the ability of the proposed algorithms to detect outlier instances among hospital prescriptions. However, any approach that seeks medication errors isolating the drug from the prescription, as we do, inherently has some limitations. For example, drug-to-drug interactions and duplicated/redundant therapy problems must use all prescription data to handle this task. Therefore, our method should be construed as part of possible mistakes in prescriptions.

Some medications are usually prescribed taking into account patient weight. In that case, the outlier algorithm should use patient information to detect relative overdoses/underdoses considering each patient's characteristics. Nonetheless, because of the simplicity of the preprocessing pipeline and the lack of information on patients' weights in the HNSC dataset, we discarded any medicines that depend on body weight.

Another limitation of our study is the outlier assumption in our experiments. All evaluation analyses set overdoses/underdoses as the only type of outlier problem in medication errors. Nevertheless, in a future experiment, an annotated dataset could be used to evaluate other kinds of outliers, such as potential errors.

Finally, the outlier detection is not intended to assess prescribing associated with diseases or clinical conditions of patients, but only to alert pharmacists of critical prescriptions that may contain an error. Non-standard prescriptions are reassessed by pharmacists, who may consider whether there is an error or only an unusual dose that is acceptable for the specific condition of that patient. It is a tool to help and to set priorities so that clinical pharmacists can focus on their activities.

4.3.6 Chapter Conclusion

A novel algorithm was developed to identify prescription outliers in electronic medical records. This algorithm can rapidly create a distribution pattern in a graph structure to detect anomalous prescriptions. A comparative experiment was conducted with five state-of-the-art algorithms to detect overdosed/underdosed prescriptions. Moreover, a qualitative

analysis indicated the proposed method with Jaccard similarity as the best approach to handle potential prescription errors.

There is a great advantage for the algorithm to learn the distribution of the institution's prescriptions. These characteristics allow the use of this method anywhere in the world with drug-specific standardization, self-adapting to the specifics of each hospital and being able to be automatically updated.

In the next chapter, we cover a real-scenario use of the DDC-Outlier. This experiment was performed in a 1,200-bed hospital with diverse patient comorbidities and profiles. Such an environment could evaluate the ability of the algorithm to improve medication review by pharmacists.

5. PRESCRIPTION PRIORITIZATION APPLICATION AND EVALUATION IN A REAL SCENARIO

In this chapter, we show how using the DDC-Outlier algorithm can help pharmacists prioritize prescriptions in a hospital environment. These findings are useful to better understand the performance of the DDC-Outlier in a real-world scenario. The pharmacists reviewed the classification of the algorithm for each posology (dose x frequency) of several medications. The primary goal of this evaluation is to measure the usefulness of the algorithm in clinical practice.

5.1 Materials and Methods

This section describes the hospital prescription dataset used in this evaluation, the changes in the algorithm's output to suit the pharmacists' needs, and the design of the experiment to evaluate the DDC-Outlier.

5.1.1 Hospital Santa Casa

We developed an evaluation in partnership with Hospital Santa Casa, from the city of Porto Alegre. Santa Casa is a hospital complex consisting of 7 hospitals, 33 departments, and 75 medical specialties, totaling 1,200 beds. The evaluation was developed in all hospital units, comprising adult, neonatal, and pediatric patients.

The data relates to the period from August to December 2020 and considered 24,702 prescriptions reviewed by the hospital's pharmacists. During this period, five pharmacists and two pharmacy students conducted medication reviews at the hospital.

Table 5.1 – Summary of the Dataset

Profile	**Prescriptions ↑**	**Medications**	**Patients**
Adult	21,810	629	1,805
Pediatric	2,139	260	137
Neonatal	756	80	108

Table 5.1 summarizes the data analyzed by the pharmacists in each patient profile. Regarding the profiles, neonatal patients vary from 0 to 2 years old; pediatric patients are aged between 2 and 17 years; adult patients are over 18 years old.

5.1.2 DDC-Outlier Score

In the previous chapter, the algorithm DDC-Outlier used only two classes as output. To present these outputs to the pharmacist, instead of considering only a binary output, we developed an outlier score scale that better classifies the prescribed medications. The score ranges from 0 to 3, where 0 (zero) is a common prescription and 3 (three) is the most abnormal prescription. This score is provided to the pharmacists when they review the medication chart, enabling them to rely on more options to review the prescribed medications. The distance metric used in these experiments was the Jaccard similarity, as it achieved better results in the outlier detection task in Chapter 4. The DDC-Outlier Score is defined by the following Equation 5.1:

$$s' = abs(round(\frac{(x - min(x))(3)}{max(x) - min(x)} - 3)) \tag{5.1}$$

For each medication, the Pagerank output is split in half using the mean measure of the PageRank index. All samples above the mean are considered inliers (most common prescription), with a score of 0. Then, Equation 5.1 is used to define the score of the outliers from 1 to 3.

The source code of the DDC-Outlier API was made available on the GitHub page [1] of the project. The code is an API Service that could be deployed and used by hospital health information systems to classify prescribed medications.

We developed a system to assist the pharmacists in the daily work of the clinical pharmacy. The following section describes the decision support system for clinical pharmacy deployed in Hospital Santa Casa.

5.1.3 Decision Support System for Clinical Pharmacy

The DDC-Outlier API can identify prescription errors, but pharmacists would not be able to use an API. To fill this gap, we developed a Decision Support System for Clinical Pharmacy that wraps the DDC-Outlier in a web user interface. We designed the system in collaboration with a clinical pharmacist and a physician. One year of medication prescription history was used to compute medication outliers.

The pharmacists can change the outlier score in the system if they think the score is not appropriate. They adjust the score to a value aligned with the best clinical practices and literature on medications. These actions were performed during the pharmacists' daily work routine. We tracked the changes to evaluate the accuracy of the DDC-Outlier Score.

[1] https://github.com/nlp-pucrs/ddc-api

In this hospital, pharmacists developed a protocol to classify the prescribed medications related to the outlier score. The protocol below was a reference to guide the evaluation of the score by the pharmacists:

- Score 0: Common prescription, according to the hospital's standard;
- Score 1: Depending on the patient's clinical condition and posology, similar to the hospital's standard;
- Score 2: Depending on the patient's clinical condition and posology, likely to be a prescription error;
- Score 3: Possible prescription error or dose above the maximum allowed by the literature;

We named the decision support system NoHarm.ai in a reference to a patient safety challenge developed by the *World Health Organization* called Medication Without Harm (Sheikh et al., 2017). The system integrates with hospital electronic health records (EHRs), allowing pharmacists to access the EHR data in real time. NoHarm.ai is an open-source software available on GitHub[2]. In the next section, we detail the pre-processing step performed in some specific medications.

5.1.4 Range- and Weight-Based Dosing

Regarding the posology, the DDC-Outlier works better when the number of possible posologies of a given medication is low and the density is high in very few of them. However, the number of possible posologies (dose and frequency) for some medications, such as liquid and weight-based drugs, is high and the density of most of them is low, thus worsening the performance of the algorithm. For instance, Dipyrone pills can be prescribed in doses of 500 mg, 1000 mg, 1500 mg, and 2000 mg and with a daily frequency that ranges from 1 to 6. Alternatively, there are several possible dosages for the morphine sulfate solution for injections: the doses range from 0.1 ml to 500 ml (with steps of 0.1 ml) and the daily frequency varies from 1 to 24.

Moreover, some medications are prescribed based on the patient's weight. In this case, the DDC-Outlier should also group the doses of weight-based medications. For instance, the weight-based protocol for Heparin suggests the prescription of 15 U/kg per hour.

With that in mind, we grouped the doses in ranges of values to reduce the number of cases and increase their density. The system enables pharmacists to choose whenever the medication should be grouped in dose intervals to reduce dose distribution or whether the medication is weight-based.

[2]https://github.com/noharm-ai

Table 5.2 – Medications with Range- and Weight-Based Dosing

Profile	Regular Medication	Range-Based (Weight-Based)	Total ↑
Adult	496	133 (127)	629
Pediatric	126	134 (128)	260
Neonatal	31	49 (48)	80

In Table 5.2, we show the distribution of medications with range-based dosing and weight-dependencies for each profile of patients in the hospital's dataset. In the pediatric and neonatal profiles, almost half of the medications had to be converted into smaller dose distributions. All medications with weight-based dosing are also considered range-based regarding their dose distributions. The weight-based dosing criterion of a given medication is found in the pharmaceutical literature, and]pharmacists use this information to configure this setting in the system. Conversely, the dose range is chosen by the pharmacists by analyzing the dose distribution of each medication. We evaluated the correlation between the dose range and several statistical information on doses, but no correlation was found so far.

5.1.5 Design of the Experiment

As explained above, pharmacists can change the score assigned to the posology of each medication in the system if they do not agree with the score assigned by the algorithm. To evaluate the performance of the DDC-Outlier Score, we used the data generated when the pharmacists changed each score. The score is assigned to the posology (dose and frequency) of each medication. When pharmacists change the score suggested by the algorithm and assign a manual score, we consider that the score of the posology is a false positive. In other words, the algorithm was not able to properly assign the score. In Figure 5.1, we show how pharmacists change the score in the system interface.

To evaluate the performance of the algorithm, we only considered medication scores that had at least one prescription reviewed by a pharmacist or that had been manually changed. This selection avoids evaluating scores that were not reviewed by the pharmacists. Besides, pharmacists can whitelist some medications — such as glucose and sodium chloride. As such, they are not reviewed because no dose is harmful to patients.

Medication	Dose	Daily Frequency	Score	Manual Score	Count
Acetaminophen Pill 500mg	1000 mg	4	0	-	90
Acetaminophen Pill 500mg	500 mg	4	0	-	26
Acetaminophen Pill 500mg	500 mg	1	2	0 ✓	11
Acetaminophen Pill 500mg	500 mg	6	2	-	10
Acetaminophen Pill 500mg	1000 mg	1	3	-	5
Acetaminophen Pill 500mg	1000 mg	3	3	-	2

Figure 5.1 – The figure shows the medication screen where pharmacists can change the scores for each posology. The column "Score" shows the value assigned by the DDC-Outlier algorithm. In column "Manual Score," the pharmacists can perform the change. The columns "Dose" and "Daily Frequency" show the posology of the medication. The column "Count" indicates the number of times the posology was prescribed in that hospital.

5.2 Results

This section discusses the evaluation of the DDC-Outlier Score. We used the manual score, assigned by the pharmacists, as a gold standard to evaluate the outlier algorithm. The scores (0, 1, 2, 3) are the classes used to determine if the algorithm matched the pharmacists' assessment. If the pharmacists did not change the score defined by the algorithm, we considered that the algorithm correctly classified the posology outlier score. In contrast, if the pharmacists changed the score, we considered that the algorithm incorrectly classified the posology.

The dataset contained a total of 3,472 posologies that were reviewed by pharmacists with an average change to the scores of 11.3%. Table 5.3 presents the number of posologies grouped according to the profiles. Additionally, the table shows the results for the F-measure and mean absolute error (MAE) metrics. The F-measure is a metric that corresponds to the harmonic mean between precision (number of instances labeled as positive that are actually positive) and recall (number of positive instances that are labeled as positive). This provides us with a metric on how well the algorithm can achieve a balance between making fewer mistakes while correctly classifying each target class. The mean absolute error is a numeric value related to the error size: the smaller the result, the smaller the distance between the true value and the predicted value. The 'Count' column shows the number of posologies evaluated for each profile. The 'Changes' column shows the percentage of scores changed by the pharmacists.

Table 5.3 – Results for Each Profile

Profile	Medications ↑	Count	Changes	F-Measure	MAE
Adult	629	2,630	12.6%	0.85	0.18
Pediatric	260	725	3.5%	0.94	0.07
Neonatal	80	117	17.9%	0.75	0.34

The overall F-measure of the DDC-Outlier in a real-world scenario surpasses 70% in all profiles. From 3,472 scores generated by the algorithm, only 386 were changed. The average mean absolute error for all profiles is 0.20, proving the stability of the algorithm among several units and medication profiles.

It is important to state that the pediatric profile had a lower number of changes than the other two profiles. This could demonstrate less attention from pharmacists to this profile. In comparison with the adult and neonatal categories, the result of pediatrics seems to be trustless. Even with more scores generated than in the neonatal category, the percentage of changes was still smaller in the pediatric profile. A percentage of changes close to 15% is expected in each profile, and pediatrics reached 3.5%. The adult profile covers more hospital units than the other profiles and totals 21,810 reviewed prescriptions. This large number of prescriptions is likely to produce more accurate results when compared to the other profiles.

Table 5.4 – Detailed Performance for Each Profile and Score

Profile	Scores	Count	Changes	Precision	Recall	F-Measure	MAE
Adult	Score 0	1,184	6.5%	0.96	0.90	0.92	0.05
	Score 1	207	14.9%	0.88	0.73	0.79	0.12
	Score 2	325	16.9%	0.86	0.81	0.83	0.22
	Score 3	914	18.5%	0.83	0.98	0.89	0.36
Pediatric	Score 0	350	0.2%	1.00	0.95	0.97	0.00
	Score 1	58	5.1%	0.94	0.88	0.90	0.05
	Score 2	109	5.5%	0.94	0.97	0.95	0.09
	Score 3	208	7.6%	0.92	1.00	0.95	0.17
Neonatal	Score 0	68	4.4%	0.98	0.85	0.91	0.01
	Score 1	8	25.0%	0.75	0.60	0.66	0.25
	Score 2	10	10.0%	0.90	0.69	0.78	0.01
	Score 3	31	48.3%	0.51	1.00	0.67	1.12

Table 5.4 details the results for each score and each profile. For all scores and profiles, the algorithms achieve an F-measure of over 65% and a mean absolute error of less than 0.40 (except for the neonatal profile, with a score of 3).

In the “Count” column in Table 5.4, we can see that the DDC splits data in half: 50% inliers (Score 0) and 50% outliers (Scores 1, 2, and 3). This approach relates to Equation 5.1, which uses the mean measure of the Pagerank index as a parameter to split data into inliers and outliers.

In addition to the performance results achieved by the DDC-Outlier Score, other analyses could be made to better understand the algorithm. Figure 5.2 shows the confusion matrix of the DDC for the scores of overall profiles.

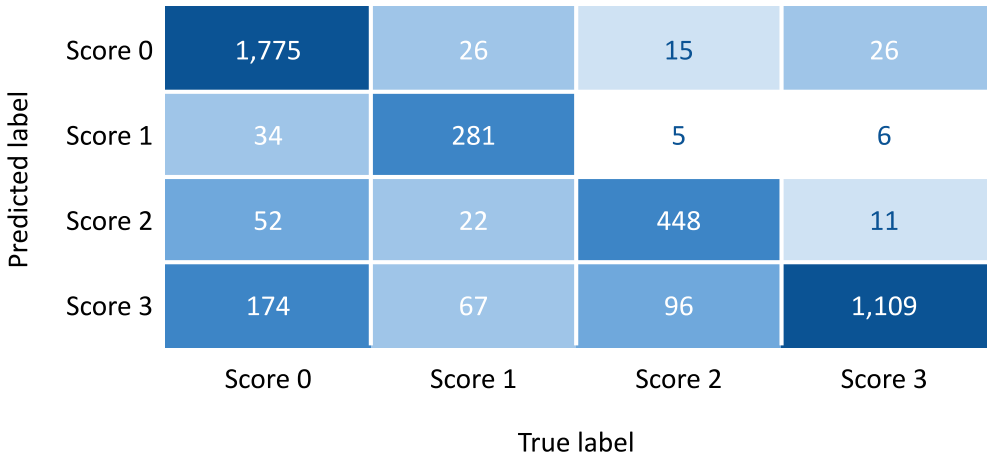

Figure 5.2 – Confusion Matrix of the Scores of Overall Profiles

The confusion matrix clearly indicates that the DDC tends to penalize posology outliers to a score of 3. Pharmacists commonly need to change a score of 3 to less critical scores such as 2, 1, and 0. Particularly in healthcare, this behavior is important as it rarely misclassifies high-scored prescriptions. For high scores, the DDC favors recall over precision, as shown in Table 5.4. In other words, the DDC-Outlier Score attempts to ensure the correct classification of the posologies with high scores even if this leads to an overestimation of posologies with lower scores.

The number of training samples is an important parameter that impacts the performance of machine learning algorithms (Beleites et al., 2013). Figure 5.3 shows the relation between the sample count, in log scale, and the F-measure of the DDC Score. After 10,000 samples (log (10K) = 4), the algorithm achieves its best result.

In the next section, we cover the journey of this real scenario experiment. We use self-reflection and writing to explore our personal experience, following a method called autoethnography (Mills et al., 2009).

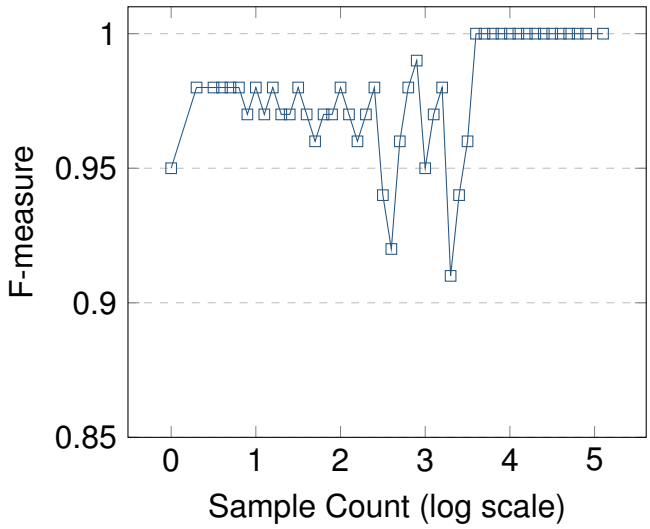

Figure 5.3 – Relation between Sample Count (log scale) and F-measure

5.3 Discussion

Our journey from research on health informatics to real-scenario applications started with the conception of the project. It was crucial to collaborate with healthcare professionals that deal with daily issues in a hospital environment. This proximity to real problems allowed us to set the objectives of the project aiming at solutions that could impact the healthcare system.

After the ethical committee approved the project, the hospital IT department promptly released a 5-year dataset with clinical notes, laboratory exams, patient demographics, and prescriptions. Then we focused our efforts on using the data to develop algorithms that solve the chosen problems.

With the good results of the project (published articles and computer science awards), we created a research group focused on solving healthcare problems using artificial intelligence. The research group was the gateway for other hospitals to join the project and for us to strengthen our relationships with other healthcare professionals.

The exchange of information with hospitals led us to understand the need to build a system, not only an algorithm, to quickly deploy the solution. Hospital Santa Casa was a great supporter of the deployment task. The pharmacy team and IT department provided a reliable environment to validate the system.

Before the Decision Support System for Clinical Pharmacy was deployed in Hospital Santa Casa, the Clinical Pharmacy department was able to validate an average of 3,200 prescriptions per month and make 140 interventions that improved patient treatment. Since the system was deployed (in April 2020), the pharmacists have been able to validate around 80,500 prescriptions and make 9,000 interventions.

The improvements in pharmacists' routines were reported in an online meeting held by the Brazilian Hospital Pharmacy Society (Guglielmi et al., 2020), in August 2020, with the staff of Hospital Santa Casa. Some testimonies given by the pharmacists are listed below:

- "The main barrier found to carry out the work of the clinical pharmacy was the time necessary to validate the prescriptions. The system we used did not provide easy access to patient information." Raquel Sinderman (39:50)
- "NoHarm presents the risks related to the prescription, summing up the score of each medication. It is able to organize the validation of the clinical pharmacy." Tatiana Hoffmann (43:08)
- "We had this difficulty: we used a lot of screens, impacting the validation time. NoHarm allows us to view a series of patient information on the same screen." Karoline Flach (45:46)
- "We have already noticed a lot of improvement in our process using NoHarm." Karoline Flach (47:57)
- "We can already see that the use of the tool is a very consistent strategy so that we can achieve the goal of validating 100% of the prescriptions in the hospital. We have already carried out some analyzes and found that we doubled our productivity in the four months since the tool was implemented with the same number of pharmacists. We had a great improvement in the number of prescriptions validated and also in the quality of the validation. We noticed a big increase in the acceptance of interventions: a jump from 16% to 75% of acceptance." Karoline Flach (19:12)

To the best of our knowledge, the evaluation of decision support algorithms in medication review in real scenarios is rarely found in the literature. The results presented in this experiment show a proper evaluation of how the DCC-Score performs in a hospital environment.

This chapter shows how to apply the machine learning algorithm developed in this thesis to a real-world scenario. In the next chapter, we finalize the thesis with our conclusions.

6. CONCLUSION

In past decades, machine learning has been an essential computer science technique to predict and describe real-world instances. The amount of data and processing power available enables the use of ML to solve problems in several areas. In this thesis, taking advantage of the field of research and of the relationship between our group and several hospitals, we tackled the problem of detecting adverse events in hospital environments. This thesis works along these lines: developing ML applications and algorithms to mitigate hospital adverse events.

We developed algorithms that could assist healthcare professionals in improving patient safety. First, we used language models and deep learning to identify sentences with fall events and highlight words that suggest fall events. We annotated thousands of clinical sentences with fall events to train a Token Classifier (TkC) to detect words within the context of falls. The model was able to correctly identify 85% of the sentences with fall events. This result shows improvements from previous work (Luther et al., 2015, Topaz et al., 2019) in number and quality: besides achieving a better F-measure (from 90% to 96%), our model is able to explain the falls it detected. Our work advances in the task by using neural networks and language models and exploring the Portuguese language. Nowadays, nurses need to manually search for events in electronic health records. Our approach could be proactive: it can alert nurses and could be extended to other adverse events. The challenge here is to annotate several sentences of each adverse event to train the deep learning model. The writing style of each hospital's crew could lead to varied results and possibly require a specific training corpus.

Second, we built an unsupervised algorithm that speeds up the medication review process for clinical pharmacy departments. The algorithm we presented can rank outlier prescriptions and help pharmacists in the screening task. Our approach uses graph structure to measure the distance, density, and centrality of the posology of prescribed medications. The DDC-Outlier, using Jaccard distance, correctly classifies 68% of prescribed medications. Our results benefit the field of medication errors by creating an approach that adapts to the history of the medications in each hospital. Even though we only evaluated patient data and the name of medications, the algorithm proved to be suitable for real application (Chapter 5).

Finally, this thesis goes beyond an evaluation of the algorithm using historical data. We integrated our knowledge of machine learning and with the expertise of healthcare experts. This type of endeavor takes time as there is a need to establish trust between all parties (Mateen et al., 2020). Therefore, in collaboration with Hospital Santa Casa, we deployed a clinical pharmacy system to improve pharmacists' work. Pharmacists manually review several prescriptions, compare literature on medications, and evaluate the clinical

condition of patients. The promising results show that the DCC-Outlier was able to correctly rank a variety of patient profiles. We developed the experiments in a 1,200-bed hospital with a diverse patient profile.

During this Ph.D. thesis, we could observe that most researchers focus their efforts on applying machine learning to patient treatment and diagnosis, which are considered critical activities. However, as we discussed herein, we believe that machine learning algorithms may also provide significant benefits for other fields of healthcare, such as risk management, a non-critical activity. It is easier for the machine to learn non-critical activities, to deploy them in real scenarios, where they face less resistance. This allows professionals to dedicate more time to other activities, where human knowledge is essential.

Additionally, to ensure better results in the application of machine learning in healthcare, professionals must be in the loop of the learning process (Wiens et al., 2019). With humans as the gatekeepers of the decisions taken by algorithms, greater safety and improvements in patient outcomes are ensured.

6.1 Contributions

The contributions of this thesis cover both experiments: supervised and unsupervised learning models were able to show improvements compared to state-of-the-art approaches. Besides advances in the field of computer science, this work also contributes to the healthcare industry, enhancing nurses' and pharmacists' ability to identify patients' risk factors.

The main contributions of this work are the following:

- Overview of possibilities for non-critical decision support systems in healthcare in Chapter 2;
- Use of explainable neural network algorithms to detect fall events in Chapter 3;
- Development of a new outlier detection algorithm to detect prescription errors applied and evaluated in a real scenario in Chapter 4;
- Deployment and report of the use of an A.I. system in a real scenario: a hospital with 1,200 beds in Chapter 5;
- Publicly available A.I. system for Clinical Pharmacy on the GitHub page of the project [1].

[1] https://github.com/noharm-ai

- Publicly available algorithms and datasets for reproducibility purposes on the GitHub page of the research group [2];

6.2 Published papers, resources and awards

During this Ph.D. research, we wrote several papers related to healthcare and computer science. The studies that cover the use of artificial intelligence in electronic health records are listed below:

- dos Santos, H. D., Ulbrich, A. H. D., Woloszyn, V., & Vieira, R. (2018). DDC-outlier: preventing medication errors using unsupervised learning. IEEE Journal of Biomedical and Health Informatics;
- dos Santos, H. D. P., Ulbrich, A. H. D., Woloszyn, V., & Vieira, R. (2018, June). An initial investigation of the Charlson comorbidity index regression based on clinical notes. In 2018 IEEE 31st International Symposium on Computer-Based Medical Systems;
- Nunes, R. O., Soares, J. E., dos Santos, H. D., & Vieira, R. (2018, July). MeSHx-Notes: Web-System for Clinical Notes. In International Workshop on Artificial Intelligence in Health;
- Quaini, T. E., dos Santos, H. D., de Abreu, S. C., Consoli, B. S. & Vieira, R. (2019, October). A study on deidentification of clinical developments. In VI Scientific Initiation Workshop on Information Technology and Human Language;
- dos Santos, H. D. P., Silva, A. P., Maciel, M. C. O., Burin, H. M. V., Urbanetto, J. S., & Vieira, R. (2019, October). Fall detection in EHR using word embeddings and deep learning. In 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering;
- Franceschini, P. M., dos Santos, H. D., & Vieira, R. (2020, July). Intrinsic and Extrinsic Evaluation of the Quality of Biomedical Embeddings in Different Languages. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems;
- Santos, J., dos Santos, H. D., & Vieira, R. (2020, July). Fall Detection in Clinical Notes using Language Models and Token Classifier. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems.
- Damasio, J. O., dos Santos, H. D., Ulbrich, A. H. D. P. S. & Vieira, R. (2021, April). Opportunities and Challenges in Fall Risk Management using EHRs and Artificial Intelligence: a Systematic Review. In 2021 23rd International Conference on Enterprise Information Systems (ICEIS).

[2] http://github.com/nlp-pucrs

All experiments above were developed using the dataset from Hospital Nossa Senhora da Conceição. Other papers written during this Ph.D. project can be found in Appendix A. The resources developed in this thesis related to healthcare are listed below:

- A real set of data containing 21 different medications with a total of 150,113 prescriptions[3].
- A biomedical language model trained with 21 million clinical sentences[4].
- An annotated dataset with 1,078 progress notes, with the presence of fall events and their structured description for replication purposes[5].
- A Decision Support System for Clinical Pharmacy[6].

This thesis has been awarded some prizes, as a project that makes an impact in the healthcare industry and makes contributions to computer science. We won the following awards:

- Health Entrepreneurship Award Highlight in 2018, granted by the Everis Foundation and Hospital Sírio-Libanês [7];
- Google Latin America Research Awards 2018[8];
- Google Latin America Research Awards 2019[9] and
- Google Latin America Research Awards 2020[10].

In the next section, we cover the limitations of this work. Despite the advancements provided by our algorithms, some aspects could be improved.

6.3 Limitations

The machine learning models we presented here have the potential to improve several daily tasks in healthcare but still face some limitations. As described in Section

[3]https://github.com/nlp-pucrs/prescription-outliers
[4]https://github.com/nlp-pucrs/cci-regression
[5]https://github.com/nlp-pucrs/fall-detection
[6]https://github.com/noharm-ai
[7]https://www.everis.com/brazil/pt-br/news/newsroom/fundacao-everis-e-sirio-libanes-anunciam-os-finalistas-da-quarta-edicao-do-premio
[8]https://brasil.googleblog.com/2018/10/LARA-2018-latin-america-research-awards.html
[9]https://brasil.googleblog.com/2019/11/os-vencedores-da-setima-edicao-do-lara-programa-de-bolsa-de-pesquisa-para-america-latina.html
[10]http://googlediscovery.com/2020/12/03/vencedores-da-lara-o-programa-de-bolsas-do-google-para-a-america-latina/

3.1.2, the main issues of ML models are the generalization problem, data sample, and data selection.

First, one of the main limitations is the generalization problem of the fall detection model. The Token Classifier (TkC) was trained and evaluated using sentences from the same hospital. There are various ways to write clinical notes, and writing protocols change depending on each hospital. To better evaluate the TkC, it is essential to train and evaluate the model in diverse contexts.

Second, the DDC-Outlier has a severe limitation in data selection. We used only the dose and frequency of a drug to classify the prescribed medications as inliers and outliers in our experiments. Medications are also prescribed in a specific route, time, and, in some cases, with a physician's note. All this information could be used to identify whether the medication is correct or not.

Finally, the outlier medication experiment considered only a single drug in the prescription to evaluate the appropriateness of the drug therapy. The prescription not only has other medications but is also prescribed for a specific patient. Assessing all medications in a prescription is vital to alert pharmacists about drug-to-drug interactions and drug duplicity. Evaluating the drug considering the patients' comorbidities is crucial to assess the appropriateness, effectiveness, safety, and adherence of the drug therapy.

6.4 Future Work

The novelty of this research and the rapid development of technology enables several possible directions for future work. In the field of natural language processing, several other outcomes could benefit from Token Classification using BiLSTM-CRF. Clinical notes feature a wealth of information about patient history (e.g., comorbidities, symptoms, allergy, vital signs) that could be detected and alert healthcare professionals. Besides, the BiLSTM-CRF topology could be used to remove names from text records in order to provide rich content for research, without identifying patients

New approaches in language model generation, such as One-shot, Few-shot, and Zero-shot Learners (Brown et al., 2020), may reduce training data and facilitate adjustments to other adverse events. Another improvement could be achieved by adding other sources of information, such as radiology images and laboratory exams, to the detection of adverse events (Rozenblum et al., 2020).

For clinical pharmacy, much could be done to improve the medication review process performed by pharmacists. A machine learning system could suggest possible new interventions in similar drug-patient cases using interventions made by previous pharmacists in prescriptions. Considering all pieces of data gathered from electronic health records, such

as laboratory exams, patient comorbidities, and other prescribed drugs, the ML models can better understand drug therapies.

Besides, the use of patients' historical medical orders and diagnostic claims as data to detect look-alike/sound-alike medication errors could improve medication error detection (Lambert et al., 2019). Medication errors could also be found in clinical notes using concept extraction and relation classification (Yang et al., 2020), upgrading risk assessment.

REFERENCES

Abadi, M. Agarwal, A. Barham, P. Brevdo, E. Chen, Z. Citro, C. Corrado, G. S. Davis, A. Dean, J. Devin, M. et al. (Mar, 2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv Preprint*, vol. 1603.04467, pp. 19.

Agarwal, D. (Oct, 2006). Detecting Anomalies in Cross-Classified Streams: A Bayesian Approach. *Knowledge and Information Systems*, vol. 11, pp. 29–44.

Ageing, W. H. O. and Unit, L. C. (2008). *WHO Global Report on Falls Prevention in Older Age*. World Health Organization, Geneva, Switzerland.

Agrawal, A. (Jun, 2009). Medication Errors: Prevention using Information Technology Systems. *British Journal of Clinical Pharmacology*, vol. 67, pp. 681–686.

Akbik, A. Bergmann, T. Blythe, D. Rasul, K. Schweter, S. and Vollgraf, R. (2019). FLAIR: An Easy-To-Use Framework for State-of-the-Art - NLP. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 54–59, Minneapolis, Minnesota. ACL Web.

Akbik, A. Blythe, D. and Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649, New Mexico, United States. ACL Web.

Akoglu, L. Tong, H. and Koutra, D. (Jul, 2014). Graph Based Anomaly Detection and Description: A Survey. *Data Mining and Knowledge Discovery*, vol. 29, pp. 626–688.

Alpaydin, E. (2020). *Introduction to Machine Learning*. MIT Press, California, United States.

Alshakrah, M. A. Steinke, D. T. and Lewis, P. J. (Jun, 2019). Patient Prioritization for Pharmaceutical Care in Hospital: A Systematic Review of Assessment Tools. *Research in Social and Administrative Pharmacy*, vol. 15, pp. 767–779.

Ashcroft, D. M. Lewis, P. J. Tully, M. P. Farragher, T. M. Taylor, D. Wass, V. Williams, S. D. and Dornan, T. (Sep, 2015). Prevalence, Nature, Severity and Risk Factors for Prescribing Errors in Hospital Inpatients: Prospective Study in 20 UK Hospitals. *Drug Safety*, vol. 38, pp. 833–843.

Baldwin, D. S. Allgulander, C. Bandelow, B. Ferre, F. and Pallanti, S. (Oct, 2012). An International Survey of Reported Prescribing Practice in the Treatment of Patients with Generalised Anxiety Disorder. *The World Journal of Biological Psychiatry*, vol. 13, pp. 510–516.

Barlow, H. (Sep-Nov, 1989). Unsupervised Learning. *Neural Computation*, vol. 1, pp. 295–311.

Bates, J. Fodeh, S. Brandt, C. and Womack, J. (Apr, 2016). Classification of Radiology Reports for Falls in an HIV Study Cohort. *Journal of the American Medical Informatics Association*, vol. 23, pp. e113–e117.

Beleites, C. Neugebauer, U. Bocklitz, T. Krafft, C. and Popp, J. (Jan, 2013). Sample Size Planning for Classification Models. *Analytica Chimica Acta*, vol. 760, pp. 25–33.

Bojanowski, P. Grave, E. Joulin, A. and Mikolov, T. (Jun, 2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146.

Bond, C. and Raehl, C. L. (Apr, 2007). Clinical Pharmacy Services, Pharmacy Staffing, and Hospital Mortality Rates. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 27, pp. 481–493.

Breunig, M. M. Kriegel, H.-P. Ng, R. T. and Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. In: *Proceedings of the ACM Special Interest Group on Management of Data*, vol. 29, pp. 93–104, Texas, United States. ACM.

Brown, T. B. Mann, B. Ryder, N. Subbiah, M. Kaplan, J. Dhariwal, P. Neelakantan, A. Shyam, P. Sastry, G. Askell, A. et al. (May, 2020). Language Models are Few-Shot Learners. *arXiv Preprint*, vol. 2005.14165, pp. 75.

Buntin, M. B. Burke, M. F. Hoaglin, M. C. and Blumenthal, D. (Mar, 2011). The Benefits of Health Information Technology: A Review of the Recent Literature Shows Predominantly Positive Results. *Health Affairs*, vol. 30, pp. 464–471.

Char, D. S. Shah, N. H. and Magnus, D. (Mar, 2018). Implementing Machine Learning in Health Care—Addressing Ethical Challenges. *The New England Journal of Medicine*, vol. 378, pp. 981.

Chu, J. Dong, W. He, K. Duan, H. and Huang, Z. (Nov, 2018). Using Neural Attention Networks to Detect Adverse Medical Events from Electronic Health Records. *Journal of Biomedical Informatics*, vol. 87, pp. 118–130.

Cohan, A. Fong, A. Ratwani, R. M. and Goharian, N. (2017). Identifying Harm Events in Clinical Care Through Medical Narratives. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 52–59, Massachusetts, United States. ACM.

Da Saúde (BR), M. (2014). Documento de Referência para o Programa Nacional de Segurança do Paciente. Source: http://bvsms.saude.gov.br/bvs/publicacoes/documento_referencia_programa_nacional_seguranca.pdf. January 2021.

D'Avolio, L. W. Nguyen, T. M. Farwell, W. R. Chen, Y. Fitzmeyer, F. Harris, O. M. and Fiore, L. D. (Jul-Aug, 2010). Evaluation of a Generalizable Approach to Clinical Information Retrieval using the Automated Retrieval Console (ARC). *Journal of the American Medical Informatics Association*, vol. 17, pp. 375–382.

De Oliveira, A. P. B. da Silva Oliveira, E. C. and de Oliveira, R. C. (Oct-Dec, 2016). Risk Management Reporting and its Contribution to Patient Safety. *Cogitare Enferm*, vol. 21, pp. 01–08.

De Silva, T. S. MacDonald, D. Paterson, G. Sikdar, K. C. and Cochrane, B. (Mar, 2011). Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) to Represent Computed Tomography Procedures. *Computer Methods and Programs in Biomedicine*, vol. 101, pp. 324–329.

De Souza Urbanetto, J. Creutzberg, M. Franz, F. Ojeda, B. da Silva Gustavo, A. Bittencourt, H. Steinmetz, Q. and Farina, V. (Jun, 2013). Morse Fall Scale: Translation and Transcultural Adaptation for the Portuguese Language. *Revista da Escola de Enfermagem*, vol. 47, pp. 569–575.

Devlin, J. Chang, M.-W. Lee, K. and Toutanova, K. (Oct, 2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv Preprint*, vol. 1810.04805, pp. 16.

Doloresco, F. and Vermeulen, L. C. (Mar, 2009). Global Survey of Hospital Pharmacy Practice. *American Journal of Health-System Pharmacy*, vol. 66, pp. s13–s19.

Emmott, A. F. Das, S. Dietterich, T. Fern, A. and Wong, W.-K. (2013). Systematic Construction of Anomaly Detection Benchmarks from Real DataSwarm Intelligent Tuning of One-Class v-SVM Parameters. In: *Proceedings of the Workshop on Outlier Detection and Description*, pp. 16–21, New York, United States. ACM.

Etges, A. P. B. d. S. de Souza, J. S. Kliemann Neto, F. J. and Felix, E. A. (Jan, 2018). A Proposed Enterprise Risk Management Model for Health Organizations. *Journal of Risk Research*, vol. 22, pp. 513–531.

Fortinsky, R. H. Iannuzzi-Sucich, M. Baker, D. I. Gottschalk, M. King, M. B. Brown, C. J. and Tinetti, M. E. (Sep, 2004). Fall-Risk Assessment and Management in Clinical Practice: Views from Healthcare Providers. *Journal of the American Geriatrics Society*, vol. 52, pp. 1522–1526.

Fung, K. W. Mcdonald, C. and Bray, B. E. (2008). RxTerms - A Drug Interface Terminology Derived from RxNorm. In: *Proceedings of the Annual Symposium American Medical Informatics Association*, pp. 227, Washington, United States. AMIA.

Gallotti, R. M. D. (Jan-Apr, 2004). Eventos Adversos: O que são? *Revista da Associação Médica Brasileira*, vol. 50, pp. 114–114.

Gandhi, T. K. Weingart, S. N. Seger, A. C. Borus, J. Burdick, E. Poon, E. G. Leape, L. L. and Bates, D. W. (Sep, 2005). Outpatient Prescribing Errors and the Impact of Computerized Prescribing. *Journal of General Internal Medicine*, vol. 20, pp. 837–841.

Goldstein, B. Navar, A. Pencina, M. and Ioannidis, J. (May, 2017). Opportunities and Challenges in Developing Risk Prediction Models with Electronic Health Records Data: A Systematic Review. *Journal of the American Medical Informatics Association*, vol. 24, pp. 198–208.

Goodfellow, I. Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press, California, United States.

Graabæk, T. and Kjeldsen, L. J. (Apr, 2013). Medication Reviews by Clinical Pharmacists at Hospitals Lead to Improved Patient Outcomes: A Systematic Review. *Basic & Clinical Pharmacology & Toxicology*, vol. 112, pp. 359–373.

Griese-Mammen, N. Hersberger, K. E. Messerli, M. Leikola, S. Horvat, N. van Mil, J. F. and Kos, M. (Aug, 2018). PCNE Definition of Medication Review: Reaching Agreement. *International Journal of Clinical Pharmacy*, vol. 40, pp. 1199–1208.

Griffin, F. A. and Resar, R. K. (2009). IHI Global Trigger Tool for Measuring Adverse Events. Source: http://www.ihi.org/resources/Pages/Tools/IHIGlobalTriggerToolforMeasuringAEs.aspx. January 2021.

Guglielmi, G. P. Ulbrich, A. H. Flach, K. Dimmer, L. Sinderman, R. and Hoffmann, T. (2020). Inteligência Artificial como Suporte na Validação Técnica das Prescrições. Source: https://www.youtube.com/watch?v=mQg9NOh1FfA. March 2021.

Han, J. Pei, J. and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier, Massachusetts, United States.

Hartmann, N. Fonseca, E. Shulby, C. Treviso, M. Silva, J. and Aluísio, S. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In: *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pp. 122–131, Minas Gerais, Brazil. ACL Web.

Hauskrecht, M. Batal, I. Valko, M. Visweswaran, S. Cooper, G. F. and Clermont, G. (Feb, 2013). Outlier Detection for Patient Monitoring and Alerting. *Journal of Biomedical Informatics*, vol. 46, pp. 47–55.

Hitcho, E. B. Krauss, M. J. Birge, S. Claiborne Dunagan, W. Fischer, I. Johnson, S. Nast, P. A. Costantinou, E. and Fraser, V. J. (Jul, 2004). Characteristics and Circumstances of

Falls in a Hospital Setting: A Prospective Analysis. *Journal of General Internal Medicine*, vol. 19, pp. 732–739.

Hochreiter, S. and Schmidhuber, J. (Aug, 1997). Long Short-Term Memory. *Neural Computation*, vol. 9, pp. 1735–1780.

Huynh, T. He, Y. Willis, A. and Rueger, S. (2016). Adverse Drug Reaction Classification with Deep Neural Networks. In: *Proceedings of 26th International Conference on Computational Linguistics: Technical Papers*, pp. 877–887, Osaka, Japan. ACL Web.

Jagannatha, A. N. and Yu, H. (2016). Bidirectional RNN for Medical Event Detection in Electronic Health Records. In: *Proceedings of the Conference Association for Computational Linguistics North American*, vol. 2016, pp. 473, California, United States. NIH Public Access.

Jensen, P. B. Jensen, L. J. and Brunak, S. (May, 2012). Mining Electronic Health Records: Towards Better Research Applications and Clinical Care. *Nature Reviews Genetics*, vol. 13, pp. 395.

Jiang, Y. Hu, C. Xiao, T. Zhang, C. and Zhu, J. (2019). Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 3585–3590, Hong Kong, China. ACL Web.

Jiang, Z. Li, L. Huang, D. and Jin, L. (Nov, 2015). Training Word Embeddings for Deep Learning in Biomedical Text Mining Tasks. In: *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pp. 625–628, Washington, United States. IEEE.

Jordan, M. I. and Mitchell, T. M. (Jul, 2015). Machine Learning: Trends, Perspectives, and Prospects. *Science*, vol. 349, pp. 255–260.

Jurafsky, D. and Martin, J. H. (2014). *Speech and Language Processing*, vol. 3. Pearson, London, United Kingdom.

Kopp, B. J. Erstad, B. L. Allen, M. E. Theodorou, A. A. and Priestley, G. (Feb, 2006). Medication Errors and Adverse Drug Events in an Intensive Care Unit: Direct Observation Approach for Detection. *Critical Care Medicine*, vol. 34, pp. 415–425.

Kruse, C. S. and Beane, A. (Feb, 2018). Health Information Technology Continues to Show Positive Effect on Medical Outcomes: Systematic Review. *Journal of Medical Internet Research*, vol. 20, pp. e41.

Lambert, B. L. Galanter, W. Liu, K. L. Falck, S. Schiff, G. Rash-Foanio, C. Schmidt, K. Shrestha, N. Vaida, A. J. and Gaunt, M. J. (Nov, 2019). Automated Detection of Wrong-Drug Prescribing Errors. *BMJ Quality & Safety*, vol. 28, pp. 908–915.

Langville, A. N. and Meyer, C. D. (Jan, 2005). A Survey of Eigenvector Methods for Web Information Retrieval. *Society for Industrial and Applied Mathematics Review*, vol. 47, pp. 135–161.

Li, Y. and Yang, T. (2017). Word Embedding for Understanding Natural Language: A Survey. In: *Guide to Big Data Applications*, vol. 26, pp. 83–104. Springer, 1 ed..

Liu, F. T. Ting, K. M. and Zhou, Z.-H. (2008). Isolation Forest. In: *Proceedings of the Eighth IEEE International Conference on Data Mining*, pp. 413–422, Pisa, Italy. IEEE.

LLC, W. K. (2017). UpToDate: Evidence-Based Clinical Decision Support. Source: https://www.uptodate.com. January 2021.

Luther, S. McCart, J. Berndt, D. Hahm, B. Finch, D. Jarman, J. Foulis, P. Lapcevic, W. Campbell, R. Shorr, R. Valencia, K. and Powell-Cope, G. (Apr, 2015). Improving Identification of Fall-Related Injuries in Ambulatory Care using Statistical Text Mining. *American Journal of Public Health*, vol. 105, pp. 1168–1173.

Mateen, B. A. Liley, J. Denniston, A. K. Holmes, C. C. and Vollmer, S. J. (2020). Improving the quality of machine learning in health applications and clinical research. *Nature Machine Intelligence*, vol. 2, pp. 554–556.

McCart, J. Berndt, D. Jarman, J. Finch, D. and Luther, S. (Sep, 2013). Finding Falls in Ambulatory Care Clinical Documents Using Statistical Text Mining. *Journal of the American Medical Informatics Association*, vol. 20, pp. 906–914.

Mendes, W. Pavão, A. L. B. Martins, M. de Oliveira Moura, M. d. L. and Travassos, C. (Sep, 2013). Características de Eventos Adversos Evitáveis em Hospitais do Rio De Janeiro. *Revista da Associação Médica Brasileira*, vol. 59, pp. 421–428.

Mikolov, T. Sutskever, I. Chen, K. Corrado, G. S. and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 3111–3119, New York, United States. ACM.

Mills, A. J. Durepos, G. and Wiebe, E. (2009). *Encyclopedia of Case Study Research*. Sage Publications, California, United States.

Moonesinghe, H. and Tan, P.-N. (Jan, 2008). OutRank: A Graph-Based Outlier Detection Framework using Random Walk. *International Journal on Artificial Intelligence Tools*, vol. 17, pp. 19–36.

Morse, J. Morse, R. and Tylko, S. (Dec-Mar, 1989). Development af a Scale to Identify the Fall-Prone Patient. *Canadian Journal on Aging / La Revue Canadienne du Vieillissement*, vol. 8, pp. 366–377.

Mubashir, M. Shao, L. and Seed, L. (Jan, 2013). A Survey on Fall Detection: Principles and Approaches. *Neurocomputing*, vol. 100, pp. 144 – 152.

Muller, E. Sánchez, P. I. Mulle, Y. and Bohm, K. (2013). Ranking Outlier Nodes in Subspaces of Attributed Graphs. In: *Proceedings of the 29th International Conference on Data Engineering Workshops*, pp. 216–222, Queensland, Australia. IEEE.

Nangle, C. McTaggart, S. MacLeod, M. Caldwell, J. and Bennie, M. (Apr, 2017). Application of Natural Language Processing Methods to Extract Coded Data from Administrative Data Held in the Scottish Prescribing Information System. *International Journal for Population Data Science*, vol. 1:243, pp. 1.

Oh, K.-S. and Jung, K. (Jun, 2004). GPU Implementation of Neural Networks. *Pattern Recognition*, vol. 37, pp. 1311–1314.

Okanda, P. and Kanyaru, J. (2014). Smartprescription: A Principled Approach Towards Eliminating Prescription Errors in Healthcare. In: *Proceedings of the Conference IST-Africa*, pp. 1–8, Le Meridien Ile Maurice, Mauritius. IEEE.

Oliver, D. (Jul, 2007). Preventing Falls and Fall Injuries in Hospital: A Major Risk Management Challenge. *Clinical Risk*, vol. 13, pp. 173–178.

Organization, W. H. (2009). The Conceptual Framework for the International Classification for Patient Safety. Technical Report, World Health Organization.

Page, L. Brin, S. Motwani, R. and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab.

Park, S. Choi, D. Kim, M. Cha, W. Kim, C. and Moon, I.-C. (Nov, 2017). Identifying Prescription Patterns with a Topic Model of Diseases and Medications. *Journal of Biomedical Informatics*, vol. 75, pp. 35–47.

Paszke, A. Gross, S. Massa, F. Lerer, A. Bradbury, J. Chanan, G. Killeen, T. Lin, Z. Gimelshein, N. Antiga, L. et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 8024–8035, Vancouver, Canada. NeurIPS.

Pedregosa, F. Varoquaux, G. Gramfort, A. Michel, V. Thirion, B. Grisel, O. Blondel, M. Prettenhofer, P. Weiss, R. Dubourg, V. Vanderplas, J. Passos, A. Cournapeau, D. Brucher, M. Perrot, M. and Édouard Duchesnay (Oct, 2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830.

Pennington, J. Socher, R. and Manning, C. D. (2014). Glove: Global Vectors for Word Representation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, Doha, Qatar. ACL Web.

Peters, M. E. Neumann, M. Iyyer, M. Gardner, M. Clark, C. Lee, K. and Zettlemoyer, L. (2018b). Deep Contextualized Word Representations. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237, New Orleans, Louisiana. ACL Web.

Peters, M. E. Neumann, M. Iyyer, M. Gardner, M. Clark, C. Lee, K. and Zettlemoyer, L. (Feb, 2018a). Deep Contextualized Word Representations. *arXiv Preprint*, vol. 1802.05365, pp. 15.

Radford, A. Wu, J. Child, R. Luan, D. Amodei, D. and Sutskever, I. (Feb, 2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, vol. 1, pp. 9.

Rash-Foanio, C. Galanter, W. Bryson, M. Falck, S. Liu, K. L. Schiff, G. D. Vaida, A. and Lambert, B. L. (Apr, 2017). Automated Detection of Look-Alike/Sound-Alike Medication Errors. *American Journal of Health-System Pharmacy*, vol. 74, pp. 521–527.

Ravì, D. Wong, C. Deligianni, F. Berthelot, M. Andreu-Perez, J. Lo, B. and Yang, G.-Z. (Dec, 2016). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, vol. 21, pp. 4–21.

Reps, J. M. Garibaldi, J. M. Aickelin, U. Soria, D. Gibson, J. E. and Hubbard, R. B. (Mar, 2014). A Novel Semisupervised Algorithm for Rare Prescription Side Effect Discovery. *IEEE Journal of Biomedical and Health Informatics*, vol. 18, pp. 537–547.

Resar, R. Rozich, J. Simmonds, T. and Haraden, C. (Oct, 2006). A Trigger Tool to Identify Adverse Events in the Intensive Care Unit. *Joint Commission Journal on Quality and Patient Safety*, vol. 32, pp. 585–590.

Rochefort, C. Buckeridge, D. and Abrahamowicz, M. (Jun, 2015). Improving Patient Safety by Optimizing the use of Nursing Human Resources. *Implementation Science*, vol. 10, pp. 11.

Rousseeuw, P. J. and Driessen, K. V. (Aug, 1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, vol. 41, pp. 212–223.

Rozenblum, R. Rodriguez-Monguio, R. Volk, L. A. Forsythe, K. J. Myers, S. McGurrin, M. Williams, D. H. Bates, D. W. Schiff, G. and Seoane-Vazquez, E. (Jan, 2020). Using a Machine Learning System to Identify and Prevent Medication Prescribing Errors: A Clinical and Cost Analysis Evaluation. *The Joint Commission Journal on Quality and Patient Safety*, vol. 46, pp. 3–10.

Sang, T. K. and Erik, F. (2002). Introduction to the CONLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: *Proceedings of the 6th Conference on Natural Language Learning*, pp. 155–158, Pennsylvania, United States. ACM.

Saseen, J. J. Ripley, T. L. Bondi, D. Burke, J. M. Cohen, L. J. McBane, S. McConnell, K. J. Sackey, B. Sanoski, C. Simonyan, A. et al. (May, 2017). ACCP Clinical Pharmacist Competencies. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 37, pp. 630–636.

Savova, G. K. Masanz, J. J. Ogren, P. V. Zheng, J. Sohn, S. Kipper-Schuler, K. C. and Chute, C. G. (Sep-Oct, 2010). Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications. *Journal of the American Medical Informatics Association*, vol. 17, pp. 507–513.

Schiff, G. D. Volk, L. A. Volodarskaya, M. Williams, D. H. Walsh, L. Myers, S. G. Bates, D. W. and Rozenblum, R. (Mar, 2017). Screening for Medication Errors using an Outlier Detection System. *Journal of the American Medical Informatics Association*, vol. 24, pp. 281–287.

Schölkopf, B. Platt, J. C. Shawe-Taylor, J. Smola, A. J. and Williamson, R. C. (Jul, 2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, vol. 13, pp. 1443–1471.

Sheikh, A. Dhingra-Kumar, N. Kelley, E. Kieny, M. P. and Donaldson, L. J. (Aug, 2017). The Third Global Patient Safety Challenge: Tackling Medication-Related Harm. *Bulletin of the World Health Organization*, vol. 95, pp. 546.

Shiner, B. Neily, J. Mills, P. and Watts, B. (Sep, 2016). Identification of Inpatient Falls using Automated Review of Text-Based Medical Records. *Journal of Patient Safety*, vol. 3, pp. e174–e178.

Solutions, M. (2017). Medication, Disease and Toxicology Management. Source: https://www.micromedexsolutions.com. January 2021.

Straková, J. Straka, M. and Hajic, J. (2019). Neural Architectures for Nested NER Through Linearization. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5326–5331, Florence, Italy. ACL Web.

Topaz, M. Murga, L. Gaddis, K. McDonald, M. Bar-Bachar, O. Goldberg, Y. and Bowles, K. (Jan, 2019). Mining Fall-Related Information in Clinical Notes: Comparison of Rule-Based and Novel Word Embedding-Based Machine Learning Approaches. *Journal of Biomedical Informatics*, vol. 90, pp. 8.

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, Inc., New York, United States.

Touchette, D. R. Doloresco, F. Suda, K. J. Perez, A. Turner, S. Jalundhwala, Y. Tangonan, M. C. and Hoffman, J. M. (Aug, 2014). Economic Evaluations of Clinical Pharmacy

Services: 2006–2010. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 34, pp. 771–793.

Toyabe, S.-I. (Dec, 2012). Detecting Inpatient Falls by Using Natural Language Processing of Electronic Medical Records. *BMC Health Services Research*, vol. 12, pp. 8.

Tremblay, M. Berndt, D. Luther, S. Foulis, P. and French, D. (Nov, 2009). Identifying Fall-Related Injuries: Text Mining the Electronic Medical Record. *Information Technology and Management*, vol. 10, pp. 253–265.

Walsh, M. Frances Horgan, N. Walsh, C. and Galvin, R. (May, 2016). Systematic Review of Risk Prediction Models for Falls after Stroke. *Journal of Epidemiology and Community Health*, vol. 70, pp. 513–519.

Wang, F. and Preininger, A. (Aug, 2019). AI in Health: State of the Art, Challenges, and Future Directions. *Yearbook of Medical Informatics*, vol. 28, pp. 16.

Wiens, J. Saria, S. Sendak, M. Ghassemi, M. Liu, V. X. Doshi-Velez, F. Jung, K. Heller, K. Kale, D. Saeed, M. et al. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, vol. 25, pp. 1337–1340.

Woloszyn, V. dos Santos, H. D. P. Wives, L. K. and Becker, K. (2017a). MRR: An Unsupervised Algorithm to Rank Reviews by Relevance. In: *Proceedings of the International Conference on Web Intelligence*, pp. 877–883, Leipzig, Germany. ACM.

Woloszyn, V. Machado, G. M. de Oliveira, J. P. M. Wives, L. and Saggion, H. (2017b). Beatnik: An Algorithm to Automatic Generation of Educational Description of Movies. In: *Proceedings of the Brazilian Symposium on Computers in Education*, pp. 10, Pernambuco, Brazil. BR-IE.

Wunnava, S. Qin, X. Kakar, T. Rundensteiner, E. A. and Kong, X. (2018). Bidirectional LSTM-CRF for Adverse Drug Event Tagging in Electronic Health Records. In: *Proceedings of the International Workshop on Medication and Adverse Drug Event Detection*, pp. 48–56, Massachusetts, United States. PMLR.

Xie, L. (2006). Swarm Intelligent Tuning of One-Class v-SVM Parameters. In: *Proceedings of the International Conference on Rough Sets and Knowledge Technology*, pp. 552–559, Chongquing, China. Springer.

Xing, W. and Ghorbani, A. (2004). Weighted PageRank Algorithm. In: *Proceedings Second Annual Conference on Communication Networks and Services Research*, pp. 305–314, Nova Brunswick, Canada. IEEE.

Yang, X. Bian, J. Fang, R. Bjarnadottir, R. I. Hogan, W. R. and Wu, Y. (Jan, 2020). Identifying Relations of Medications with Adverse Drug Events using Recurrent Convolutional Neural

Networks and Gradient Boosting. *Journal of the American Medical Informatics Association*, vol. 27, pp. 65–72.

Yang, Z. Dai, Z. Yang, Y. Carbonell, J. Salakhutdinov, R. R. and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining For Language Understanding. In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 5754–5764, Vancouver, Canada. NeurIPS.

Yimam, S. M. Gurevych, I. de Castilho, R. E. and Biemann, C. (2013). Webanno: A Flexible, Web-Based and Visually Supported System for Distributed Annotations. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 1–6, Sofia, Bulgaria. ACL Web.

APPENDIX A – OTHER RELATED PUBLICATIONS

The following section enumerates other related works published by the author of this thesis in collaboration with several students and advisors, during the course of this Ph.D. These studies consolidate knowledge and machine learning and natural language processing techniques.

During this Ph.D. candidature, we established other partnerships with research groups and published other papers. Some studies relate to NLP, Core-NLP, ML, electronic health records, and other areas of healthcare, as listed below in chronological order:

A.1 Portuguese Personal Story Detection and Analysis in Blogs

Diary-like content expressing authors' personal experiences and sentiments relating to various topics is generated every day and made available on the Internet. This rich content can be used for psychological analysis and knowledge discovery regarding human-related issues in several ways. This paper presented a Brazilian Portuguese corpus, using blog posts, to analyze and detect personal stories. We presented an analysis of psycholinguistic categories across personal-story and non-story posts, discussing their similarities and differences. We also studied the use of these psycholinguistic categories as classifying features. Then we described the evaluation of several machine learning approaches and the process of applying them to identify personal stories based on our dataset. Finally, we investigated the central topic-related polarity of individual narrative posts.

Published in the Proceedings of the International Conference on Web Intelligence, August 2017, in collaboration with student Vinicius Woloszyn, advised by Prof. Renata Vieira.

A.2 PLN-PUCRS at EmoInt-2017: Psycholinguistic Features for Emotion

Linguistic Inquiry and Word Count (LIWC) is a rich dictionary that maps words into several psychological categories, such as Affective, Social, Cognitive, Perceptual, and Biological processes. In this work, we used LIWC psycholinguistic categories to train regression models and predict emotion intensity in tweets for the EmoInt-2017 task. Results showed that LIWC features may boost emotion intensity prediction based on a standard dimension set.

Published in the Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, September 2017, advised by Prof. Renata Vieira.

A.3 Wheel of Life, an Initial Investigation

User-generated content is a rich source of information regarding human behavior in social media. Sentiment analysis is a powerful tool to understand human psychological meanings in the text. Visualizing these sentiments and knowledge about users is crucial to figure out the trends in data and use it to make decisions. This work presented an initial investigation about a visualization chart considering topic-related polarities in Brazilian bloggers' personal stories. Visualizing these sentiments allows specialists to rapidly understand user-affected areas of life.

Published in the Proceedings of the 11th Brazilian Symposium on Information and Human Language Technology, October 2017, in collaboration with students Greice P. D. Molin and Jackson Pinheiro, advised by Prof. Renata Vieira.

A.4 Blogset-BR: A Brazilian Portuguese Blog Corpus

The rich user-generated content found on blogs has always attracted the interest of the scientific community for many different reasons, such as opinion and sentiment mining, information extraction, and topic discovery. Nonetheless, an extensive corpus is essential to perform most of the natural language processing involved in these tasks. This paper presented BlogSet-BR, an extensive Brazilian Portuguese corpus containing 2.1 billion words extracted from 7.4 million posts from 808,000 different Brazilian blogs. Additionally, a survey was conducted with authors to draw a profile of Brazilian bloggers.

Published in the Proceedings of the Eleventh International Conference on Language Resources and Evaluation, May 2018, in collaboration with student Vinicius Woloszyn, advised by Prof. Renata Vieira.

A.5 Cross-Framework Evaluation for Portuguese POS Taggers

This work compared POS and parsing systems for the Portuguese language. We analyzed available features and tagsets and compared the results of POS tagging and syntactic structure identification using both intrinsic and extrinsic evaluation methods. To do so,

we used well-known metrics for parser evaluation, such as bracket cross, besides leaf ancestor for intrinsic evaluation. We also applied these parsers to the extrinsic evaluation task of noun phrase identification. The comparison proposed in this paper considers the different linguistic theories and frameworks each parser subscribes to, but it is not dependent on any particular one.

Published in the 19th International Conference on Computational Linguistics and Intelligent Text Processing, May 2018, in collaboration with students Sandra Collovini, Thiago Lima, Evandro Fonseca, Bolivar Pereira, and Marlo Souza, advised by Prof. Silvia Moraes and Prof. Renata Vieira

A.6 Annotating Relations between Named Entities Crowdsourcing

This paper described how the CrowdFlower platform was used to build an annotated corpus for relation extraction. The obtained data provides information on the relations between named entities in texts in Portuguese.

Published in the International Conference on Applications of Natural Language to Information Systems, May 2018, in collaboration with students Sandra Collovini and Bolivar Pereira, advised by Prof. Renata Vieira.

A.7 An Initial Investigation of the Charlson Index Regression

The Charlson comorbidity index (CCI) is widely used to predict mortality for patients with many comorbid conditions. The index is also used as an indicator of the patients' complexity inside a hospital. This paper evaluated a variety of feature extraction and regression methods to predict the CCI from clinical notes. We used a tertiary hospital dataset with 48,000 hospitalizations featuring the CCI annotated by physicians. In our experiments, Dense Neural Networks with Word Embeddings proved to be the best regression method, with a mean absolute error of 0.51.

Published in the IEEE 31st International Symposium on Computer-Based Medical Systems, July 2018, in collaboration with students Ana Helena D. P. S. Ulbrich and Vinícius Woloszyn, advised by Prof. Renata Vieira.

A.8 MeSHx-Notes: Web System for Clinical Notes Information

We introduced MeSHx-Notes, MeSH eXtended for clinical notes, a multi-language web system based on the Django framework to present selected terms in clinical notes. MeSHx-Notes extended Medical Subject Headings (MeSH) terms with Word Embeddings with similar words. Since MeSH is available in 15 languages, MeSHx-Notes is easily extendable by replacing the MeSH thesaurus with the target language (plus the generation of the corresponding WE for the new language). Our version deals with Portuguese and English.

Published in the First International Workshop of Artificial Intelligence in Health, July 2018, in collaboration with students Rafael O. Nunes and João E. Soares, advised by Prof. Renata Vieira.

A.9 A Study on Deidentification of Clinical Developments

Medical records of patients are essential in the field of medical research. However, to obtain the identity of a patient, the Health Insurance Portability and Accountability Act (HIPAA) is required. It must be removed before the study. The manual de-identification of large amounts of medical record data is expensive, time-consuming, and error-prone, requiring large-scale automated de-identification methods. This paper presented an analysis of the problem in Brazilian Portuguese for a task of disidentification of electronic medical records. We compared the main types of business rule identification with an approach based on a list of names specially built for the task. The list of names was developed from the database, by using the embedded words to specialize the names through the semantic similarity between words.

Published in the VI Workshop of Scientific Initiation in Information Technology and Human Language, October 2019, in collaboration with students Thaila Elisa Quaini, Sandra C. de Abreu, and Bernardo S. Consoli, advised by Prof. Renata Vieira.

A.10 Fall Detection in EHR using Word Embeddings and Deep Learning

Electronic health records (EHR) are an essential source of information to detect adverse events in patients. In-hospital fall incidents represent the largest category of adverse event reports. The detection of such incidents leads to a better understanding of the event and improves patient healthcare quality. This work evaluated several language models with state-of-the-art recurrent neural networks (RNN) to detect fall incidents in progress notes.

Our experiments showed that the deep-learning approach outperforms previous works in the task of detecting fall events. The vector representation of words in the biomedical domain was able to detect falls with an F-measure of 90%. Additionally, we made available an annotated dataset with 1,078 de-identified progress notes for replication purposes.

Published in the IEEE 19th International Conference on Bioinformatics and Bioengineering, October 2019, in collaboration with students Amanda P. Silva, Maria Carolina O. Maciel, and Haline Maria V. Burin, advised by Prof. Janete S. Urbanetto and Prof. Renata Vieira.

A.11 Cross-Media Sentiment Analysis in Brazilian Blogs

The use of social media is becoming highly present in our lives. Through images, texts, and videos, human beings try to communicate in social networks and express their opinions in the face of everyday events. Due to the increased volume of data transmitted over the Internet, doing a human analysis of this content becomes difficult. For this reason, it is necessary to automate the task of classifying feelings. Although the area of classification of feelings in images and texts is well developed and applied in the social network context, the classification of feelings from images together with texts is still under development. A challenge is to build algorithms and methods that can infer feelings just like humans perceive them. Firstly, we presented a cross-media corpus of Brazilian blogs, the dataset we built based on BlogSet-BR, whose goal was to have a data ground truth (based on subjects' opinions) concerning feelings perceived in texts and images when analyzed separately as well as when presented together. Therefore, we tested some available technologies to detect sentiment polarities in texts and images and compared them with the ground truth. Besides, we conducted research specifically on contradictory posts, i.e., when, in the same blog post, the image is positive and the text is negative. The results indicated that subjectivity affects emotional judgments because there are variances between cultures.

Published in the 14th International Symposium on Visual Computing, October 2029, in collaboration with student Greice P. Dal Molin, advised by Prof. Isabel H. Manssour, Prof. Renata Vieira, and Prof. Soraia R. Musse.

A.12 Multivariable Prediction Model to Predict Subjective Refraction

The study aimed to test machine learning models to predict subjective ocular refraction from patients' demographics and ophthalmological data and compare the performance of the model with an automatic refractometer. The dataset consisted of ophthalmic examination data of 17,039 eyes from TeleOftalmo, a teleophthalmology project in the Brazilian

public health system. We collected the following variables to be tested as attributes in the predictive model: age, gender, race, symptoms, uncorrected visual acuity, best-corrected visual acuity, pinhole visual acuity, intraocular pressure (Visuplan, Zeiss, Germany), dioptric power of current spectacles, keratometry measurements, and automatic refraction (Visuref, Zeiss, Germany). Same-day subjective refraction performed by an ophthalmologist was defined as the target attribute. Subjective refraction was converted into power vectors (M, J0, and J45). We used the Orange Data Mining Toolbox in Python to run the tests. The performances of Random Forest, Linear Regression, and Neural Network (Multi-Layer Perceptron) algorithms in predicting subjective refraction were assessed in terms of mean absolute error (MAE) and root mean square error (RMSE). We determined the automatic refraction error for comparison purposes, defined as the difference of the component M between automatic refraction and subjective refraction without a predictive model.

Published in the Annual Meeting of the Association for Research in Vision and Ophthalmology, June 2020, in collaboration with Aline Lutz de Araujo, Daniel Sganzerla, Roberto Nunes Umpierre, and Paulo Schor.

A.13 Machine Learning Early Warning System Evaluation

Early recognition of clinical deterioration is one of the main steps to reduce inpatient morbidity and mortality. The challenging task of identification of clinical deterioration in hospitals lies in the intense daily routines of healthcare practitioners, in the unconnected patient data stored in electronic health records (EHRs), and in the use of low accuracy scores. Since hospital wards are given less attention than the Intensive Care Unit (ICU), we hypothesized that when a platform is connected to a stream of EHRs, there would be a drastic improvement in the awareness of dangerous situations, which could thus assist the healthcare team. With the application of machine learning, the system can consider all patients' histories, and an intelligent early warning system is enabled through the use of high-performing predictive models. In this study, we used 121,089 medical encounters from 6 (six) different hospitals and 7,540,389 data points. We compared popular ward protocols with six different scalable machine learning methods (three are classic machine learning models, logistic and probabilistic-based models, and three are gradient boosted models). The results showed an advantage in AUC (Area Under the Receiver Operating Characteristic Curve) of 25 percentage points in the result of the best machine learning model compared to the current state-of-the-art protocols. The generalization of the algorithm demonstrates this result with leave-one-group-out (AUC of 0.949) and the robustness through cross-validation (AUC of 0.961). We also performed experiments to compare several window sizes to justify the use of five patient timestamps. A sample dataset, experiments, and code are available for replicability purposes.

Published in the IEEE 33rd International Symposium on Computer-Based Medical Systems, July 2020, in collaboration with students Jhonatan Kobylarz Ribeiro, Felipe Barletta, and Mateus Cichelero da Silva, advised by Prof. Renata Vieira, Prof. Hugo M. P. Morales, and Prof. Cristian da Costa Rocha.

A.14 Intrinsic and Extrinsic Evaluation of Biomedical Embeddings

Lately, language models have been applied to several tasks in biomedical natural language processing. Some language models are available online, each built with different corpora. This paper evaluated different public word embedding models trained with both general and biomedical corpora for English and Portuguese. We presented intrinsic evaluations based on semantic analogies that use word pairs extracted from the MeSH biomedical thesaurus and benchmarks available for general-domain evaluation. For extrinsic evaluations, we relied on a classification task over electronic health records. Our experiments showed that biomedical embeddings can better capture semantics for biomedical analogies in both languages. Conversely, based on classification tasks using the language models, larger general textual corpora were equally or more effective for extrinsic evaluations.

Published in the IEEE 33rd International Symposium on Computer-Based Medical Systems, July 2020, in collaboration with student Paula M. Franceschini, advised by Prof. Renata Vieira.

A.15 Implementations of Fuzzy Logic for Knee Rehabilitation

Since its beginning, artificial intelligence (AI) followed a strategy to mimic human cognition. In the physical rehabilitation area, studies include AI to process, estimate, and classify physical activity levels. In order to improve the professional-patient relationship, this study aimed to develop and compare the implementations of Fuzzy Logic (FL) of Sugeno (SFL) and Mamdani (MFL) types to assist the physical therapist in deciding, with more data, whether patients can safely return to their activities. The implemented systems consist of a sequence of fuzzy rules (if – then) and four inputs taking into account range of motion, extension and flexion; pain; and muscle strength to generate an output on the capability of the knee. The qualitative requirements of the systems are taken into account, along with the processing time, the precision, and the reliability of the responses. When comparing MFL and SFL, the Sugeno method obtained more reliable responses regarding the level of pertinence; however, both systems showed agreement between the values reported in six hypothetical clinical cases and the resulting concepts of capability.

Published in the XXVII Congress Brazilian Biomedical Engineering, October 2020, in collaboration with student Thiago Susin, advised by Prof. Rafael Baptista and Prof. Fabian Vargas.

A.16 Analysis of the Agreement between Electronic Medical Records and Notifications in the Record of Falls: a Cohort Study

We analyzed the agreement between the daily clinical notes in the electronic medical records of the patients and the notifications in the Computerized Notification System in the record of falls. This retrospective cohort study was carried out in a public hospital in the city of Porto Alegre, in the state of Rio Grande do Sul, Brazil. The study comprised 367 patients, 441 voluntary notifications, and 441 evolutions. Data collection took place in the online annotation tool WebAnno, from September to December 2018. An instrument for the collection was developed. Data analysis was performed using descriptive statistics. Among the patients, 316 had one fall and 51 had two or more falls. The study included 441 reports of falls. Of these, 43.9% were not recorded in the electronic medical record on the day of its occurrence. Regarding the assessment of the risk of falls, only 3 (three) (0.7%) evolutions contained the record. When analyzing the records in the notifications and electronic medical records, more complete reports were identified in the notifications. The local variables of the fall stand out, registered in all notifications and in 13.8% of the evolutions; the degree of damage was recorded in all notifications and in only 1.6% of evolutions. We identified a gap in the records of falls in the medical records. The results point to an aspect of extreme relevance in the issue of communication via the patient's medical records, which may directly impact the planning and implementation of effective care.

Published in the journal Research, Society and Development, December 2020, in collaboration with students Amanda P. Silva, Maria Carolina O. Maciel, and Haline Maria V. Burin, advised by Prof. Janete S. Urbanetto and Prof. Renata Vieira.

A.17 Fall Risk Prediction and Fall Detection: a Systematic Review

We searched for articles that reported using EHRs and artificial intelligence techniques to identify in-hospital falls in several digital libraries. Three authors of this work selected articles. We compiled information on study design, use of EHR data types, and methods. We identified 19 articles — 11 about fall risk prediction and 8 covering fall detection. Studies varied according to sample size (from 750 to 57,678) and used diverse sources of information (text, administrative data, patient history, and medication data). All studies used validation methods to evaluate the performance of the model, and most showed their

performance (n = 17 of 19). However, studies limited their use of EHR data, picking some items of data as predictors instead of all available data. All studies that developed fall event detection used textual information; however, studies focusing on fall risk prediction generally used structured data (n = 9 of 11). A small portion made a multicenter study evaluation (n = 4), while the rest of the studies were validated through self-assessment. EHR data shows opportunities and challenges for fall risk prediction and in-hospital fall detection. There is room for improvement in developing such studies.

Published in the International Conference on Enterprise Information Systems (ICEIS), April 2021, in collaboration with student Juliana D. Oliveira and Ana Helena D. P. S. Ulbrich, advised by Prof. Renata Vieira.

A.18 Nephrotoxicity and Formula for Vancomycin in a Tertiary Hospital

A retrospective cohort study was conducted in a public tertiary Brazilian hospital. We analyzed 930 courses of vancomycin therapy in 2016. We developed formulas using the relationship between the daily dose of vancomycin (mg) and the product of the patient's calculated endogenous creatinine clearance (ECC) (ml/min/1.73 m^2) and weight (kg): F-MDRD, F-CKD-EPI, and F-CG, when MDRD, CKD-EPI, and Cockroft-Gault were used as the ECC calculation equations, respectively. The accuracy of the formulas and the vancomycin serum level (SL) were evaluated and compared to predict the following outcomes: AKI and AKI requiring hemodialysis (AKI-HD).

Not published yet, in collaboration with Ana Helena D. P. S. Ulbrich, Clarissa B. Pinto, Cláudia R. Ames, Fernando P. Junior, Jaqueline Pandolfo, Marina B. Oliveira, and Graziella G. Baiocco, advised by Prof. Paulo R. M. Rosa.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br