UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÀTICA

Linking patient data to scientific knowledge to support contextualized mining

Ricardo Miguel Serafim de Carvalho

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Prof.a Doutora Cátia Luísa Santana Calisto Pesquita
Prof.a Doutora Daniela Patrícia dos Santos Oliveira

2022

Acknowledgments

To all that are or were in my life, you are the reason I got this far. Your words, your advice, your example or simply your presence meant everything to me, and weather you know it or not, you molded my character and made me the knowledge angry and ambitious person i am today, and for that I thank you all. With this being said there are a couple of people, to whom I owe special thanks.

First, to my supervisors, Prof. Cátia Pesquita and Daniela Oliveira, thank you for the opportunity, guidance and advise. You were the true back bone of this work, without your help, dedication and availability this would never be possible.

To my parents Carla and Paulo, because if my supervisors were the back bone in this work, you were, and still are, the back bone of my life, teaching, caring and pushing me to new heights that I never tough were reachable. Since the very beginning, you sacrificed your self's for me, so that I could have the best education possible to reach any dream I had. For the sacrifice, care, love and trust I can never thank you enough but take this work as a prof that your efforts were not in vain.

To my life long friends, Bia, Bernardo, Gil, Vicente, Duarte and Silvestre, thank you. You all were there in the best moments of my journey but you all were also there in the worse, helping and motivating me get back up and keep working. I extend this thanks to my dearest friends Susana, Vera, João and Andre, because during our degrees we were always together, sharing sleepless nights and disappointments but above that we shared the good moments and supported each other.

Finally, I wold like to thank my girlfriend Beatriz, because during this journey you were there though it all. During this we were true life partners, sacrificing our dates, having late night conversations about frustrations and together reviewing every content made in this work. When things went well we celebrated together, and when thing when bad we suffered together and got back up together, you never let me be alone and for that I can never thank you enough.

Dedicatória.

Resumo

As admissões em Unidade de Cuidados Intensivos Unidades de Cuidados Intensivos (UCI) geralmente estão associadas a condições graves, doenças ou complicações devido a uma doença. As readmissões de UCI acontecem porque doenças graves detectadas anteriormente se repetem e o paciente requer novo atendimento médico após uma internamento anterior (Lai et al., 2012). Uma readmissão imlica diversos fatores, como novas prescrições ou novo exame, o que representa mais custos e gasto de recursos, tornando-se facilmente um peso financeiro para as instituições de saúde. Mais mais importante, é que cerca de 1 em cada 10 pacientes que recebem alta de UCI em países desenvolvidos acabam por ser readmitidos durante o mesmo internamento (Correa et al., 2017), e essas readmissões tendem não apenas a aumentar os custos financeiros, mas também levar a resultados de saúde negativos para os pacientes readmitidos. Tais resultados incluem o aumento do tempo de permanência, aumento da mortalidade e aumento do suporte hospitalar necessário. Há, portanto, uma clara necessidade de reduzir o número de pacientes readmitidos e, consequentemente, as taxas de readmissão UCI, que são estimadas em torno de 4% a 14% para o paciente habitual, mas podem aumentar substancialmente para pacientes com comorbidades (Lai et al., 2012). As readmissões representam uma falha em fornecer o melhor padrão de cuidados de saúde, e a redução desse risco é do interesse dos hospitais não apenas para diminuir os riscos e custos de saúde associados, mas também para diminuir o tempo total gasto na UCI. Além disso, a taxa de readmissão tem sido proposta como um marcador para medir a qualidade da assistência, podendo também impactar outros marcadores, como tempo de internamento e mortalidade.

A decisão de dar alta a um paciente da UCI pode levar em consideração uma variedade de dados e fatores, todos ponderados pelo clínico responsável. À medida que os hospitais se tornam cada vez mais orientados a dados com a adoção de Registros Eletrónicos de Saúde Registos de Saúde Eletrónicos (RSE), testemunhamos um aumento no desenvolvimento de abordagens computacionais para apoiar a decisão clínica. O uso de RSE como formato de dados para informações clínicas cria a oportunidade de fazer avaliações de risco proativamente e aplicar intervenções eficazes para mitigar esse risco com base em dados (Jamei et al., 2017) porque os médicos registram o estado do paciente por meio de dados qualitativos e quantitativos e observações para que os RSE sejam capazes de capturar informações sobre todos os aspectos do cuidado, garantindo ao mesmo tempo

uma representação clara de acordo com vocabulários controlados padronizados. À medida que mais dados são registados em RSE, as oportunidades de explorar esses dados para desenvolver abordagens preditivas também crescem (Jensen et al., 2012; Goldstein et al., 2016).

Esforços foram feitos no sentido de desenvolver abordagens eficazes de aprendizagem automática para fazer previsões em várias configurações clínicas Intensive Care Unit (ICU) (Suresh et al., 2018), como prever mortalidade e tempo de internamento (Huang et al., 2019), prever sepsis (Scherpf et al., 2019), e previsões para detectar alto risco de readmissão dentro de 30 dias após a alta. Apesar desses esforços crescentes, as abordagens de aprendizagem automática ainda exploram dados de RSE diretamente, sem levar em consideração o seu significado ou contexto. O conhecimento médico não é acessível a esses métodos, que trabalham cegamente sobre os dados, sem considerar o significado e as relações dos objetos de dados. Ontologias e grafos de conhecimento podem ajudar a preencher essa lacuna entre dados e contexto científico, uma vez que são artefatos computacionais que representam formalmente as entidades de um domínio e como elas se relacionam entre si.

Esta oportunidade motivou o objetivo deste trabalho: investigar de que forma o enriquecimento de dados de RSE com anotações semânticas baseadas em ontologias e a aplicação de técnicas de aprendizagem automática que as exploram podem impactar a previsão do risco de readmissão em UCI em 30 dias. Para isso, várias contribuições foram desenvolvidas, incluindo: (1) Um enriquecimento do conjunto de dados MIMIC-III com anotações para várias ontologias biomédicas; (2) Uma nova abordagem para prever o risco de readmissão em UCI, que explora as incorporações do grafos de conhecimento para representar os dados do paciente tendo em consideração as anotações semânticas; (3) Uma variante da abordagem preditiva que visa diferentes momentos para apoiar a previsão de risco durante toda a permanência na UCI.

O trabalho procurou responder a três questões de investigação.

Para a primeira questão **Q1: Como podem os dados do paciente ser devidamente anotados com as ontologias?**, com base nos resultados obtidos pode-se concluir que fazer anotações apenas com a ontologia National Cancer Institute Thesaurus (NCIT) é a melhor solução, e em particular, usar todas as informações disponíveis fornece melhores resultados do que usar apenas as informações de diagnóstico que eram as únicas que, de acordo com o conjunto de dados MIMIC-III, poderiam ser mapeadas para NCIT. Para chegar às anotações, os termos referentes a todas as informações da estadia devem ser mapeados para a ontologia NCIT, com o *NCBO annotator*, para que as classes da ontologia correspondam aos termos e as anotações possam ser extraídas. Embora esses resultados possam parecer inesperados com base simplesmente no ajuste das ontologias, isso porque a solução alternativa para a abordagem apenas com NCIT propõe que, com base no tipo de informação que cada termo fornece, deve ser mapeado para a ontologia específica que descreve essa informação, a abordagem com apenas NCIT superou a solução alternativa, não apenas em precisão e Receiver Operating Characteristic (ROC)-Area Under the Curve (AUC) mas também nas medidas de desempenho Precision/Recall (PR), garantindo que este método não é apenas mais preciso e valioso, mas também mais robusto e reprodutível.

Quanto à segunda pergunta **Q2: Como podem essas anotações ser exploradas pelos embeddings RDF2Vec para melhorar a previsão do risco de readmissão?**, os resultados demonstram claramente que, quando comparados com os embeddings International Classification of Diseases, Version 9 (ICD9) pré-treinados, o RDF2vec com qualquer conjunto de anotações, supera a incorporação do ICD9 nas diferentes experiências. A integração do RDF2vec embedding com anotações RSE, como forma de fornecer enriquecimento semântico, tem um impacto substancial nos resultados da previsão, ajudando a melhorar o desempenho independentemente das anotações ou ontologias usadas para fazer os embeddings (seções 5.5 e 5.6). A melhor configuração de ontologias resulta em um ROC-AUC de 0,82, melhorando 0,2 acima da linha de base.

Para a pergunta final **Q3: Como é influenciado o valor preditivo quando as previsões são feitas em diferentes momentos durante a estadia de um paciente na UCI e com informações variadas disponíveis?** Os resultados mostram que o procedimento pode ser aplicado em um cenário mais realista, com AUC geralmente melhorando à medida que mais informações estão disponíveis e com as melhores abordagens demontrando sempre AUC acima de 0,8.

Atendendo aos resultados positivos obtidos e ao meu interesse pessoal, foi criado um projeto de empreendedorismo para explorar a oportunidade de traduzir os resultados científicos desta dissertação num produto tecnológico. Este projeto incluiu uma série de tarefas: (1) criação de uma ideia de produto; (2) desenvolvimento de um plano de negócios em colaboração com a TecLabs; (3) preparação de um deck de slides para apresentação pública; (4) participação no concurso internacional de empreendedorismo'*H-INNOVA-Health INNOVAtion Award*'. Mais informações sobre o prêmio disponíveis no Hinnova Hub Website.

RedHealth.AI é um produto que visa ajudar os médicos a decidir se um paciente está pronto para receber alta da UCI para evitar mais complicações e reinternamentos. RedHealth.AI apresenta aos clínicos, de forma compreensível, informações chave para ajudar na tomada de decisão e estas incluem; a evolução do risco de reinternamento, pacientes semelhantes, detecção de pacientes de alto risco, gráfico do paciente com informações e relações de permanência e outros. RedHealth.AI foi selecionado como finalista para o *H-INNOVA-Health INNOVAtion Award* entre mais de 170 candidatos de todo o mundo e atingiu o Top 5.

Palavras-chave: Anotação semantica, Ontologias Biomédicas, Readmissões em UCI, Apredizagme automática, Embedding de grafos de conhecimento.

Abstract

ICU readmissions are a critical problem associated with either serious conditions, illnesses, or complications, representing a 4 times increase in mortality risk and a financial burden to health institutions. In developed countries 1 in every 10 patients discharged comes back to the ICU. As hospitals become more and more data-oriented with the adoption of Electronic Health Records (EHR), there as been a rise in the development of computational approaches to support clinical decision.

In recent years new efforts emerged, using machine learning approaches to make ICU readmission predictions directly over EHR data. Despite these growing efforts, machine learning approaches still explore EHR data directly without taking into account its meaning or context. Medical knowledge is not accessible to these methods, who work blindly over the data, without considering the meaning and relationships the data objects. Ontologies and knowledge graphs can help bridge this gap between data and scientific context, since they are computational artefacts that represent the entities in a domain and how the relate to each other in a formalized fashion.

This opportunity motivated the aim of this work: to investigate how enriching EHR data with ontology-based semantic annotations and applying machine learning techniques that explore them can impact the prediction of 30-day ICU readmission risk. To achieve this, a number of contributions were developed, including: (1) An enrichment of the MIMIC-III data set with annotations to several biomedical ontologies; (2) A novel approach to predict ICU readmission risk that explores knowledge graph embeddings to represent patient data taking into account the semantic annotations; (3) A variant of the predictive approach that targets different moments to support risk prediction throughout the ICU stay.

The predictive approaches outperformed both state-of-the-art and a baseline achieving a ROC-AUC of 0.815 (an increase of 0.2 over the state of the art). The positive results achieved motivated the development of an entrepreneurial project, which placed in the Top 5 of the H-INNOVA 2021 entrepreneurship award.

Keywords: Anotação semantica, Ontologias Biomédicas, Readmissões em UCI, Apredizagme Semantic annotation, Biomedical Ontologies, ICU readmissions, Artificial intelligence, Knowledge graph embeddings.

Contents

1 Introduction 1
1.1 Motivation 1
1.2 Goals 3
1.3 Contributions 3
1.4 Structure of the document 5

2 Concepts 7
2.1 Ontologies 7
2.2 Knowledge graphs 8
2.3 Semantic Annotations 9
2.4 Knowledge Graph Embeddings 10
2.5 Machine Learning 12
2.6 Data Cleaning 13

3 Related Work 15
3.1 Semantic Annotations for EHR 15
3.2 Mining EHR 16

4 Methods 21
4.1 Data Collection and Pre-processing 23
4.1.1 Data Acquisition and Filtering 23
4.1.2 Data Cleaning and Pre-Processing 25
4.1.3 Snap Shot split 27
4.2 Semantic Feature Generation 28
4.2.1 Ontology Selection 29
4.2.2 Ontologies and Thesauri 30
4.2.3 Semantic Annotations 32
4.2.4 Knowledge Graph Embeddings 35
4.3 Readmission Prediction 36
4.3.1 Prediction Task 36
4.3.2 Prediction Evaluation 36

5 Results and Discussion **39**
5.1 Data Cleaning Outcomes . . . 39
5.2 Experimental Design . . . 39
5.3 Reproduction Results . . . 41
5.4 RDF2vec Embeddings and Diagnosis Annotations . . . 42
5.5 NCIT Annotations for all the information . . . 43
5.6 NCIT replacement with MIMIC-III proposed Ontologies . . . 44
5.7 ICU stay simulation . . . 46
5.8 KG embeddings contribution analysis . . . 49
5.9 Entrepreneurship Projection . . . 50

6 Conclusions, Future Work and Limitations **53**
6.1 Conclusions . . . 53
6.2 Limitations . . . 55
6.3 Future Work . . . 55

References **64**

Appendices **65**

A Folds Average Performance **67**

B Ontology coverage **69**

C Embedding Combination approaches **73**

List of Figures

1.1 Relations captured between terms with the semantic component of the data. 4

2.1 Knowledge Graph (KG) representation of a concept, from the NCIT ontology with the edges and nodes that define him. 8

2.2 Graphic explanations of a general graph embedding exploration 11

4.1 Schema representation of the methodology. 22

4.2 Prediction moments on an ICU stay, according to the MIMIC-III data set, as well as the new information added in each moment 28

4.3 Amount of terms covered with each individual ontology, for all the information used on the method. 30

4.4 Elastic Search query format, to target label and synonyms of a term. . . . 33

4.5 Distribution of the NCIT Elastic Search matched classes, with Levenshtein Distance to measure similarity. 34

5.1 One fold representative ROC and PR values for the reproduction of the (Lin et al., 2019) method. 41

5.2 One fold representative ROC and PR values for the method replacing just the embeddings with RDF2vec embeddings and the ICD9 information with NCIT information. 43

5.3 One fold representative ROC and PR values for the method mapping all the information gathered during an ICU stay, to the NCIT Ontology. . . . 44

5.4 One fold representative ROC and PR values for the method mapping each set of the information gathered during an ICU stay, to the matching ontology proposed on the MIMIC-III data set. 45

5.5 One fold representative ROC and PR values for the different stages of an ICU, with the data mapped to the NCIT ontology. 48

5.6 Comparison of the prediction values at discharge on a normal implementation and without chart events 49

5.7 Comparison of the discharge predictions with RDF2Vec embeddings and without RDF2Vec embeddings 50

5.8 Preliminary interface for the RedHealth.AI project. 51

B.1 Amount of initial diagnoses terms covered by each ontology. 70
B.2 Amount of drug prescriptions terms covered by each ontology. 70
B.3 Amount of laboratory events terms covered by each ontology. 71
B.4 Amount of procedures covered by each ontology. 71
B.5 Amount of final diagnoses terms covered by each ontology. 72

C.1 Model performance values using weighted average as the embedding combination approach. 74
C.2 Model performance values using multiplication as the embedding combination approach. 75
C.3 Model performance values using the absolute values as the embedding combination approach. 76
C.4 Model performance values using the concatenation of the values as the embedding combination approach. 77

List of Tables

1.1 Predicted estimates of the costs per categories of disease (Carey and Stefos, 2015) 2

3.1 Pool of papers that respect the selection criteria for the introduction on the mining EHR related work. 19

4.1 Demographic features used on the model, with the possible values each feature can have. Source: Lin, Yu-Wei et al.(Lin et al., 2019) 24
4.2 Sample of Prescriptions table from the MIMIC-III data set. Source: https://mit-lcp.github.io/mimic-schema-spy/index.html. 24
4.3 Sample of Lab events from the MIMIC-III data set. 25

5.1 Difference in paths found when doing RDF2vec Embeddings with just NCIT or with the multiple ontologies proposed on the MIMIC-III 46

A.1 Average performance results for the reproduction. 67
A.2 Average performance results for one ontology strategy at admission. . . . 67
A.3 Average performance results for one ontology strategy at treatment. . . . 68
A.4 Average performance results for one ontology strategy at diagnostic. . . . 68
A.5 Average performance results for one ontology strategy at discharge. . . . 68
A.6 Average performance results the multiple ontology strategy. 68

Acronyms

UCI Unidades de Cuidados Intensivos

RSE Registos de Saúde Eletrónicos

ICU Intensive Care Unit

EHR Electronic Health Records

AI Artificial Intelligence

MedDRA Medical Dictionary for Regulatory Activities Terminology

ML Machine Learning

SVM Support Vector Machine

ROC Receiver Operating Characteristic

AUC Area Under the Curve

LR Linear Regression

RF Random Forest

NB Naive Bayes

TPR True Positive Rate

FPR False Positive Rate

PR Precision/Recall

PRC Precision/Recall Curve

NCIT National Cancer Institute Thesaurus

ICD9CM International Classification of Diseases, Version 9 - Clinical Modification

ICD9 International Classification of Diseases, Version 9

LOINC Logical Observation Identifier Names and Codes

NDC National Drug Code

DRON The Drug Ontology

KG Knowledge Graph

NLM U.S. National Library of Medicine

FDA US Food and Drug Administration

ChEBI Chemical Entities of Biological Interest

NCBO National Center for Biomedical Ontology

URI Uniform Resource Identifier

RDF Resource Description Framework

CNN Convolutional neural network

RNN Recurrent neural networks

LSTM Long short-term memory

LOCF Last-Observation-Carried-Forward

SNOMEDCT Systematized Nomenclature of Medicine - Clinical Terms

MeSH Medical Subject Headings Thesaurus

EFO Experimental Factor Ontology

NCR Neural Concept Recognizer

HPO Human Phenotype Ontology

KNN K-Nearest Neighbors

Chapter 1

Introduction

1.1 Motivation

Intensive Care Unit ICU admissions are usually associated with either serious conditions, illnesses or complications due to illness. ICU readmissions happen because serious illnesses previously detected reoccur and the patient requires new medical attention after a previous admission (Lai et al., 2012). Due to factors implied on a readmission like new prescriptions or new examination, ICU readmissions represent further costs and expenditure of resources, easily turning them in a financial burden to health care institutions. About 1 in every 10 patients discharged from ICU units across developed countries end up being readmitted during the same hospital stay (Correa et al., 2017), and these readmissions tend to not only increase the financial costs but also lead to poor health outcomes to the patients readmitted. Such outcomes include increased length of stay, increased mortality and increased in-hospital support needed. There is thus a clear need to reduce the number of readmitted patients and consequently the ICU readmission rates, which are estimated to be around 4% to 14% for the usual patient but can increase substantially for patients with comorbidities (Lai et al., 2012). Readmissions represent a failure to provide the best standard of care, and the reduction of this risk is in the best interests of the hospitals not only to lower associated health risks and costs, but also to decrease total time spent in the ICU. Moreover, the rate of readmission has been proposed as a marker to measure the quality of care, and it can also impact others markers such as length of stay and mortality.

In the United States of America, about 5 million patients are readmitted to ICU units per year costing an extra \$7000 per patient in each stay (Carey and Stefos, 2015), and the hospital systems could save an expected \$2140 per each average 30-day readmission avoided, leading to an expected saving of millions of dollars which could be put to better use. Some specific departments like cardiology have even more to gain avoiding readmissions as some of the most frequent diagnosis found in readmission episodes can be associated with this specific department (Lai et al., 2012), as shown in Table 1.1, where the cost of both Heart Attack and Heart Failure have both bigger costs for readmission,3.432$ and 2.488$ respectively, and for a full hospitalization, 10.814$ and 8.607$, than the overall groups.

Table 1.1: Predicted estimates of the costs per categories of disease (Carey and Stefos, 2015)

	Readmission Probability	Hospitalization Cost ($)	Readmission Cost ($)
Overall	19.6%	8.940	2.140
Heart Attack	29.3%	10.814	3.432
Heart Failure	24.9%	8.607	2.488
Pneumonia	21.2%	9.047	2.278

The decision to release a patient from the ICU can take into account a variety of data and factors all weighed by the clinician in charge. As hospitals become more and more data-oriented with the adoption of Electronic Health Records EHR, we have witnessed a rise in the development of computational approaches to support clinical decision. The use of EHR as the data format for clinical information creates the opportunity to proactively make risk assessments and apply effective interventions to mitigate that risk based on data (Jamei et al., 2017) because medical practitioners record the patient status through qualitative and quantitative observations so the EHR is capable of capturing information on all aspects of care, whilst ensuring a clear representations according to standardized controlled vocabularies. As more data is collected in EHR, the opportunities to explore this data to develop predictive approaches also grow (Jensen et al., 2012; Goldstein et al., 2016).

Efforts have been made to develop effective machine learning approaches to make predictions under several ICU clinical settings (Suresh et al., 2018), such as predicting mortality and length of hospital stay (Huang et al., 2019), predicting sepsis (Scherpf et al., 2019), and predictions to detect high risk of readmission within 30 days from discharge. Despite these growing efforts, machine learning approaches still explore EHR data directly without taking into account its meaning or context. The medical knowledge is not accessible to these methods, who work blindly over the data, without considering the meaning and relationships the data objects. An example of missed context can impair the exploration of EHR data can be seen when comparing two diagnoses: **'Aortic Valve Disease'** and **'Coronary Artery Disease'**. Using categorical analysis these two diagnosis have no similarity, and with a string similarity analysis they have low similarity, sharing only the less informative word 'disease'. However, it is common knowledge that these two diagnoses are closely related. When controlled vocabularies are used, we gain an extra layer of information given by the standardization (two entries with the same code mean the same thing) and also by the hierarchy that organized the vocabulary. However, controlled vocabularies are limited in their contextual richness, and moreover, in a single EHR multiple domains can be covered by different controlled vocabularies which makes their concerted analysis more difficult.

1.2 Goals

Ontologies can help bridge this gap between data and scientific context, since they are computational artefacts that represent the entities in a domain and how the relate to each other in a formalized fashion. (Gruber, 1994). Biomedical ontologies have become quite popular in the last decades to support the annotation of the massive amounts of data produced by gene sequencing technologies. At the same time, clinical ontologies have also been developed to tackle the limitations of controlled vocabularies and allow for a fuller semantic representation. The opportunity here is that by linking EHR data to the ontologies through semantic annotations, we can feed this extra later if information about the meaning of the data to machine learning systems. Looking again at the example of the **'Aortic Valve Disease'** and **'Coronary Artery Disease'** relation, now with the semantic components is possible to understand that they both are for instance, heart diseases, and non-neoplastic heart disorders, sharing after all a considerable amount of similarity and relationships, facts that are hidden on raw analysis (Fig 1.1).

In healthcare, the existence and detection of correlations grants the ability to predict future patient outcomes, and predictive clinical modelling uses machine learning methods to build models from clinical data and make inferences on the data (Goldstein et al., 2016). The addition of data descriptions and semantic annotations might be very impacting because they will supply richer information to the machine learning models and likely lead to an improvement on the prediction results. However most of the studies that handle EHR data for predictions on clinical scenarios (Huang et al., 2019; Scherpf et al., 2019; Suresh et al., 2018), ignore the semantic context of the data.

The aim of this work is then to investigate how enriching EHR data with ontology-based semantic annotations and applying machine learning techniques that explore them can impact the prediction of 30-day ICU readmission risk.

Additionally, there is an opportunity to build models that more closely align with the reality of the ICU. To achieve this, the project also aims to establish models that work with information limited to specific points in time of an ICU stay. So instead of making predictions only for the moment of discharge, predictions are made throughout the stay, allowing clinicians and health care practitioners to keep track on the 30-day risk of readmission and have it updated as more information on the patient becomes available.

1.3 Contributions

The main contributions of this work are:

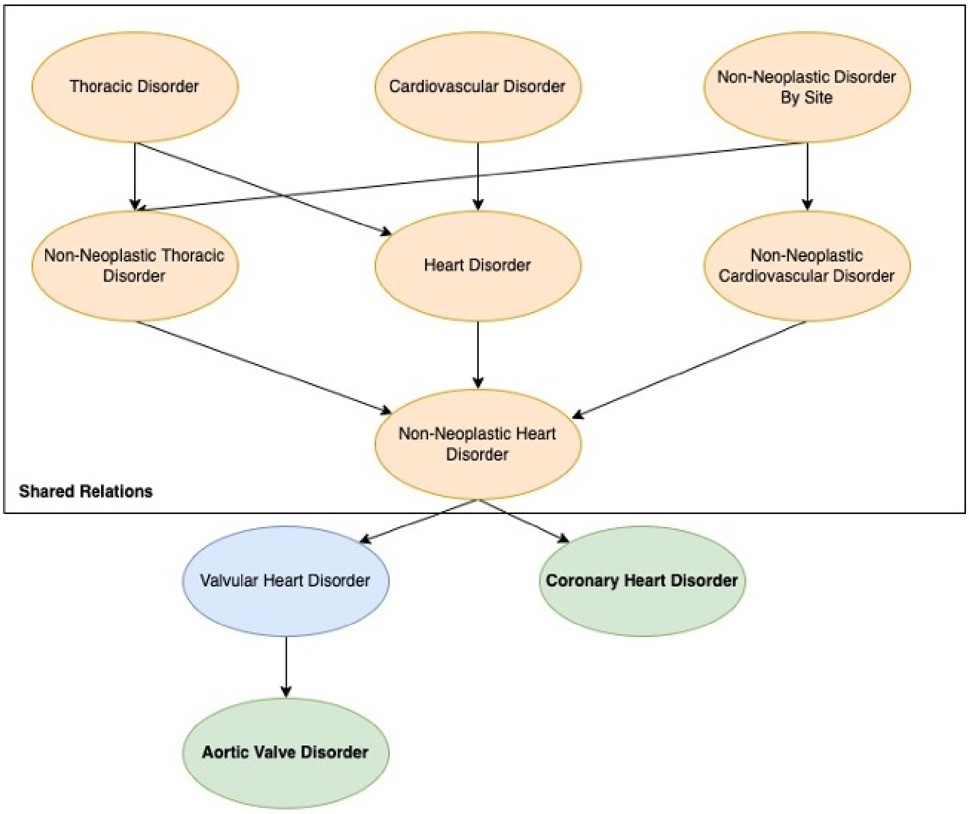


Figure 1.1: Relations captured between terms with the semantic component of the data.

- An enrichment of the MIMIC-III data set with annotations to several biomedical ontologies. Available in our GitHub repository.
- An ElasticSearch querying system, capable to match input terms to NCIT classes with adjustable similarity score thresholds.
- A novel approach to predict ICU readmission risk that explores semantic information and improves on the state-of-the-art.
- A variant of the predictive approach that targets different moments to support risk prediction throughout the ICU stay, and that is suitable to be used by health care providers during an the stay.
- A business model plan based on the technological and scientific advances achieved in the thesis which placed in Top 5 of the H-INNOVA 2021 entrepreneurship award.
- An oral presentation, consecrated with best oral communication award at the 11^{th} edition of the bioinformatics open days 2022.

1.4 Structure of the document

Going forward, the document is organised as follows:

- Chapter 2 - Defines and explains the fundamental concepts needed to understand the employed methods and techniques.
- Chapter 3 - Surveys the works developed in the various fields covered by this work.
- Chapter 4 - Presents the methods developed in this work, with descriptions of the tools and procedures, as well as clarifications on the reasoning behind the steps taken and resources used.
- Chapter 5 - Presents and discusses the results obtained, according to the experimental steps established, with brief resolutions on the reasoning for the results.
- Chapter 6 - Summarizes the main conclusions of this work, contextualizing the outcomes, presenting some of the limitations and reflecting on the future of the work.

Chapter 2

Concepts

2.1 Ontologies

Since the later part of last century and particularly the 1990's, ontologies became a topic of interest on small communities within the Artificial Intelligence (AI) field of investigations and later the interest grew to other areas of knowledge such as information integration or knowledge management. This growing interest is due to the promise that ontologies allow for a common understanding of domains, shared across computers and people (Studer et al., 1998), promoting connectivity and accessibility to knowledge.

An ontology can be described as a specification of a conceptualization (Gruber, 1993), meaning that they define a set of representational primitives to model a domain of knowledge or discourse, these primitives include detailed information about their meaning and constraints on their logically consistent application. These specification need to be explicit, so that the concepts used, and the constraints are explicitly defined, thus preventing illogical relationships or definitions, but also formal, so that the ontology is machine readable (Studer et al., 1998). The utility of capturing relationships between concepts is that they convey semantics, identifying how concepts relate in the hierarchical knowledge space (Schuurman and Leszczynski, 2008).

Due to the tremendous progress in the domain of biological sciences there has been a rapid increase in the amount of available biological data. To handle this increase, ontologies are increasingly used for biological data annotation since domain-specific knowledge can be encoded in ontologies (Konopka, 2014). In a biomedical scenario, the power of ontologies lies in their capacity to provide context for biological semantics, being semantic models for real biological domains.

Ontologies provide controlled vocabularies, but so do thesauri, which define terms, synonyms and alternative spellings (Adams et al., 2014). A thesaurus describes terms that are organized into a hierarchical structure and adds non-hierarchical relationships between concepts and other properties to each concept. Ontologies take a step further and express axioms and restrictions, thus meaning that a thesaurus are less expressive

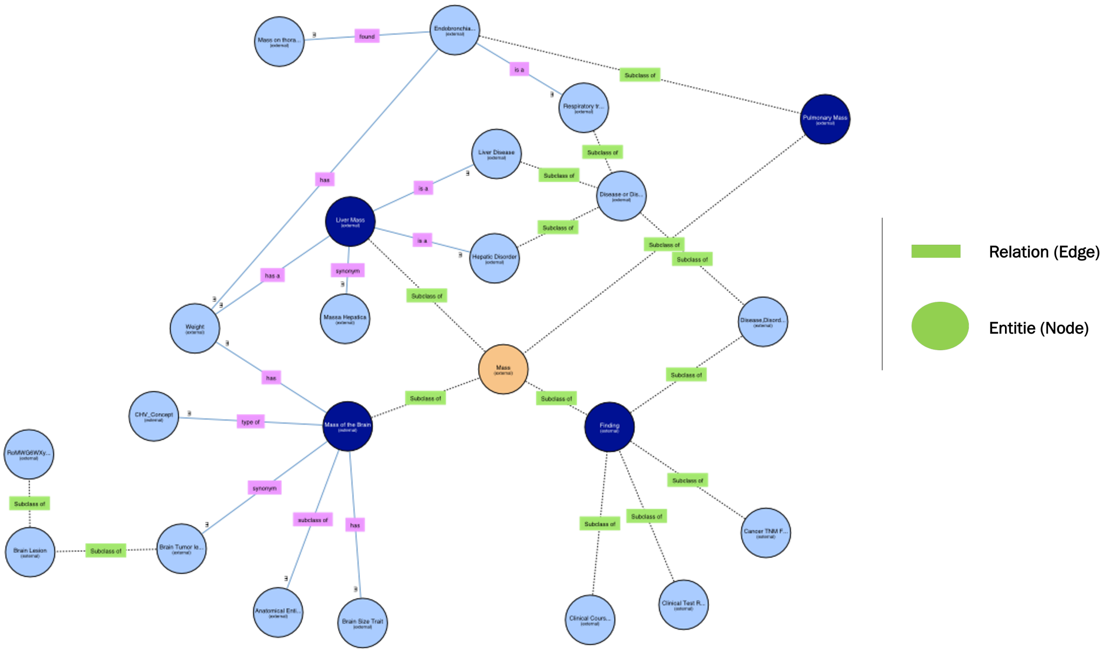

Mass on thora... (external)
found
Endobronchia... (external)
is a
Respiratory tr... (external)
Subclass of
Pulmonary Mass (external)
has
Liver Disease (external)
is a
Liver Mass (external)
Disease or Dis... (external)
is a
Hepatic Disorder (external)
synonym
has a
Massa Hepatica (external)
Weight (external)
Disease,Disord... (external)
Mass (external)
has
CHV_Concept (external)
type of
Mass of the Brain (external)
Finding (external)
RoMWG6WXy... (external)
synonym
subclass of
Brain Lesion (external)
Brain Tumor le... (external)
Cancer TNM F... (external)
Anatomical Enti... (external)
Brain Size Trait (external)
Clinical Cours... (external)
Clinical Test R... (external)
∃
Relation (Edge)
Entitie (Node)

Figure 2.1: KG representation of a concept, from the NCIT ontology with the edges and nodes that define him.

than ontologies, though still useful (Adams et al., 2014). Many thesauri have been re-engineered as ontologies, attending to the differences and similarities between controlled vocabularies and ontologies (Adams et al., 2014).

The knowledge that ontologies (and thesauri) provide may be used in predictive models without prior data analysis or mining, and the major benefit they may have is the possibility of enriching or expanding features, making information available for the machine learning methods that would not be available without relying on ontologiesKulmanov et al. (2020). This is especially true in the life science domain where there are more than nine-hundred ontologies available, spanning almost all domains of biological and biomedical research. With this growth biomedical ontologies are able to provide controlled vocabularies for characterizing majority of biological phenomena with formalized domain descriptions and link them to other related domains Smith et al. (2007).

2.2 Knowledge graphs

A KG represents the relationships between real-world entities in a graph structure, thus meaning the graph gives a conceptualization of a entity based on the entities and relationships presented on the graph (Ehrlinger and Wöß, 2016). A KG is composed of entities (nodes) and relations (edges) like *partOf* or *isA*. An KG edge is represented as a triple or fact, structured as head entity, relation and tail entity (*subject-predicate-object*), indicating that two entities are connected by a specific relation Wang et al. (2017).

Ontologies are essential elements of the KG, giving them a structured representation of the domain (Ehrlinger and Wöß, 2016). Ontologies formally define the types of entities and relations a KG can represent (Auer et al., 2018).

Resource Description Framework (RDF) is the most popular language to define a KG(Ontotext, 2021). It connects data pieces via the three positional statements on the edge, *subject-predicate-object*, these triplets can with the help of RDF statements, express just about any facts, relationships and data in a structured and uniform manner. The classes, predicates and named graphs on an RDF representations are all defined as Uniform Resource Identifier (URI)'s. This way they can appear as nodes in the graph, get their descriptions, i.e. instance data and schema can be managed and accessed in an uniform model . Employing RDF to encode the KG supports expressiveness, formal semantics, good performance,interoperability and standardization (Wang et al., 2017).

2.3 Semantic Annotations

Semantic annotation is the process of annotation of an object by associating it with concepts that have well-defined semantics(Jovanovic and Bagheri, 2017). Semantic annotations provide a way to enrich the data since they are used to discover and add information to data, thus meaning that we can to real entities associate links to their semantic description (Kiryakov et al., 2004). In healthcare data, although large parts of the records exist as structured data, the use of dis-coordinated vocabularies is common and moreover, a significant proportion exists as unstructured free texts (Roberts et al., 2007).

For the unstructured data, semantic annotators first extract the entities from clinical data such as EHR and then couple them to the appropriate corresponding entry on the biomedical ontology, thus getting the semantic description. For instance, in the sentence "**The patient has a brain tumor**" the annotator will recognize "**brain tumor**" as an entity and identify it as a disease by connecting it the concept **C0006118** of the Medical Dictionary for Regulatory Activities Terminology (MedDRA), that corresponds to the semantic description of "**Brain tumor**" thus also inheriting all the other descriptions associated with this entity and making the semantic context of the EHR data available (Jovanovic and Bagheri, 2017). For the structured data, the terms in the EHR belong to controlled vocabularies used to annotate the domain, facilitating the match.

There are several challenges in the semantic annotation of EHR:

- Clinical text often contains incorrect grammar and does not respect syntactic or spelling rules.

- Biomedical terms are often polysemous and thus prone to ambiguity.

- Biomedical terms are often replaced by abbreviations or acronyms that tend to be ambiguous.

To address these challenges, semantic annotators (Jonquet et al., 2009; Tanenblatt et al., 2010). often rely on a combination of text processing, large-scale knowledge bases, semantic similarity measures and machine learning techniques (Jovanovic and Bagheri, 2017). In EHR semantic annotation is typically based on a term-to-concept matching approach or approaches based on Machine Learning (ML) methods. Term-to-concept matching also referred to as dictionary lookup, is based on matching specific segments of text to a structured dictionary, knowledge base, or ontologies. The drawback of the annotators that implement this approach is the lack of disambiguation due to the fact that the terms recognized in texts are connected with several possible meanings. This problem is commonly addressed by using a human- based decision where once the match is done, one of the meanings is chosen as the most adequate to the particular purpose of the research.

2.4 Knowledge Graph Embeddings

Embedding is a technique that transforms a higher dimensional space into a lower dimensional space (Ristoski et al., 2018). One example being the transformation of high dimensional vectors, like the ones that represent all the occurrences of a word with all the other words in a corpus, into smaller vectors that retain meaningful properties of the original space 2.2. Embeddings are particularly useful since they allow the construction of continuous word vectors that capture both syntactic and semantic information of terms that can enrich models with information. Once the information is embedded in vectors with a smaller dimensional space, it tends to use fewer memory resources and most importantly it stores the information in one place, stopping the spacing of information that happens when the information on a database is in distant spaces making the access difficult.

The idea behind using embeddings to preserve particular structures, is that the second structure may enable different or additional operations which are not possible in the first structure. This type of embedding can applied to KGs allowing the information in the KG to be exploited by ML models (Craven et al., 2006). While there are many structures in which ontologies can be embedded, the main interest for this project is in embedding ontologies within real-valued vector spaces so that modern optimization and ML algorithms can be applied to perform prediction tasks.

KG embeddings embed the KG components into continuous vector spaces, so that the manipulation is simplified but at the same time preserve the inherent structure of the KG (Wang et al., 2017). A typical KG embedding technique has three steps; the first specifies the way entities and relations are represented on the vector space, entities are usually represented as vectors and relations are taken as operations and represented as vectors or matrices; the second step defines scoring function on each fact to measure the

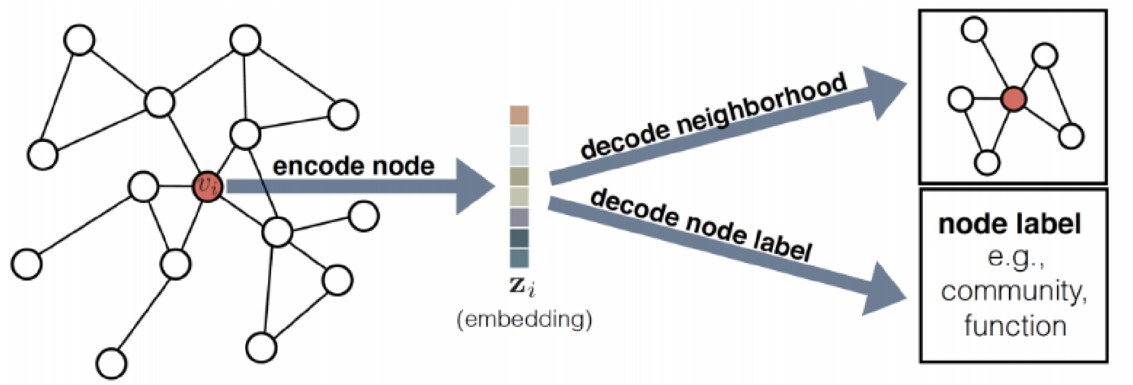


Figure 2.2: Graphic explanations of a general graph embedding exploration

plausibility, where facts observed on the KG tend to have higher scores than not observed facts; and on the final step, to learn the entity and relation representations, this step solves the optimization so that the plausibility of the observed facts is maximized (Wang et al., 2017). The two main types of KG embeddings are translational embeddings and the random walks embeddings. (Wang et al., 2017).

In KG random walks embeddings random walks are used to explore the neighborhood of each node in the graph creating a number of paths or walks in the graph. The set of walks is used as the basis of the embeddings (Kulmanov et al., 2020). Some potential methods to perform random walks are:

- DeepWalk (Perozzi et al., 2014) generates a corpus of sentences with random walks from each node, and then applies Word2Vec on the resulting corpus to obtain embedding vectors.
- Node2Vec (Grover and Leskovec, 2016) is a model that explores the original graph through biased random walks and can force walks to remain within a certain distance of the origin node.
- OWL2Vec (Chen et al., 2021) is a model that uses the biased random walks from Node2Vec to embed graphs generated from axioms.
- RDF2Vec (Ristoski et al., 2018) uses language modeling approaches to extract features from word sequences, and adapts them to RDF graphs.

The other type of graph embeddings are the translational embeddings, and these methods are a family of representation learning methods which model relations in the knowledge graph as translation operations between graph node embeddings. $f\eta$ being a graph embedding, the methods represent knowledge graphs as a set of three edges (s, p, o) and define a translation operation which translates $f\eta(s)$ to $f\eta(o)$ depending on the relation p (Jovanovic and Bagheri, 2017). Some examples of methods that do translational embeddings are:

- TransE uses a vector representation for relations that have the same dimensions as vectors representing nodes, and define the translation operation as the addition of the relation vector to the node vector.

- TransH extends TransE by moving the translation operation to a relation- specific hyper-plane.

2.5 Machine Learning

Machine learning is a common term used to refer a broad range of algorithms that perform intelligent predictions on a data set (Nichols et al., 2018). There is a large variety of ML algorithms, also referred as *models*.In supervised learning, objects with a set of features are presented with the desired outcome to be learned, so the functions of ML map the input to an output based on exemplar input-output pairs. To do, this algorithms split the data in two portions; training data, to find the parameters that produce model results. The other portion, the test data, used to assess the performance of the model and must not be used to influence the parameters or other wise will compromise the model. Some practitioners also reserve a portion of validation data, used to select the best performer of a collection of models that may use completely different algorithms or training methods (Nichols et al., 2018).

The choice of a particular model is determined by the data characteristics and by the outcome desired, which in the case of can be a classification, meaning a prediction of a qualitative label, or a regression, a prediction of a continuous variable (Nichols et al., 2018). For instance, smaller data sets perform better with robust classical techniques like Linear Regression, or Decision-tree methods, and larger data sets require more computational demanding deep learning algorithms like Convolutional neural network (CNN) or Long short-term memory (LSTM) (Nichols et al., 2018).

In terms of supervised learning and considering classic techniques, some of the examples are:

- **Linear Regression (LR)**, that is the simplest form of machine learning(Nichols et al., 2018). The model assumes the linear function $f(x, \theta) = \beta x + m$ where β is the slope and m the intercept, and the calculation of the slope and intercept is training the model and this are calculated with simple closed-form calculations (Nichols et al., 2018).

- **Support Vector Machine (SVM)**, this is a powerful method that creates a decision boundary between two classes and makes prediction from one ore more vectors. The decision boundary is known as hyper-plane, and it is oriented to be as far way as possible from the closest data point from each of the two classes. For SVM a labeled training data set can be: $(x_1, y_1), ..., (x_n, y_n), x_i \in R^d$ and $y_i \in (-1, +1)$, where x is the feature vector, y the class label for i. The optimal hyper-plane can than be describes as: $wx^t + b = 0$, where w is the weight vector, x is the input feature vector, and b is the bias (Huang et al., 2018).

- **Naive Bayes (NB)**, is a simple algorithm that is based on Bayes's rule $P(y|x) = P(y)P(x|y)/P(x)$, that assumes that the attributes are conditionally independent given the class. And although the Independence assumption is often violated, the NB still offers competitive and accurate classifications. NB provides a mechanism that estimates the posterior probability $P(y|x)$ of each class y given an object x, and this estimations are used for classification or implemented in other decision support applications (Webb et al., 2010).

- **Random Forest (RF)**, has is a versatile and accurate classifier, that requires little tuning and is able to provide valuable outputs. RF is a collection of classification and regression trees ensemble, trained on data sets randomly re-sampled from the training data set it self, with the same size as the training set, *bootstraps*. RF follows specific rules for tree handling and development, thus it is robust to over-fitting and more stable in the presence of outliers (Sarica et al., 2017).

2.6 Data Cleaning

A pretty well established conception on the scientific world, is that the insights and analysis made are only as good as the data used to make them, and that good data provides good insights and bad data provides bad insights as well as unreliable analyses (Chu et al., 2016). Thus meaning that treating and cleaning the data is one of the most important steps to any data implementation, so that data quality is ensured.

Data cleaning is the process of fixing and removing unwanted data form a data set. This can include incorrect, corrupted,duplicate, incomplete or deformed data. When dealing with real-world data collected daily, they most likely include lots of typos and errors and the data cleaning is time-consuming, and requires to master techniques of data processing (Shi et al., 2021). Therefore, a establish template is crucial for the data cleaning process, to ensure it is done the right way every time.

There is no one absolute way to prescribe the exact steps in the data cleaning process, because it depends on the data being treated and on the goal desired. The only certainty is that to every target error, two stages exist, the detection of the error and the repairing of the error. With this, a proposes guideline establishes the steps as follow:

- **Filter Unwanted Data**: Some observation may not fit the purpose of the investigation, this because hither the data does not fit the analyses or some restrictions prevent the use of some information. An example may be an age restriction that prevents the use of certain patients information or event all the information regarding that particular individual (TABLEAU SOFTWARE, 2021).

- **Fix Structural Error**: These are a kind of structural errors, where strange conventions,typos or capitalization happen undesirably. An example of this being the presence on ’N/A’ and ’Not Applicable’, that are different and should be analyzed in the same category (TABLEAU SOFTWARE, 2021).
- **Remove Duplicates**: These happen when two or more occurrences, of the same information happen undesirably on the data set. Duplicate observations are the error occurring most often on the data collection, because when combining data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data (TABLEAU SOFTWARE, 2021).
- **Handle Missing Data**: A vast majority of algorithms can not deal with missing data, and the ones that can most of the time have poor outcomes. Thus stretching the importance of dealing whit this types of data. The solution to solve this error, is to ether drop the information, based on other observations or input the missing information with the most common value, with the risk of losing integrity (TABLEAU SOFTWARE, 2021).

Data cleaning is thus an effective solution to fight data errors and inconsistencies, and although the errors it can fine are set, the way they are solved is arbitrary.

Chapter 3

Related Work

In this dissertation, the method takes advantage of all EHR as to offer to make ML predictions. However these types of data have challenges as they are complex to deal with, leading some times to volatile results that may vary from prediction to prediction. The correct mining of EHR, correct annotation, correct ML implementation, and use of ontologies for all the processes is fundamental to prevent such volatile result. Once proper methods are selected, stability can be achieved and the predictions can be made.

The following section presents what are related works on annotation for EHR and handling of EHR.

3.1 Semantic Annotations for EHR

The annotation of EHR is still challenging and variations in performance happen often, usually due to the chosen annotation method and the knowledge graphs. To solve thisGazzotti et al. (2020) propose a ML approach based on DBpedia to extract medical subjects from EHR and evaluate the impact of augmenting the features used to represent EHR with these subjects, with access to the metadata of the relations and descriptions, in the task of predicting hospitalization. To improve the DBpedia spotlight detection, both words and abbreviations are used, because with the proper rule set and semantic similarity the annotator can recognize the instances and retrieve classification labels, capturing the complexity of both structured and unstructured data Gazzotti et al. (2020).

The annotation of structured EHR data although complex is not as challenging as the annotation of unstructured data (Arbabi et al., 2019). To handle this Arbabi et al. (2019) present a ML approach for automated medical concept recognition on unstructured data called Neural Concept Recognizer (NCR). The NCR is trained on the information provided by the ontologies, such as concept names, synonyms and taxonomic relations between concepts, and does so using a CNN that predicts if a phrase or a specific word is synonymous to a concept on an ontology of choice, providing a fitting term. This work in particular benefited from the information provided by the Human Phenotype Ontology (HPO) and Systematized Nomenclature of Medicine - Clinical Terms (SNOMEDCT), but it can be used with other ontologies. Most concepts cannot generalize well to unseen synonyms and typically require large corpora of annotated text that cover all classes of concepts. The great advantage of NCR is that it can be efficiently generalized to new synonyms without the need to large corpora of annotated text Arbabi et al. (2019). However, the method in unable to utilize contextual information for concept recognition, limiting the performance.

3.2 Mining EHR

The following is focused on the use of ML over the MIMIC-III data set which is openly available. This supports reproducibility and comparison of the results. Moreover, MIMIC-III contains categorical and textual data that can be subject to semantic annotation. These works will be used to find a comparison basis for the impact of the methodology being proposed on this dissertation.

A search of works in this area focused on: (1) The methodology must use ML resources to make predictions over EHR data and in particular use the MIMIC-III data set; (2) The prediction tasks of the ML approaches are health related and are done to predict an outcome or probability related to the patient's health and care.

The pool of relevant works found is very diverse and represented in Tab 3.1

Javan et al. (2019) predict cardiac arrest on adult patients with sepsis using 30 h clinical data of every sepsis patients on the Mimic III database. This method in particular used EHR data with classical ML techniques (SVM, Decision Tree, LR, K-Nearest Neighbors (KNN), GaussianNB) and ensemble methods (Gradient Boosting, XGBoost, RF, Balanced Bagging Classifier and Stacking). This work shows high potential to use ML in a prognostic systems for sepsis patients, but is limited to the ICU information.

(Lu et al., 2019) predicted the readmission of a patient in the ICU using clinical data from the database MIMIC-III. (Lu et al., 2019) used patient discharge reports and represents them with multi-view graphs enhanced by the Unified Medical Language System Metathesaurus. These are coupled with a graph CNN for representation learning used to predict ICU readmission. This work demonstrates the importance of incorporating knowledge graphs into ML pipelines for clinical prediction.

Xu et al. (2019) studied the impact of class imbalance on the performance of Acute Kidney Injury prediction. This work happens because research with patient's EHR, usually has class imbalance problems, where the number of Acute Kidney Injury prediction cases is usually much smaller than the controls. The method proposed a class balancing that lead to improved ML prediction, and showed that class balancing could improve EHR ML approaches.

Beaulieu-Jones et al. (2018) takes advantage of the small batches were clinicians perform and record actions to trace a patients trajectory during a stay. Using MIMIC-III as the EHR data set to extract features and unsupervised ML approaches to learn regular embeddings, predict the patient's survival within a 1-year period from admission with traditional and deep learning ML algorithms. Although successful this approach and deep learning approaches in general require a large patient sample to outperform the traditional learning approaches.

Anand et al. (2018), does a predictive modeling to examine the relative influence of diabetes, diabetic health maintenance, and comorbidities on outcomes in ICU patients. This work uses the MIMIC-III database, machine learning and binomial logistic regression modeling to predict risk of mortality in ICU patients, with the associated comorbidities and diabetic conditions. This work uses both RF and LR models and with positive results is able to improve the predictions.

Lin et al. (2019), uses machine learning methods clinical data from the MIMIC-III to predict the ICU readmission of patients within 30 days of their discharge. This method incorporate multiple types of clinical data features and pre-maid ICD-9 embeddings, on recent machine learning techniques, such as Recurrent neural networks (RNN) with LSTM, to incorporate multivariate features of EHR and capture sudden fluctuations on clinical measurement. The proposed LSTM-based solution can better capture high volatility and unstable status in ICU patients, granting ML models the ability to improve our ICU decision-making accuracy. This work in particular is set to have substantial clinical impact by augmenting clinical decision-making for physicians and ICU specialists.

Although none of these work successfully use KG embeddings to account the meaning or context of EHR, they ether use other kinds of embeddings obtained from a series of events, or base on a text corpora, and none of them take advantage of the graph structures to make embeddings. The work (Lin et al., 2019), was chosen as the starting point of the methodology since this method is the only with a the Python code publicly available, which makes their experiments easily reproducible and comparable. Moreover, (Lin et al., 2019) and (Lu et al., 2021) are the only ones to include clinical text in the form of clinical notes and discharge reports, respectively, in the embeddings. However, since this work specifically targets clinical notes, (Lin et al., 2019) is a better basis for developing the methodology than (Lu et al., 2019).

Lin et al. (2019) use machine learning approaches to predict ICU readmission within a 30 day period from discharge, to do so the methods are done on comprehensive and longitudinal data from the MIMIC-III data set. This work incorporates multiple types of features like chart events, demographic features and pre-maid ICD9 embeddings and uses diversified ML approaches like RNN, LSTM, CNN and a couple of conventional supervised learning approaches like RF, SVM, NB and LR. The methodology can be divided in to three different parts; (1) The **data set construction**, where the authors construct based on the MIMIC-III data set a specific sub set that is the one the ML approaches use to make predictions. In this part the patients represented on the data set are categorized in to positive and negative cases, where the positive cases are the patients that would benefit from the prediction approach before a transfer or a discharge. Based on these categories and also based on other constraints like age, the data set is filtered and the authors are left with a sub set of 35,334 patients that represent the set for prediction. (2) The second part of the method is a **feature extraction**, where the features and time series window are introduced. For temporal information, the authors selected a 48h window from where the information can be extracted and data obtained after that window of the ICU stay is not included. To cope with any possible missingness on that window the authors also used a Last-Observation-Carried-Forward (LOCF) imputation method to deal with it. Within this 48h window the information extracted from the data set are of two categories; the chart events and demographic information and to develop the readmission prediction model to these two categories the authors added the ICD9 Embeddings that were pre-trained. These pre-trained embeddings provide a easy access to the ICD9 terminology with the terms already computed on the embedding were the is no room for enrichment or change in the content, making semantic enrichment impossible on these embeddings. (3) The final part is the **machine learning** (ML) where the authors construct based on the features extracted three different approaches; the baseline models, that include the conventional machine learning approaches RF, SVM, NB and LR, and are implemented with regularization penalties; the CNN to analyse longitudinal EHR data; and LSTM that is the best fitted approach to deal with predictions based on time series data.

Table 3.1: Pool of papers that respect the selection criteria for the introduction on the mining EHR related work.

Authors	Prediction task	MIMIC-III features	Embeddings	Ontology for Enrichment	ML Methods
(Anand et al., 2018)	Predicting mortality in diabetic ICU patients	Laboratory events, Demographic data, Descriptive statistics, Admission information.	Does not use Embeddings	Indirect use of ICD9 ontology, though the ICD9 codes.	Random forest models, Logistic regression.
(Beaulieu-Jones et al., 2018)	Predicting patient's survival within a 1-year period from admission.	Statistical data, Laboratory events, Procedural information, Streaming data measured.	Uses t-Stochastic Neighbor Embeddings that are extracted from a series of events.	No ontology directly used on the method.	Multi-layer perceptron, Deep neural network, Random forest, Logistic regression, Support vector machine, Long short term memory networks.
(Lin et al., 2019)	Predicting patents readmission within a 30 day period from discharge.	Statistical features, Demographic features, ICD-9 codes, Chart events.	Uses pre-maid ICD9 embedding, that are based on text corpus to detect terms.	Indirect use of the ICD9 ontology, trough the pre-maid ICD9 embeddings. No direct use of ontologies on the method.	Random forest, Logistic regression, Support vector machine, Naive Bayes, Long short term memory networks, Convolutional neural network.
(Javan et al., 2019)	Cardiac arrest prediction model for adult patients with sepsis.	Multivariate features, Time series, latent features.	Does not use Embeddings.	No ontology directly used on the method.	Support vector machine, Decision tree, Logistic regression, K-nearest neighbors, Gaussian NB, XG Boost, Random forest, Balanced bagging classifier.
(Xu et al., 2019)	Predict the risk of Acute Kidney Injury in Critical Care.	Demographic features, Medication information, Comorbidities, Chart events, Laboratory events.	Does not use Embeddings.	No ontology directly used on the method.	Support vector machine, Random Forest, Gradient Boost Classifier, Multi-Layer Perceptron, Long short term memory networks, Convolutional neural network.
(Lu et al., 2019)	30-day unplanned ICU readmission risk.	Discharge summaries	Graph-based text classification model , i.e., MedText.	No ontology directly used on the method.	CC-LSTM, Graph CNN.

Chapter 4

Methods

To accomplish the identified goals, this work follows a detailed methodology that is supported by the following research questions:

- RQ1: How can patient data be properly annotated and linked to ontologies?
- RQ2: How can these annotations be explored by KG embeddings RDF2Vec to improve prediction of readmission risk?
- RQ3: How is predictive value impacted when predictions are made at different moments through a patient's ICU stay and with varying information available?

The methodology was set up with different components and strategies that although linked, support the evaluation of different components and an elucidation of the research questions separately.

The methodology has 3 main components: (1) data collection and pre-processing, (2) semantic feature generation, (3) machine learning models implementation and evaluation. Each part plays a role that generates contributions to the overall methodology. The data collection gathers all the ICU information needed for the ML model and semantic enrichment from the MIMIC-III database. Then the semantic feature generation step performs semantic annotations of the data and generates vector representations (i.e., embeddings) that can be processed by the ML models. Finally, several ML algorithms were applied to the feature set and predictions were evaluated and compared with a state-of-the-art work(Lin et al., 2019). These phases and each one of their steps are represented on Fig 4.1 and detailed in the next sections.

It is relevant to note that this work builds upon Lin et al. (2019) but goes beyond this work by both considering additional relevant information from the MIMIC-III data set and by enriching this information with semantic representations of features based on ontology embeddings. While Lin et al. (2019) built a readmission prediction model based on three specific categories of features in the MIMIC-III data set, namely, chart events, ICD9 embeddings and demographic information for each patient, this work also includes the

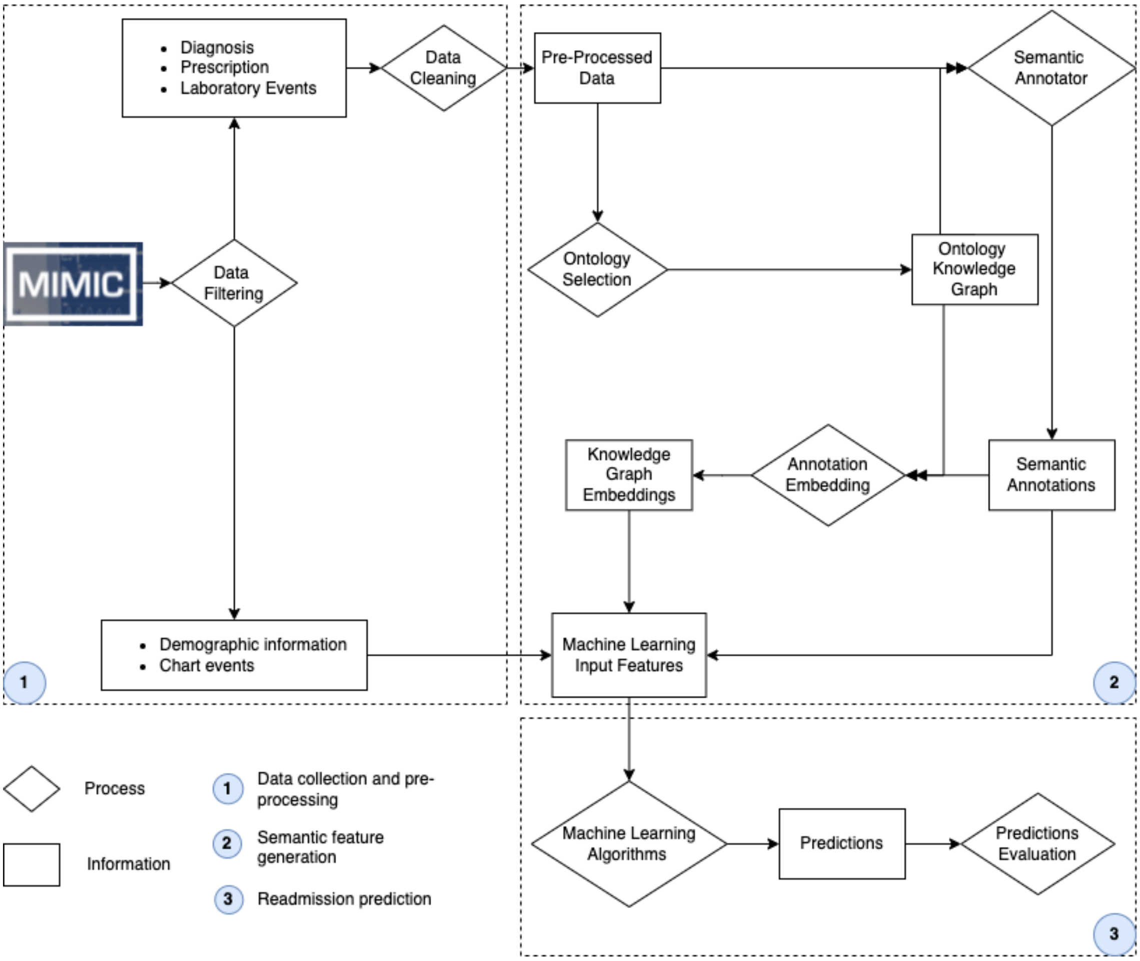


Figure 4.1: Schema representation of the methodology.

prescriptions information, initial and final diagnosis information, procedures information and laboratory events. This work also replaced the pre-trained ICD9 embeddings with RDF2vec embeddings (Ristoski et al., 2018) for all the considered features, because on the timeline of an ICU stay the ICD9 information can only be obtained at the end of the stay and thus prediction is delayed until the last possible moment. Moreover, this work supports predictions at different moments of the ICU stay as more data becomes available.

4.1 Data Collection and Pre-processing

This project deals with a large amount of information, regarding medical data on the format of EHR, as well as more comprehensive information about the medical terms being dealt with, collected with the use of ontologies that help categorize and describe the information based on the representations of the concepts that the ontologies and controlled vocabularies define.

The medical data used on this project is the EHR data from the MIMIC-III data set. This is a large, freely available database comprising de-identified health-related data associated with patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The MIMIC-III contains data associated with $53,423$ distinct hospital admissions for adult patients (aged 16 years or above) and includes information such as demographics, vital sign measurements made at the bedside (1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (including post-hospital discharge) (Johnson et al., 2016).

4.1.1 Data Acquisition and Filtering

The MIMIC-III data is available as a set of csv files. These files were processed to extract the necessary features to describe each patient and their ICU stay.

Patients' demographics are registered at admission and include information such as insurance type for billing proposes, language and religion for communication and procedures warnings, ethnicity, gender, age, etc. of all the information available, the ones selected were the age, gender and ethnicity as general demographic features and insurance type as Lin, Yu-Wei et al. (Lin et al., 2019), believe that this specific feature can influence discharge/transfer rate, knowing that different insurance companies may have different policies. The possible values for each feature can be found in Tab 4.1.

Chart events are the events occurring on a patient's chart such as notes, laboratory test, and fluid balance, spread through several tables like $Output_events$ table that contains all the measurements related to a patient's output during an ICU stay. Like Lin, Yu-Wei et al. (Lin et al., 2019), extracted 17 types of features from the chart events and used the features them self's as well as the normal median values in humans for machine learning interpretation. The mean values are adjusted to the American population represented on

Table 4.1: Demographic features used on the model, with the possible values each feature can have. Source: Lin, Yu-Wei et al.(Lin et al., 2019)

Features	Dimension	Options
Gender	2	Male/Female
Age	1	18-120
Insurance type	5	Government, Self, Medicare, Private, Medicaid
Race	6	Asian, Black, Hispanic, White, Other, No information
Total	14	

the MIMIC-III data set. The information extracted accounts for a total of 59 dimensions with 17-dim binary to overcome missing data and evaluate the existence of the chart event. **Prescriptions** are the medication related orders attributed to a patient and have a detailed description of the consumption with start and end date as well as dosage in value and unit. Adding to this information is also possible to find the type of drug prescribed as well as the representation of the drug in specific drug coding systems. Of the information available the method will target the drug name and National Drug Code (NDC) Tab 4.2, as these two information gives the possibility to map each code to a specific concept.

Table 4.2: Sample of Prescriptions table from the MIMIC-III data set. Source: https://mit-lcp.github.io/mimic-schema-spy/index.html.

Features	Comments	Example
Subject ID	Patient identification	6
Hadm ID	Hospital stay identification	107064
Start date	Start of the prescription	11.06.75 00:00
End date	End of the prescription	12.06.75 00:00
Drug name	Name of the drug	Warfarin
NDC code	National drug code	56017275
Rout	Rout of administration	PO (Oral administration)

The **Diagnosis** is captured at admission together with the demographic information and provides a preliminary, free text diagnosis for the patient being admitted to the unit, and do not use systematic ontologies. A final diagnosis is coded at the discharge and presented when billing the patient. The first diagnosis can be very helpful when researching the status of the patient and although it may be vague with proper semantic annotation, a first prediction on the stay can be made has soon as the patient is admitted to the unit instead of waiting on the report of the full stay.

Procedures contain the ICD9 procedures for patients, more specifically ICD9 procedures. These codes are generated for billing purposes at the end of each hospital stay and are recorded for every hospitalization on the MIMIC-III data set. For this information the target is to collect the ICD9 code corresponding to each medical procedure performed on the patient during a stay.

Lab events data contain information regarding laboratory-based measurement. These measurements are associated to an analysis on a specific fluid from a site of the patient's body (e.g., blood from an arterial line, urine from a catheter, etc.), and have a coded bar associated with the patient and timestamped to record the time of the fluid acquisition. The Lab events contains both in-hospital laboratory measurements and out of hospital laboratory measurements from clinics which the patient has visited Tab 4.3.

Table 4.3: Sample of Lab events from the MIMIC-III data set.

Columns	Description	Example
Subject ID	Patient identification	156541
Item ID	Charted item identification	50912
Lable	Label item identification	Creatine
Value	Value of the event	4.4
Unit	Unit of measurement	mg/dL
Flag	Indication of abnormality	abnormal

4.1.2 Data Cleaning and Pre-Processing

Similarly, to Lin, Yu-Wei et al. (Lin et al., 2019), the data set will be filtered by removing patients in MIMIC-III that are under the age of eighteen years old and are in the ICU, and this results in a total of 35,334 patients with 48,393 ICU stays. Then the processed patients are split into training (80%), validation (10%), and testing (10%) partitions to train the model and perform a five-fold cross-validation.

The patients left after the filtering are categorized according to their corresponding ICU stays records into positive or negative cases. Positive cases are the ones where the patients could benefit from a prediction of readmission before being transferred or discharged. Negative cases are those where the patient does not need ICU readmission including those who were transferred or discharged, did not return and are still alive within the next 30 days. According to the criteria for patient's selection, the following cases are considered to be positive patient stays (Lin et al., 2019):

- The patients that were transferred to low-level wards from ICU, but returned to ICU again.
- The patients that were transferred to low-level wards from ICU, and died later.
- The patients that were discharged, but returned to the ICU within the next 30 days.
- The patients that were discharged and died within the next 30 days.

After filtering the patients, a selective group of individuals are left. With them comes all the data that is traceable to their individuals stay, thus meaning that now the only information available is the information of the stay associated with this group of people. This data, corresponds to all the chart events and demographic information, the prescriptions information, initial and final diagnosis information, procedures information and laboratory events. Demographic information is not subject to this processing.

Although the patients are filtered, their information is still raw. Meaning that the classes are untreated and all the data not associated with a particular terminology is still in the original format, such as the initial diagnosis that is a free text variable recorded by the clinicians with possible misspellings, incompleteness or noise in the form of unreadable characters, that when transformed to an EHR format are not accounted or transformed in other characters. To advance in the methodology, first the MIMIC-III data needs pre-processing, in order to prepare the data for the machine learning models. The focus is on data cleaning, to treat the miss handled classes and raw free text data that has incomplete terms, noise and inconsistent data (Li, 2019). In this last category are included the initial and final diagnosis and chart events. To these the treatment is as follows;

- To the empty terms, inconsistent data, and noisy information, the strategy is to exclude and eliminate such information, as these cannot be deciphered correctly leading to meaningless data or even to incorrect data.
- For the incomplete terms, they are kept to further perform semantic similarity so that they can be matched with proper complete terms on an ontology and provide insightful information.

Fortunately the MIMIC-III data is mostly composed of information associated with a terminology, meaning that most of the information is mapped to a class and identified with a code that has a corresponding predetermined term preventing the misspellings error associated with free text variables not subjected to a specific terminology. Although these classes are not included in the first part of the data cleaning, they need to be treated due to possible computational error and miss handling, that leads to empty classes (no code), double codes (one code for two terms), wrong codes (codes in the wrong format) or incomplete labels. In this category are included the prescriptions, lab events and procedures where the treatment is as follows;

- The empty classes are excluded and eliminated as it is impossible to establish the correct class because, even if there is a label, the classes can be defined by dosage or sub-type and with no information is impossible to correctly match the class. For instance, C0488132 (Brach a-R BP dias) and C0488131 (Brach a-L BP dias), that are labeled in most cases on the MIMIC-III labeled as Brach a BP dias, is impossible to definitely attribute the right class because the crucial information is missing.
- The double codes are fairly common in the MIMIC-III data set, and are due to miss labeling because the labels are incomplete or are wrong. To these, the solution is to

keep the codes because the they map proper unique classes and also keep the labels to perform semantic similarity and try to find classes on specific ontologies, the differences are not problematic because with string similarity they will be trimmed to a similar string.

- Incomplete labels are kept to perform the same procedure as the incomplete term, semantic similarity so that they can be matched with proper complete terms on an ontology and provide insightful information.

- The wrong format for codes only happens on the classes of the procedures and final diagnoses, were the codes are presented on the MIMIC-III as strings with no characters. And to these the solution is to follow the guideline for ICD9-CM and edit the codes accordingly. The guidelines are the ICD9-CM ones because this is the ontology the MIMIC-III is mapping this concepts to.

4.1.3 Snap Shot split

An ICU stay is not a stagnant process solvable in one particular moment, contrariwise it has multiple stages, because a patient that enters the unit, starts ill and with serious conditions and evolves with the treatment, drug intake and care, leading ideally to a successful discharge and recovery. This means that as a patient progresses, new information is generated, as drugs are prescribed, tests are prescribed and done, or procedures are done, and so, in terms of data, an ICU stay is an incremental process of information addition through time. In a prediction scenario that tries to be valuable, a 30-day readmission prediction after the stay is for certain valid and useful, but when considering an ICU stay as an evolutionary process, it would be better to take in to account these increments of information. Thus changing what would be a prediction made after, to multiple predictions made during the stay.To make multiple predictions, appropriate moments need to be selected based on when new is information being added, fortunately clinicians usually collect and record information at particular times, helping the finding snap shots of time appropriate for the prediction.

To capture the essence of the stay four moments (snap shots) of predictions are selected, Fig 4.2, as these are the particular moments where more information becomes available. To note that these moments of predictions are based on the MIMIC-III data set and can be changed to fit other scenarios and data sets, changing the timings and information available.

- At **Admission**, where the information available is the ***Initial diagnoses***, *the* ***Demographic information*** of the patient and the ***Chart events***. In the timeline of an ICU this moment is when a patient is admitted to an ICU unit, his information is collected and a preliminary diagnosis is made.

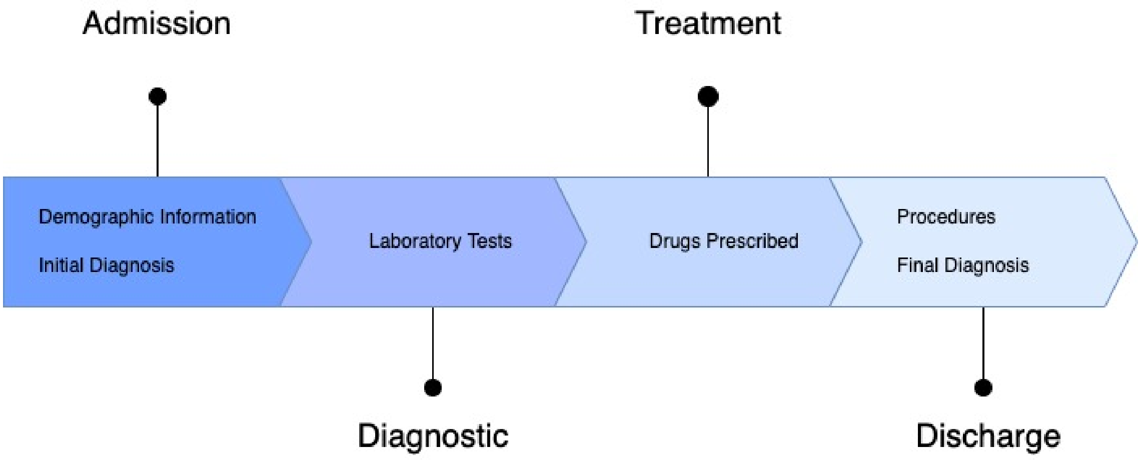


Figure 4.2: Prediction moments on an ICU stay, according to the MIMIC-III data set, as well as the new information added in each moment

- The second moment of prediction is the **Diagnostic**. This moment is on an ICU stay timeline, the moment where tests are ran to try and confirm or reconsider the diagnosis, thus meaning that in terms of the information now there is also the ***Laboratory tests***.
- The Third moment is the **Treatment**, where the patient is prescribed drugs to treat the disease. Here the information incremented is the ***Drug prescriptions***.
- The final moment is at **Discharge**, when the patient is released. In this moment the information incremented are the ***Final diagnosis*** and ***Procedures made***, although this increment does not match what would be an expected timeline, on the MIMIC-III data set this information are only revealed when billing the patient, thus meaning that only at discharge they are revealed.

4.2 Semantic Feature Generation

Following the data collection, cleaning and pre-processing semantic feature generation is performed based on the features collected. To do so, there are three main steps, Fig 4.1: (1) Ontology Selection, where ontologies that provide adequate coverage of the feature's domains are selected; (2) Semantic Annotation, where textual features are mapped to ontology classes that describe them; and (3) Annotation Embedding, where each feature's annotation is processed using a knowledge graph embedding approach that represents it in a numerical vector that reflects the meaning of the particular class within the ontology. Each patient's stay is then represented through an aggregation of the several embeddings vectors that describe it, as well as the original textual and numerical features.

4.2.1 Ontology Selection

With the data cleaned and processed, now available are the sets of terms for prescriptions, diagnosis, procedures and laboratory events. To these sets, an ontology to map the terms need selection, each term will then through an annotation process be mapped to a class on the ontology.

The BioPortal Recommender platform (Martínez-Romero et al., 2017) was used to support ontology selection. The BioPortal Recommender is a service created and managed by the National Center for Biomedical Ontology (NCBO), that receives a biomedical text corpus or a list of keywords, for instance a set of EHR terms and suggests ontologies appropriate for referencing the indicated terms (Martínez-Romero et al., 2017).

For this data, the inputs are the textual features, prepared as lists of terms or keywords, according to the BioPortal's recommendations. The goal is then to select the ontology that has the best coverage of the sets. An important benefit of the BioPortal Recommender is that it can be used on both complete and incomplete terms because it can handle minor misspelling errors due to the tokenization it does to perform string similarity between the terms and the labels of the classes, thus meaning that it can handle incomplete terms matching them with specific classes.

Another relevant benefit of the BioPortal's Recommender is the diversity in fields of knowledge, providing ontology recommendations of the various areas and consequently recommending more than just biomedical ontologies. In this implementation, and to prevent attribution of non relevant ontologies, a pre-selected group of ontologies of interest was selected; NCIT, SNOMEDCT, Medical Subject Headings Thesaurus (MeSH) and RxNORM were selected based on relevance attributed in a previous work (Bodenreider, 2008); Logical Observation Identifier Names and Codes (LOINC),The Drug Ontology (DRON) and International Classification of Diseases, Version 9 - Clinical Modification (ICD9CM) were selected due to their presence on the MIMIC-III data set; MedDRA and Experimental Factor Ontology (EFO) were selected as extra relevant biomedical ontologies.

After running the BioPortal recommender for all the information, the NCIT ontology is the best (Fig 4.3), with a good overall coverage, this because the NCIT is very inclusive and within the health care data most of the terms have a matching class on the ontology. Thus meaning that when exploring the Knowledge graphs more paths can be found with the NCIT than with the rest, Tab 5.1. The other ontologies have very dispersed coverage, performing well in some sets, but performing bad in other, NCIT although fluctuate is the one that present good coverage for the most amount of sets (see Annex B).

Although NCIT is the best ontology for this prediction scenarios, the MIMIC-III data set has mappings to specific ontologies, the LOINC used for the laboratory events such has blood or urine analyses, the ICD9CM, used for procedures performed or prescribed during a specific ICU stay as well as final diagnoses, and the NDC used for drug prescriptions.

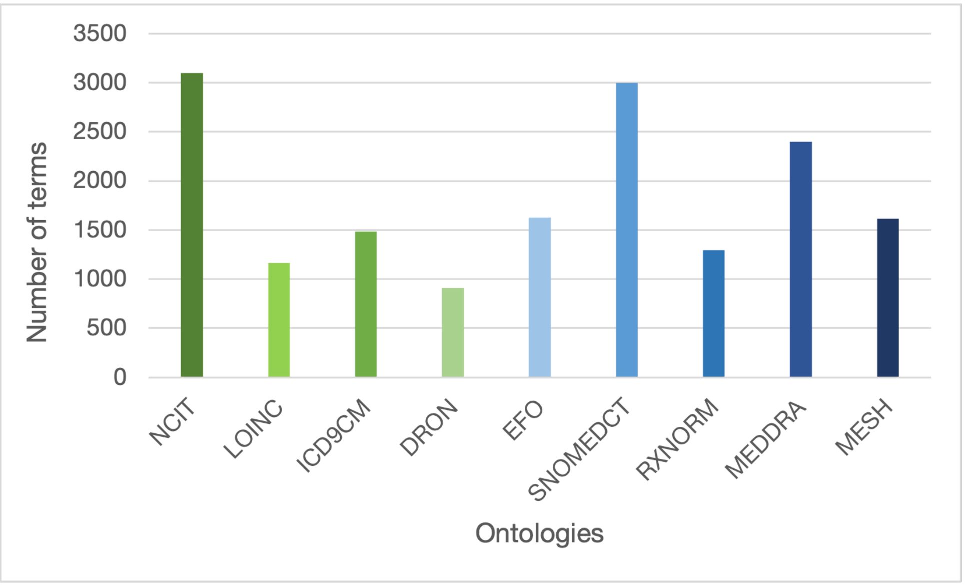


Figure 4.3: Amount of terms covered with each individual ontology, for all the information used on the method.

However, a concern regarding these ontologies is their accessibility because the NDC is not openly available. To address this issue the NDC ontology was replaced with the DRON, this because they concern the same information, have mappings between them and the DRON is not only public but also smaller.

With these conditions there is clearly an opportunity to extend this experiment and make two different prediction scenarios, one where the four different ontologies (NCIT, LOINC,ICD9M and DRON) are used and a second one where the annotations are made for all the data using just the NCIT.

4.2.2 Ontologies and Thesauri

The NCIT, that is the National Cancer Institute's reference terminology and ontology. NCIT is a thesaurus that provides the terminology concepts used on the national cancer institute semantic infrastructure, with responsiveness and with terminology based on scientific knowledge. This thesaurus covers the terminology associated with clinical care, public information and administrative activities. This thesaurus has many purposes, the annotation of data in the National Cancer Institute's repositories and search and retrieval operations in the repositories (Ceusters et al., 2005). One of the most advantageous parts of the NCIT and the main reason for a rich graph of knowledge is the interaction with other knowledge outside the cancer institute semantic. The NCIT is linked to other resources, both internal National Cancer Institute systems as caCore,caBio or MGED and also external systems such as the Gene Ontology or SMOMED-CT. Due to the fact that the NCIT is part of an open biomedical ontology and is considered open source, it the sense that it has a free licence download, the NCIT is a candidate to deliver vocabulary services in cancer-related biomedical informatics applications in the future (Ceusters et al., 2005), but as of now it is also recognized as the standard for coding and reference by a variety of partners. Due to it being a thesaurus, one can thus expect the utilization of the information in researches engaged in biomedical data annotations, but at the same time due to the fact the the ontological underpinnings are designed to open the possibility for more complex uses in liking heterogeneous resources and information retrieval, attentions is being drone to the NCIT, and in particular by the biomedical research community that find this particular ontology very useful (Ceusters et al., 2005).

The ICD9CM, was created by the Centers for Medicare and Medicaid Services (CMS) and by the United States Department of Health and Human Services as an extension of the ICD9, that was created and maintained by the World Heath Organization (WHO) to track morbidity and mortality statistics across the world. The CM version was created with the intent to be used in billing, and also in accordance whit the purpose ICD9 was created to, the ICD9CM codes have been used to record patient's diagnoses in clinical practice and health management for decades (Wei et al., 2017). ICD9CM uses between three and five digits to describe diseases and syndromes where the first three digits describe the general condition of a patient and therefore have been commonly used to represent disease categories, and the 2015 edition has 22,401 distinct codes that describe these diagnoses and syndromes. These codes are arranged hierarchically into nineteen large chapters, 160 sections, and 1,247 3-digit categories (Wei et al., 2017).

DRON, is a drug modular, extensible ontology of drug products that was built due to existing artifacts failures, that did not match the requirements and had systematic and ontological errors (Hogan et al., 2013), and to enable comparative effectiveness and health services researchers to query NDC's (Bona et al., 2019). DRON is an OWL 2.0 artifact, with manual curation at the upper layers and also with automatic curation based on RxNorm, a US specific terminology maintained by the U.S. National Library of Medicine (NLM) with all the drugs available in the market, and based on Chemical Entities of Biological Interest (ChEBI)) (Bona et al., 2019). The classes are populated with NDC's and other classes from RxNorm using only content created by the National Library of Medicine, these are in accordance with the February 2013 version of RxNorm and describe ingredients and ingredient relationships, semantic clinical drug forms, semantic clinical drugs, and semantic branded drugs (Hogan et al., 2013)

DRON thrives in quality of information, due to linkage with the NDC's, that are numeric codes issued by the US Food and Drug Administration (FDA) and published in a National Drug Code Directory that is updated daily (Bona et al., 2019), as well as the RxNorm maintained by the NLM. Thus meaning that DRON is a valuable publicly available solution to the private drug ontologies.

LOINC, is a set of identifiers, names, and codes for health measurements, observations, and documents that work as a universal standard for identifying laboratory observations (Adamusiak and Bodenreider, 2012). Thus meaning that LOINC is rich catalog of measurements, such as laboratory tests, clinical measures, standardized survey instruments and more measurements. Adding to the measures LOINC also saves the codes for collections of these items formatted as panels, forms and documents (LOINC, 2021).

LOINC has more than 15,000 users in 145 countries meaning that it can be considered the *lingua franca* of clinical observation exchange. It is recommended as part of the Meaningful Use and endorsed by American Clinical Laboratory Association and College of American Pathologists (Adamusiak and Bodenreider, 2012). LOINC has many benefits, perhaps the biggest is the ability to enable the exchange and aggregation of clinical results for care delivery, outcomes management, and research by providing a set of universal codes and structured names to unambiguously identify things you can measure or observe (LOINC, 2021).

4.2.3 Semantic Annotations

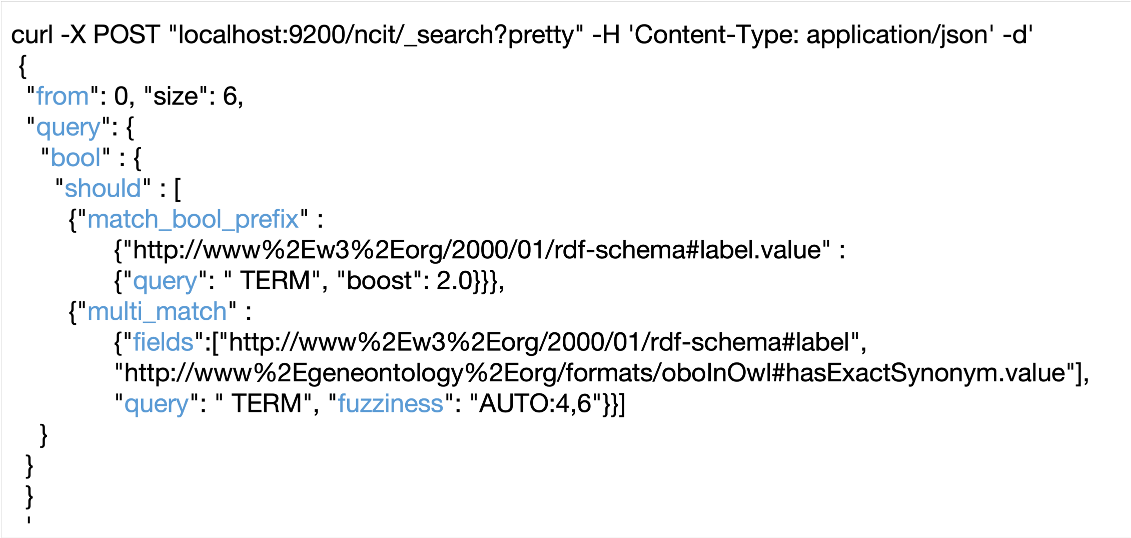
For the first ontology scenario, where four different ontologies are used, the semantic annotation procedure is simpler because the mappings already exist on the MIMIC-III data set for the laboratory events (LOINC), final diagnosis and procedures (ICD9CM) and drug prescriptions (NDC) classes. It is only necessary to map (NDC) to (DRON) and to find the mappings to NCIT that correspond to the initial diagnoses set.

For the second ontology scenario the procedure is more complex, because there are no mappings to the NCIT ontology in any of the sets, thus meaning that there is a necessity to find the NCIT mappings for all the sets, initial diagnoses, drug prescriptions, procedures, laboratory events and final diagnoses.

The mapping between (LOINC) and (DRON) was accomplished using the BioPortal Recommender, this time with the NDC classes as inputs and with the goal to search for the mappings to the DRON classes that will replace the original mappings on the MIMIC-III data-set.

A semantic annotation tool based on ElasticSearch (Kononenko et al., 2014) was developed to find the diagnosis terms mappings to the NCIT ontology and perform semantic annotation. Elastic Search is an open source full-text search engine designed to be distributive, scale-able, and near real-time capable, what make it a good tool to be used in this scenario were annotations need to be extracted on demand.

ElasticSearch works as a schema free and index dependent platform that stores documents as JSON objects on a database like format data, that supports indexing and searching through a server (Kononenko et al., 2014). The queries are structured to search standard JSON files and allow Elastic Search to establish specific parameters that help filter the results based on a similarity score, done comparing the query input with the targeted field, and to access the server the user uses a standard REST API.



```
curl -X POST "localhost:9200/ncit/_search?pretty" -H 'Content-Type: application/json' -d'
 {
  "from": 0, "size": 6,
  "query": {
    "bool" : {
      "should" : [
        {"match_bool_prefix" :
              {"http://www%2Ew3%2Eorg/2000/01/rdf-schema#label.value" :
              {"query": " TERM", "boost": 2.0}}},
        {"multi_match" :
              {"fields":["http://www%2Ew3%2Eorg/2000/01/rdf-schema#label",
              "http://www%2Egeneontology%2Eorg/formats/oboInOwl#hasExactSynonym.value"],
              "query": " TERM", "fuzziness": "AUTO:4,6"}}]
    }
  }
  }
  '
```

Figure 4.4: Elastic Search query format, to target label and synonyms of a term.

To extract the semantic annotations for the sets, the full NCIT ontology network is loaded to establish the nodes and clusters to search. Then it can be queried to find the specific nodes that match the terms, now represented as regular JSON entities with a *field* that can be searched. In this scenario, the *field* is the class label that is equivalent to the terms. These queries are designed to extract, based on the similarity score between the class label and term, the URI of the matched class, so that a gateway to the knowledge graph is possible (mapping) and consequently to the metadata.

The queries used for the sets of data, follow the same format the one as shown in Fig 4.4, replacing the ”query” with the intended term. On these queries the parameters are;

- **”From”**, that restricts the number of results, and in this case they are restricted to the six best scoring terms. More than six results, can lead to matches with low similarity.

- **”Bool”**, that allows the construction of Boolean query, that instead of a simple query allows the combinations of other queries. To this Bool parameter a **”Should”** parameter is added, and this is used to accept results with and without the exact query on the field, as opposed to the **”Must”** parameter that only allows results with the exact query on the field. This is done to extend the list of result and use the incomplete term that are not exact matches but with this parameter can be used.

- **”Match_Bool_Prefix”**, this parameter is intended to focus of multi word queries, and it works just as a normal match but creates a prefix query out of the last term in the query string. The scores on under this parameter are lower, and to make them comparable a **”Boost”** is applied to double the score and allow comparability.

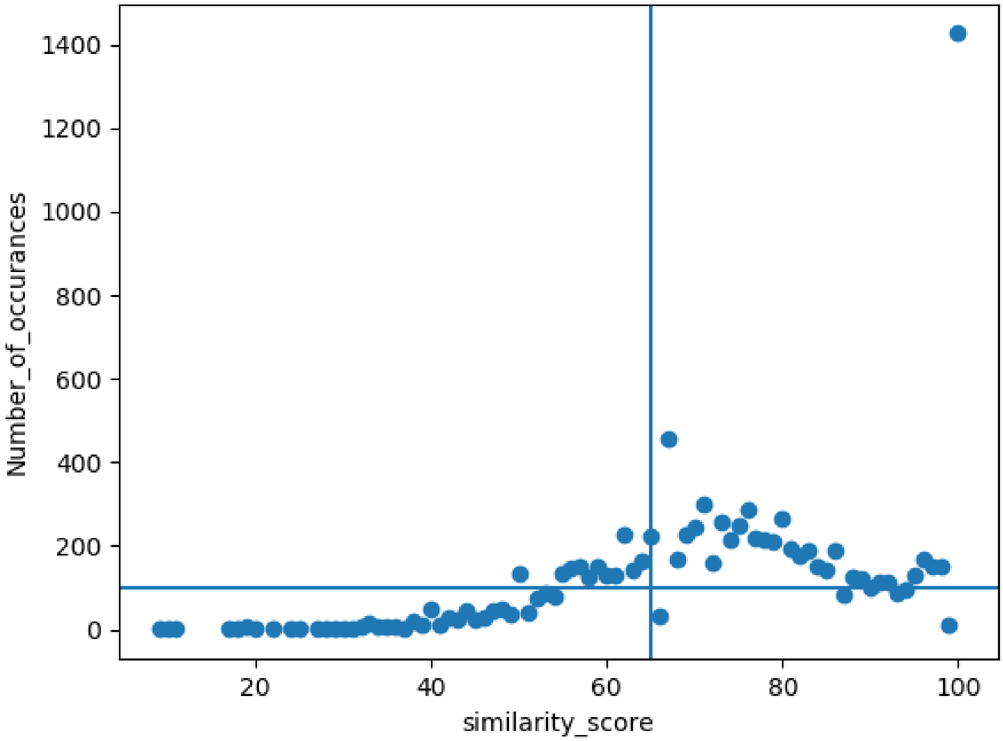


Figure 4.5: Distribution of the NCIT Elastic Search matched classes, with Levenshtein Distance to measure similarity.

- **"Multi_Match"**, this parameter allows the query to be run in more than one field. Although the target field is the label, because the terms on the sets do not always match the exact label a search on their synonyms can also be useful to capture terms that are intended to a particular label but are defined on the terms by the less common definition. To also support the search on the incomplete terms, the **"Fuzziness"** is also introduce to allow a predetermined number of misspellings and match the incomplete terms to a class.

After establishing a functional ElasticSearch server and prepare the query format, the search for the annotations is ready to start. The search is done with the individual terms from the initial diagnosis set, drug prescriptions set, procedures set, laboratory events set, and final diagnoses set as inputs to find classes. The outcome of the Elastic Search queering is, for each term a list of the six best scoring matched classes with the corresponding labels and URI.To these classes, and in order to select one as the matching class, the Levenshtein Distance between the label of each class and the input term is calculated, and the class with the smallest values is chosen as the class that matches the term. The Levenshtein distance between two strings is calculated as the minimum number of single-character edits required to change one word into the other, thus meaning that when comparing a list of strings to one target string, the one with the smallest Levenshtein Distance is the closest in terms of syntax. The ideal scenario is to have all the matches with similarity scores above 65%, ensuring that the terms are similar enough to have minor differences, and Fig 4.5 indicates that Elastic search annotation with Levenshtein Distance to measure similarity on the MIMIC-III data set has the vast majority of the results above that desired value, thus meaning that this is a successful procedure in this scenario.

This process is done for all the terms of the information sets, one set at the time, and the final result is all the terms mapped to the NCIT ontology as desired, and both ontology scenarios have now their information mapped and annotated.

4.2.4 Knowledge Graph Embeddings

Knowledge graph embeddings are learned for each annotated class of each ontology used, resulting in five sets of vector embeddings (four for ontology scenario one and one for ontology scenario two). The embedding technique used on this work is a graph-based approach, RDF2vec (Ristoski et al., 2018), which receives as input knowledge graphs as RDF graphs. RDF2Vec is the clear choice for the embedding technique because it is tailored to handle the specific semantics of RDF graphs and preserves the relationship between different entities, a particularly useful feature for semantic annotation Ristoski et al. (2018).The embedding vectors have a size of 300, with 500 walks per entity and a max depth of walk of 4, meaning that in a hierarchy format the walk can only move a maximum of 4 levels. If an ICU stay of a patient is annotated by more than one class in one ontology, then the vectors for each annotated class are summed. This aggregation approach follows the one used by Lin et al. (2019) for the ICD9 embeddings. Concatenation is used to combine vectors from different ontologies. This results in a vector describing an ICU stay of a patient with 1200 dimensions for scenario one and 300 dimensions for scenario two. All the combination approaches were tested (see Annex C) but concatenation achieved the best results.

Now has input information the models have the demographic features, chart event and RDF2Vec embeddings to be used based on the ontology scenario. These information are separated in different NumPy arrays, and need to be joined so that the final result presented to the ML models is a single array. To do so, the method uses a NumPy function called hstack that specifically concatenates NumPy arrays horizontally, or column wise. The functions concatenates the information, in a way that a patient can be described on a single horizontal line, interpret-able by a machine learning model. To concatenate the first array NumPy. Hstack uses the chart events, then the Embedding information and then the demographic information.

4.3 Readmission Prediction

4.3.1 Prediction Task

The prediction task is formulated to correctly predict if a patient is likely to be readmitted to an ICU unit within 30-days after release. Each instance corresponds to a patient and their ICU stay, and the class corresponds to whether the patient was readmitted or not within 30 days. Each instance is described by features extracted directly from the data and embeddings vectors. Predictions are made with varying sets of features according to the ICU timeline stay (see Section 4.1.2) to evaluate the evolution of prediction quality as more data becomes available during the stay.

Four classical machine learning methods are used: LR, RF, NB, and SVM. These are the same methods used by Lin et al. (2019) as baseline models. No hyperparameter optimization is applied, thus using just the default parameter of each algorithm, to ensure a more direct comparison to other works.

4.3.2 Prediction Evaluation

Following the base method proposed by Lin, Yu-Wei et al. (Lin et al., 2019), the evaluation of the predictive performance was done with a five-fold cross-validation and measurement of the ROC-AUC along with operating points corresponding to sensitivity (True Positive Rate (TPR)), specificity (False Positive Rate (FPR)) of the algorithm (Pepe et al., 2016).

To avoid over-fitting or bias a five-fold cross-validation was used. Performance of the model in each fold was measured and then results from all five folds were averaged to produce a single estimation for the model's performance.

ROC is a graph that represents the performance of a classification model at all classification thresholds, plotting the sensitivity TPR and specificity FPR and the area under the plotted curve is the AUC.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$ROC - AUC = \int_0^1 TPR(FPR^{-}1(x))dx$$

Additionally and to complement the ROC-AUC, the performance is also evaluated with the PR-AUC. Much like the ROC curve, the Precision/Recall Curve (PRC) evaluates the performance of algorithms and provide a graphical representation of a classifier's performance at all classification thresholds, plotting the precision against the recall (Storey et al., 2001). The AUC here, like on the ROC-AUC, summarizes the information in a single value, and the higher the score, the better a classifier performs for the given task with the added intention to study robustness and the trade-off between precision and recall.

$$Precision = PPV = \frac{TP}{TP + FP}$$

$$Recall = TPR = \frac{TP}{TP + FN}$$

$$PR - AUC = \int_0^1 PPV(TPR^{-}1(x))dx$$

Chapter 5

Results and Discussion

In this chapter of the work, the results obtained after the implementation are analysed and discussed, with the intent to see if valuable results were generated on each of the experiments. Adding to the results, this chapter, also present the reasons that most likely lead to the improvement or retrogression of the performance values. This chapter also aims to explain each of the experiments analysed, this because different scenarios are done for different reasons, these may be to see if a change in data or method leads to improvements or simply see if a implementation change that seems obvious has a good or bad impact on the performance.

5.1 Data Cleaning Outcomes

Before the cleaning of the terms, 14916 different one were found on the data set that were distributed as follows; 3567 unique diagnosis, 574 different laboratory tests, 2782 different drugs prescribed and 7993 different final diagnosis and procedures. Because these last to information's are on the data set analysed and used together, their proportions are undisclosed. After the treatment there are 12737 different terms that can be annotated, representing a total of 85.4% of the original terms, thus meaning that the data cleaning erased a total of 14.6% of terms that were, according to the restrictions, unusable. The proportions on the terms that could be annotated are as follows; 2709 diagnosis, 568 laboratory testes, 2782 drugs prescribed and 6678 final diagnosis and procedures. These proportions were expected because the diagnosis are free text variable made by the physicians and because of that are more prone to contain syntactic errors that ultimately lead to unusable terms.

5.2 Experimental Design

The experimental steps established to answer the research questions outlined in the previous chapter are as follow:

- **Reproduction Results**; Due to computational differences between the hardware and software used on this dissertation and used on the Lin et al. work, a direct comparison can not be made, and to solve this a reproduction with the computational power available is done so set a fair and reliable baseline for comparison.

- **RDF2vec Embeddings and Diagnosis Annotations**; With a fair baseline for comparison the next experiment is set to introduce the first semantically enriched features to the machine learning and see the impact these have on the performance values. In this first experiment and because the NCIT ontology is the best for this data set, the only information used is the diagnoses that is already mapped to the NCIT. To do so some changes need to happen, namely, the replacement of the pre-maid ICD9 embeddings with the new NCIT RDF2Vec embeddings and the replacement of the ICD9 terms with the diagnosis information to the input features.

- **NCIT Annotations for all the information**; After testing the addition of a NCIT semantically enriched feature, the next logical step is to test the incorporation of all the features on the machine learning. As established NCIT is the best ontology and so all the features will be mapped to the NCIT ontology. This increment is information will be matched with an increment on the information used to make the RDF2Vec embeddings, and so on this model all the features will be added and the embeddings will be constructed with all the enriched features.

- **NCIT replacement with MIMIC-III proposed Ontologies**; The NCIT is clearly the best fitted ontology to annotate this data set, and after experimenting with the addition of all the features mapped to the NCIT ontology, is time to see how the NCIT annotations compare with the annotation to a set of different ontologies. This experimentation exists because on the MIMIC-III, each type of feature is annotated to a specific ontology that according to the ontology description is the best to describe the feature. This experiment will reveal if a multi ontology approach results in better performance than a single ontology approach.

- **ICU stay simulation**; This experimentation aims to simulate a real world scenario where the information flows and changes as time advances. An ICU stay has specific time were new information is known, thus meaning that in terms of data an ICU stay is an incremental process where at the beginning few information is available, and at the end a much more substantial amount of information is available. Because this dissertation deals with a specific data set, the timings to make predictions are adjusted to the data set. This experimentation is done because just as Lin et al. (Lin et al., 2019) work, this dissertation as the potential to generate a substantial impact on the decision making of physicians on ICU, and by making predictions that come as close as possible to the real world case, more reliable this impact is.

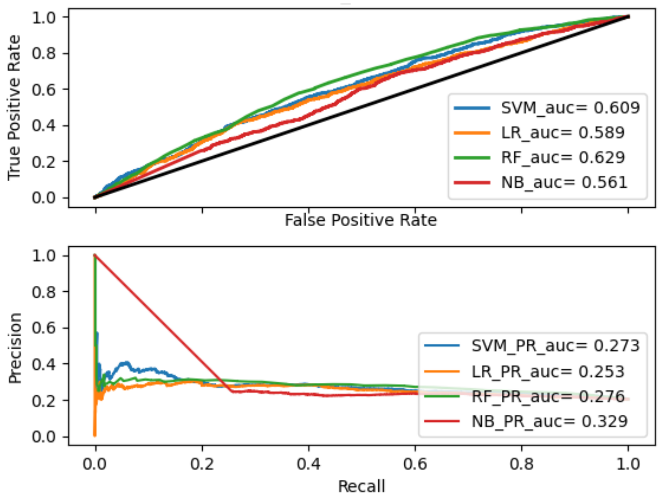


Figure 5.1: One fold representative ROC and PR values for the reproduction of the (Lin et al., 2019) method.

- **Impact analysis**; The final experimentation done on the procedures, is done to test the impact the features have on the results and to see if the RDF2Vec embeddings, that are the main focus of this dissertation, together with the liked data on the form of semantic annotations are the real reason the results are impacted.

The following sections describe in greater detail each of the experimentation's done, and analyse the results obtained with a discussion on the reasons for the results as well as a small discussion on what these results mean to the work. Section A tables report on the average ROC-AUC and PR-ROC, whereas figures report on results of a representative fold.

5.3 Reproduction Results

To evaluate the performance of the proposed methodology a baseline comparison based on (Lin et al., 2019) was established based on the sourced code made available. This allows for a more transparent evaluation, rather than just using the reported values since a reproduction is done under the same computational conditions.

The reproduction results on Fig 5.1 and Tab A.1 are somewhat lower from the ones reported. The authors were contacted about this, but no further support was given. It is possible this is due to details in the data set processing that were not reported by the authors.

The mean results obtained were ROC-AUC values of 0.5916 for SVM, 0.5714 for LR, 0.6176 for RF and 0.5578 for NB, meaning that all four prediction models are relatively poor in terms of performance which is also clearly visible in the plot of ROC curves presented in Fig 5.1. All the curves are close to a 45º angle meaning a high trade of between TPR and FPR, and consequently a very low predictive value. These low values also indicate that the models can not distinguish between what are readmissions and no readmission.

The PR trade off was also analysed with the PRC and with the PR-AUC's. The ideal values should be for AUC, greater than 0.5 meaning that there is a good enough trade of between precision and recall, and the model is robust and with a balanced data set. In this case we see mean values of 0.2576 for SVM, 0.2378 for LR, 0.2634 for RF and 0.3196 for NB, telling us that to achieve high precision we must sacrifice the recall so these models are not very robust.

5.4 RDF2vec Embeddings and Diagnosis Annotations

The study of the information available in the stay with the NCBO recommender showed that the Ontology with the best coverage is the NCIT, and with semantic similarity it was possible to find the annotations that better fit the terms. The terms that are originally mapped to the NCIT ontology are only the initial diagnoses and to do the first introduction of annotated information to the machine learning models this is the only information being added. This happens because on the original method the ICD9 information maps the final diagnosis, and to keep the prediction in the same domain of information the initial diagnosis is the best. Another important factor on the decision to replace the ICD9 information with the initial diagnoses information on the first introduction is that the ICD9 terms used on the base line methodology for both the model and embeddings, are only obtained at the end of the stay, meaning that predictions are restricted to the 30 day period after discharge and replacing the terms with the NCIT diagnosis information, that is information available from the very first moment of the ICU stay allows for an extension in this period that now starts form the admission until the 30 day period after discharge. To do this change, the ICD9 annotations are replaced with NCIT annotations and the ICD9 embeddings are also replaced with RDF2Vec embeddings that use the NCIT terms. The rest of the information containing the Demographic features and Chart events is kept.

Looking at the results on Fig 5.2 and Tab A.2 there is an impact in performance with an overall increase in mean ROC-AUC, now all the values are over 0.6 with an angle far from the 45º angle that is the minimum for an acceptable model, and in particular RF performed the best with a mean AUC of 0.6612, showing that the changes in the annotations and in the embeddings impacted positively the accuracy of the predictive models with a very acceptable trade of between TPR and FPR, and bring a high predictive value. This means also that there is valuable information in the initial diagnosis, which when semantically enriched is able to outperform the baseline that has access to the final diagnosis, which

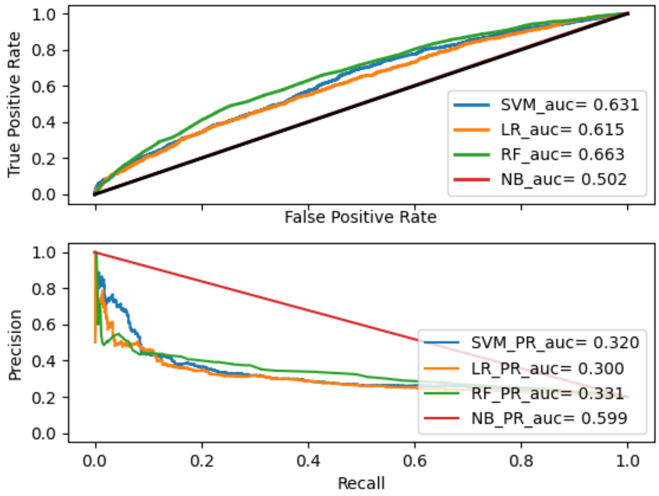


Figure 5.2: One fold representative ROC and PR values for the method replacing just the embeddings with RDF2vec embeddings and the ICD9 information with NCIT information.

one could hypothesize has better predictive value for the readmission risk.

On the other end when analyzing the PRC, although there is an improvement on the AUC mean values, that are now all above 0.3, only the NB has a mean value greater than 0.500, that is the targeted minimum value to consider a trade-off between precision and recall good enough and consider a model robust. In this situation although very close to that target the outcome still is that to achieve a high sensitivity there must be a sacrifice on the precision values Fig 5.2, because as soon as the recall gets higher than 0.01 the precision starts dropping. Looking at the results is safe to assume that the change in methodology with the replacement of pre-made ICD9 Embedding with RDF2vec NCIT Embedding with semantic enrichment is an improvement in all aspect to the base methodology.

5.5 NCIT Annotations for all the information

As established before NCIT has the best coverage out of all the ontologies for all the annotations on the MIMIC-III data set and as seen on the last section, using just the initial NCIT diagnosis improves the results. With this analysis made and with the results obtained, the next logical step is to see if by extending the NCIT to all the information available during the stay there is an impact on the prediction results. This will mean that the initial diagnoses, prescriptions, procedures, final diagnoses, and laboratory events information sets are all mapped to NCIT with the corresponding annotations and the RDF2Vec Embedding are extended with all this enriched features.

After implementing the models, the results on Tab A.5 clearly show that, using all the

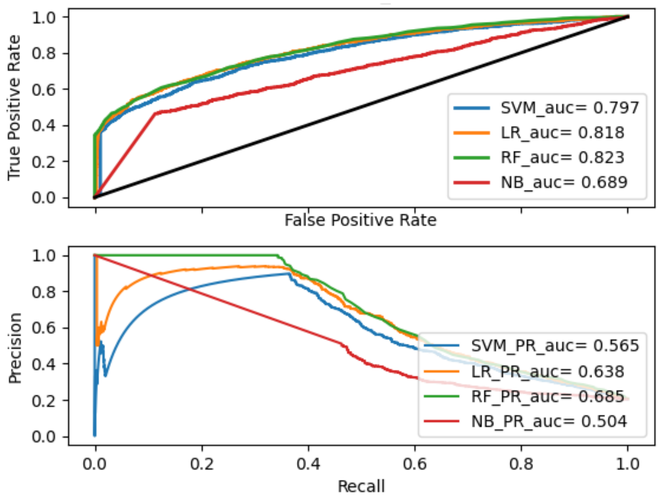


Figure 5.3: One fold representative ROC and PR values for the method mapping all the information gathered during an ICU stay, to the NCIT Ontology.

information available on the stay is beneficial, leading to higher overall mean ROC-AUC values with RF as the best performing value with a mean AUC of 0.826 and all models achieving values above 0.8, with exception of the NB model with a 0.6974 AUC value that although low, maintains the tendency of being the worst performing model.

The ROC values improved, but the biggest impact of using all the features was felt on the PRC where there was a growth to significantly better performance values. Now all the models are over the 0.5 desired limit and RF is also in this aspect the best performing model with a mean 0.685 AUC value, and if an analyses of Fig 5.3 is done, is possible to see that all PR-AUC curves, except for NB, show a growth of precision simultaneous to the recall growth until a very reasonable value of 0.4 where the decrease in performance begins but with a low accentuation thus showing a low trade-off between precision and recall throughout all the model. These values mean that the models are both highly valuable in terms of prediction value and accuracy, but also that these are very robust with no need to compromise any aspect of the model, ultimately meaning that the use of all the features is encouraged and beneficial.

5.6 NCIT replacement with MIMIC-III proposed Ontologies

Although NCIT is the best ontology overall in terms of coverage, the MIMIC-III data set proposes that individual information sets have individual specific ontologies to map the data so that the terms are within a specific vocabulary, laboratory tests mapped to the LOINC ontology, drug prescriptions mapped to NDC but once it is a privet ontology the

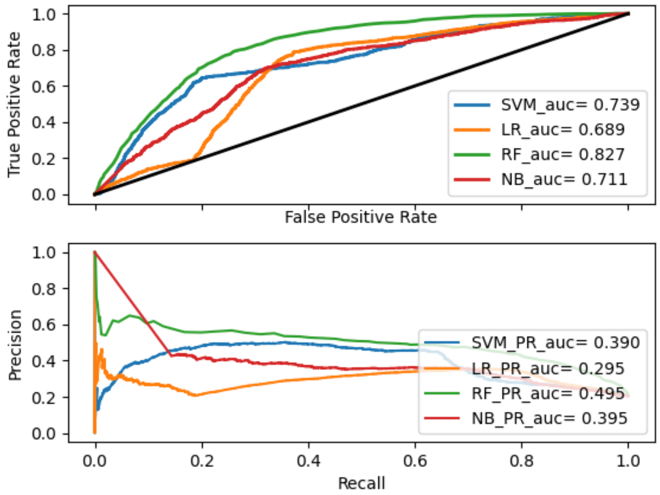


Figure 5.4: One fold representative ROC and PR values for the method mapping each set of the information gathered during an ICU stay, to the matching ontology proposed on the MIMIC-III data set.

terms are mapped to the best public ontology DRON, and procedures and final diagnosis mapped to the ICD9CM ontology. These are the proposed ontologies for the data, and adding to these, the NCIT ontology that maps the initial diagnosis, that had no specific ontology attributed on the MIMIC-III data set.

In this scenario each ontology requires specific conditions thus requiring four different RDF2vec embeddings each specific to the annotations and ontology appropriate, changing the scenario from a one embedding scenario to a four-embedding scenario. This change requires a change in the machine learning development because, at this time of the development the models were only able to deal with one embedding, to work around that challenge and to avoid large changes the embeddings will be joined prior to the implementation of the models, so that we have once again only one embedding is used. After analyzing all the options to join embeddings, it was clear that no matter the way these were joined, the final result remained unchanged and so the choice was to do a concatenation of the elements, thus joined the embeddings via concatenation, as this is the option that is more computational friendly. Adding to the change in embeddings the annotation also changed because now there are four sets of annotations instead of one, and to these the solution is easier as the only change needed is to load the annotations one set at the time without a need to join them because the models can easily be adjusted to deal with multiple annotation sets.

The results on Fig 5.4 and Tab A.6, show that the performance is worse in this scenario, this is most likely due to the quality of the embeddings, the NCIT produces good results not only because it is the ontology with the best coverage leading to more paths

Table 5.1: Difference in paths found when doing RDF2vec Embeddings with just NCIT or with the multiple ontologies proposed on the MIMIC-III

	Initial Diagnosis	Laboratory Tests	Drug Prescription	Procedures and Final Diagnosis	Total Paths
MIMIC-III Ontologies	5262287	78819	2456866	19322554	27120526
NCIT Ontology	5262287	7100581	6690821	8340000	27393689

found on the ontology Tab 5.1, that enriches the embedding, but another factor and most likely the most impactful on the quality of the embedding generated, NCIT has a better KG than the rest of the ontologies, the KG for NCIT is very rich and informative when compared to the ICD9CM, DRON or LOINC, this meaning the most natural outcome is that the annotations when ran with the NCIT find more paths lead to fitter embeddings for the prediction scenario that uses this sets of data.

In term of actual values, all the models except RF decrease their ROC-AUC values and consequently both their accuracy and prediction values decreased by adding different ontologies to the data. The trade-off between TPR and FPR although still very good with ROC curves all above the 45° angle desired, is worse when compared with the trade-off obtained when using just the NCIT in the same conditions. The exception of RF, that has the best recorded ROC-AUC mean value so far out of all the experiments, 0.825, can mean that RF in these conditions is the optimal solution to the prediction scenario under analyses, but as all the other models indicate otherwise and the difference between RF with NCIT and RF with the four ontologies is very small, the saves and most obvious withdraw is that using just NCIT is still the best solution.

The PRC results back the conclusions taken when looking at the ROC curves, and extend them in terms of the quality and robustness of the model, because with the four ontologies, the models once again goo under the desired values of PR-AUC, 0.5, and the models cannot be considered good or robust with the exception once again of RF that has the best recorded PR-AUC mean value, 0.4694, and as highlighted before, although this particular value is very good the rest are very poor and the method with four ontologies has no robustness, thus meaning that is better to use just the NCIT.

5.7 ICU stay simulation

An ICU stay has multiple stages and the information is not all available from the get-go, the information is added has time goes through and the patient is prescribed drugs, prescribed tests, or does procedures, and so, in terms of data an ICU stay is an incremental addition of information through time. In a prediction scenario that tries to be valuable a 30-day readmission prediction after the stay is for certain valid and useful, but when considering an ICU stay as an all, it would be better to take in to account these increments of information and change what would be a prediction made after discharge and within a 30 day period, to multiple predictions made during the stay that can also be extended to the 30 day period after discharge. This change can impact the targeted beneficiaries of this predictions, because what was useful to prepare for a possible readmission, can now be used to adjust treatments, adjust exams or simply go back to the original purpose and prepare for a readmission, because with predictions done during the stay, at all the stages the likelihood a patient has, to be readmitted, is known as well as how it changed though a full stay, thus meaning that there is no need to release a patient or wait for the final stage of the stay to be imapactful, as a matter of fact if the method is proven viable as soon as an admission is made, precautions can be taken to try to lower the likelihood of readmission.

With the goal then to make good predictions that can be used by medical practitioner, to follow the risk of readmission of a patient during an ICU stay, four moment of prediction are proposed Fig 4.2, at admission, diagnostic, treatment and discharge. If this predictions are proven valid and robust at every stage, it can be said that a tool to help medical petitioners prevent readmission form the moment of admission is not only doable but also that this methodology takes a good step on the right directions.

After all the previous results the clear choice of scenarios is to use all the information sets mapped to the NCIT ontology as well as the RDF2vec embedding with the NCIT KG. To capture the essence of the stay for the four moments of predictions, the annotations will incrementally be added, adjusting to the information available at that moment. To note that these moments of predictions are based on the MIMIC-III data set and can be changed to fit other scenarios and data sets, changing the timings and information available. At admission the only information available is the initial diagnoses, on the second moment, diagnostic, we have initial diagnoses and laboratory tests, and this information is loaded together and used together to make the embeddings, at the third moment, treatment, we have initial diagnoses, laboratory test and drug prescriptions and at discharge we have initial diagnoses, laboratory test, drug prescriptions, procedures and final diagnoses.

The results for the admission (Tab A.2) are the same as the first prediction made with the NCIT ontology with just the initial diagnosis and so just like that prediction the results are good and predictions are valid but these are not very robust with PR-AUC values near the desired 0.5 value, witch is ideal but with this values is safe to say that from the beginning valid predictions can be made, but these risk predictions should not be used by them self's to assume the risk and rather wait for a second moment of prediction to ensure the assessment, but never the less is possible to start contributing to the risk assessment evolution with the first value.

The second moment of prediction, diagnostic (Tab A.4), has very appealing results because the ROC mean values stay consistent, with all the AUC mean values higher than 0.7 and majority of them around 0.8. These values are very good in term of predictions hinting that from the second moment accurate predictions can be done, and with such results a high prediction value is added. In addition to the ROC mean values, the PR mean results are also very good, with all the values higher than the 0.5 desired value, thus

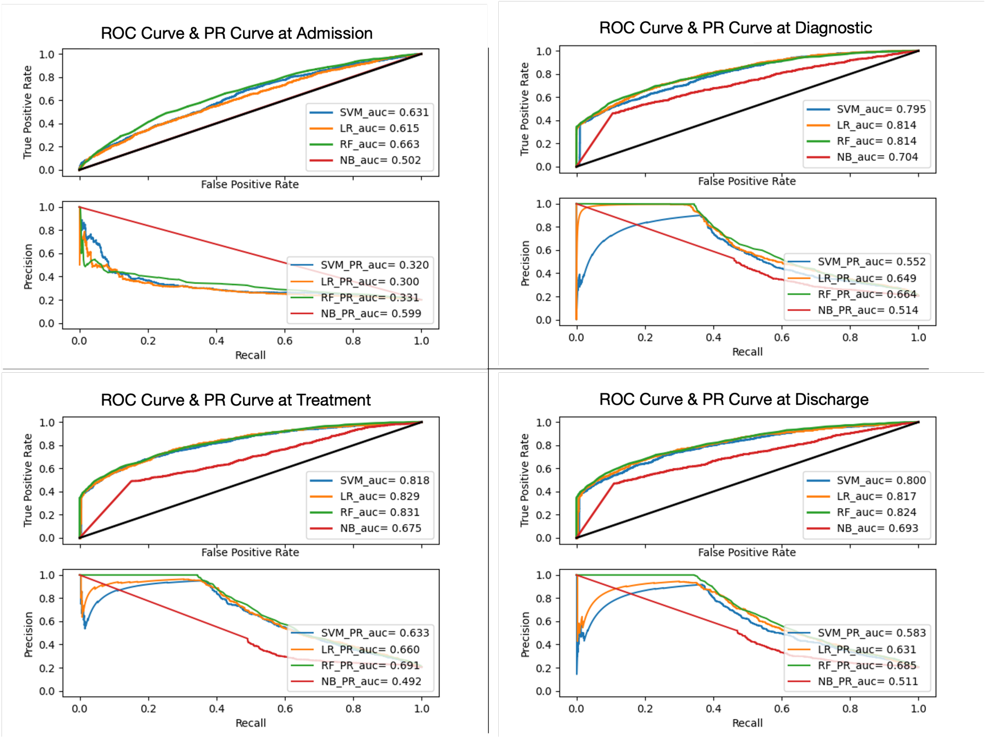


Figure 5.5: One fold representative ROC and PR values for the different stages of an ICU, with the data mapped to the NCIT ontology.

meaning that the predictions are not only accurate, but also robust. These results are of extreme importance because they mean that in this real life scenario, predictions can be done without a compromise in nether precision or recall, already from the second moment of prediction.

The third moment, treatment (Tab A.3), once again has very good results where again the ROC mean values remain high with all the values above an excellent 0.8 now, except for the NB model that has established will most likely always perform the worse, but nevertheless these results are very positive and on this third model is safe to assume that yet again is possible to produce very valuable and accurate predictions. This is a very good sign that the proposal is valid because this is the third moment out of the four were the predictions are valuable. The PR mean values are in accordance with the ROC mean values, and the distributions is around the same but now with all the values higher than a great 0.63, with the expected exception of on NB that was kept around 0.5. With these results is safe to say that the third moment can also produce good predictions and the ICU timeline predictions start to take shape and the possibility of making predictions though the ICU stay become now valid and the risk can most definitely by tracked through the stay. The final moment of prediction (Tab A.5), as the rest, has very good results, keeping all the ROC-AUC mean values above 0.8 except for NB, and PR mean values all above 0.5.

These results are in favor that the method proposes can be transported to a real ICU

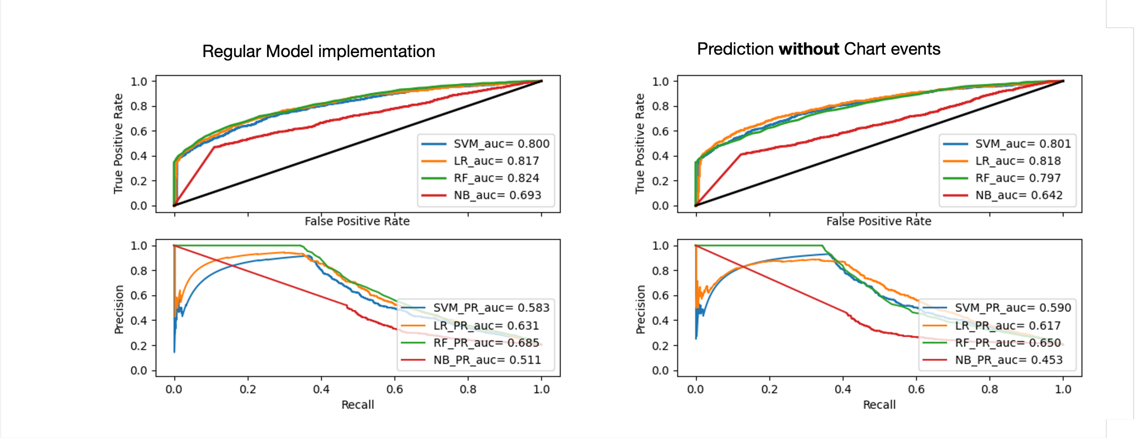


Figure 5.6: Comparison of the prediction values at discharge on a normal implementation and without chart events

stay scenario, with a continuous assessment of the risk of readmission so that medical practitioners can take advantage of the predictions and use them to benefit the well being of the patients by making decisions to help lower the risk from the moment the patients step in the ICU.

5.8 KG embeddings contribution analysis

By replacing the ICD-9 based embeddings with the KG embeddings, it is still possible that the performance gains observed are not necessarily due to the KG embeddings, but rather to the chart events features.

To test the true impact these two features have on the results, two different experiments were done: first, remove the RDF2Vec embeddings from the method to see the impact the absence of the RDF2Vec has on the results, and second, remove the chart events from the methods and keep the embeddings.

The experiment without the chart events uses as input the demographic features, and the semantic annotations and RDF2Vec embeddings at discharge, and compared the performance with the performance of a regular implementation also at discharge, that also includes the chart events. The results show that, although promising in theory, chart events proved are not contributing to the prediction power because the performance values stayed for the most part unchanged (Fig 5.6) when compared with a regular implementation. It is however possible that with other ML approaches, chart events can be better explored.

The second experiment compared the prediction at the end of the ICU stay with and without the RDF2Vec embeddings (Fig 5.7). It is clear that the RDF2Vec embeddings are the biggest contributors to the method performance, since the prediction performance without the embeddings drops drastically to a value close to the initial reproduction results (section 5.3) that also do not feature the RDF2Vec embeddings.

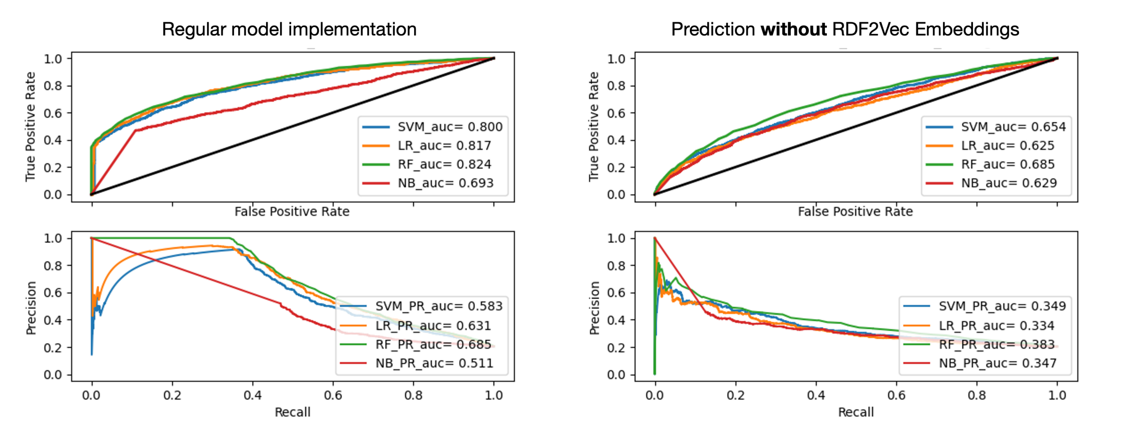


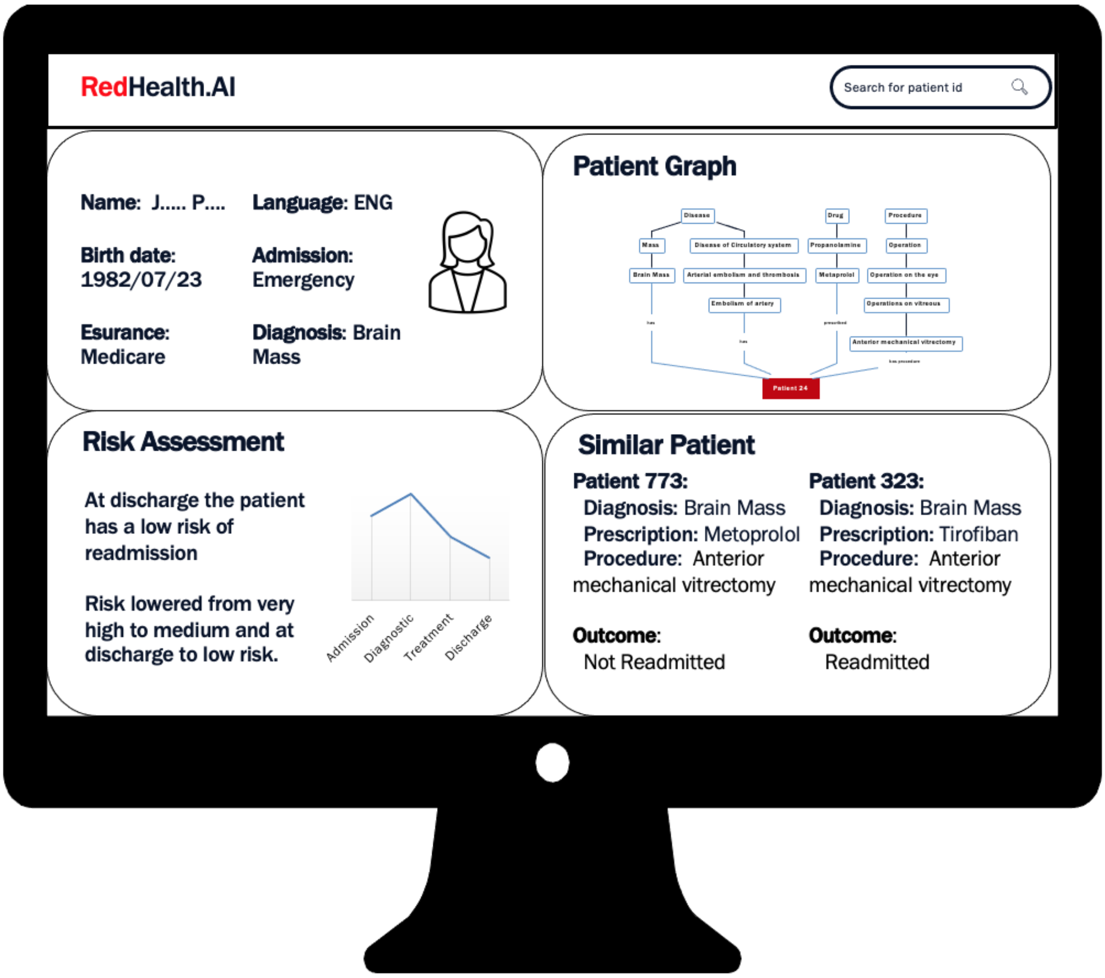
Figure 5.7: Comparison of the discharge predictions with RDF2Vec embeddings and without RDF2Vec embeddings

5.9 Entrepreneurship Projection

Given the positive results obtained and my personal interest, an entrepreneurship project was created to explore the opportunity of translating the scientific results of this dissertation into a technological product. This project included a number of tasks: (1) creation of a product idea; (2) development of a business plan in collaboration with TecLabs; (3) preparation of a slide deck for public presentation; (4) participation in international entrepreneurial contest'*H-INNOVA-Health INNOVAtion Award*'. More information about the award available on the Hinnova Hub Website.

RedHealth.AI is a product that aims to help clinicians decide whether a patient is ready to be discharged from the ICU in order to prevent further complications and readmissions. RedHealth.AI presents to the clinicians, in an understandable manner, key information to help the decision making and these include; the evolution of the risk of readmission, similar patients, detection of high risk patients, the patient graph with the stay information and relations and others. A preliminary prototype of the interface can be seen in Fig 5.8.

RedHealth.AI was selected as a finalist for the *H-INNOVA-Health INNOVAtion Award* out of more than 170 candidates from around the world, and placed in the Top 5.



RedHealth.AI

Name: J..... P....

Language: ENG

Birth date: 1982/07/23

Admission: Emergency

Esurance: Medicare

Diagnosis: Brain Mass

Patient Graph

Risk Assessment

At discharge the patient has a low risk of readmission

Risk lowered from very high to medium and at discharge to low risk.

Similar Patient

Patient 773:
Diagnosis: Brain Mass
Prescription: Metoprolol
Procedure: Anterior mechanical vitrectomy

Outcome: Not Readmitted

Patient 323:
Diagnosis: Brain Mass
Prescription: Tirofiban
Procedure: Anterior mechanical vitrectomy

Outcome: Readmitted

Figure 5.8: Preliminary interface for RedHealth.AI.

Chapter 6

Conclusions, Future Work and Limitations

ICU readmissions is a complex problem in healthcare, impacting both the patient lives and health and healthcare institutions management and finances. The growing adoption of EHR and the recent developments in ML applied to clinical data present an opportunity to address this issue by generating accurate predictions of readmission risk that help reduce the number of readmissions, improving health outcomes and minimizing institutional burden.

Despite these growing efforts, machine learning approaches still explore EHR data directly without taking into account its meaning or context. Medical knowledge is not accessible to these methods, who work blindly over the data, without considering the meaning and relationships the data objects. Ontologies and knowledge graphs can help bridge this gap between data and scientific context, since they are computational artefacts that represent the entities in a domain and how the relate to each other in a formalized fashion. This work investigated how enriching EHR data with ontology-based semantic annotations and applying machine learning techniques that explore them can impact the prediction of 30-day ICU readmission risk.

This chapter summarizes the main conclusions drawn from this work, discusses its main limitations and opens up new opportunities for future work.

6.1 Conclusions

The methodology and experiments conducted by this work enabled an elucidation of the proposed research questions.

For the fist question **RQ1: How can patient data be properly annotated and linked to ontologies?**, based on the results obtained it can be concluded that making annotations with just the NCIT ontology is the best solution, and in particular using all the information (section 5.5) provides better results than using just the diagnosis information that was the only that according to the MIMIC-III data set could be mapped to NCIT (sections 5.4). To get to the annotations, the terms regarding all the information of the stay must be mapped to the NCIT ontology, with the NCBO annotator, so that ontology classes match the terms and annotations can be extracted. Although these result may come across as unexpected based simply on the fit the ontologies have, this because the alternative solution to just NCIT approach proposes that based on the type of information each term provides, it should be mapped to the specific ontology describing that information, the approach with just NCIT outperformed the alternative solution (section 5.6), not only on accuracy and ROC AUC but also on the PR performance measures, ensuring that this method is not only more accurate and valuable but also most robust and reproducible.

As for the second question **RQ2: How can these annotations be explored by KG embeddings RDF2Vec to improve prediction of readmission risk?**, the results clearly demonstrate that when compared with the pre-made ICD9 embeddings (section 5.3), RDF2vec with either set of annotations (sections 5.4 and 5.5), outperforms the ICD9 embedding across the different experiments. The integration of RDF2vec embedding with EHR annotations, as a way to provide semantic enrichment, has a substantial impact on the prediction results, helping improve the performance no matter the annotations or ontology used to make the embeddings (sections 5.5 and 5.6. The best configuration of ontologies results in a ROC-AUC of 0.82, improving 0.2 above the baseline.

With the previous results and conclusions, although clear that the procedure that integrated RDF2Vec embeddings lead to improvements, no clear prof was presented to justify that the responsibility of the improvements was on the Embeddings. So the question was then, are the RDF2Vec embeddings really contributing to the predictions. The goal of the experimentation in section 5.8 was just that see the impact the changing features have on the results, and the results are clear and indicate that the RDF2Vec embeddings provide major contributions to the predictions and in deed in comparison with the chart events performance they are the features with the bigger role in the method this because the chart did provide a relevant value to the prediction. With this results is clear that the RDF2Vec implementation was very successful, leading to improvements on the predictions and with a clear reasoning that the embedding were the biggest contributors to this development.

For the final question **R3: How is predictive value impacted when predictions are made at different moments through a patient's ICU stay and with varying information available?** The results discussed in section 5.7 show that the procedure can be applied in a more realistic scenario, with AUC's generally improving as more information is available and with the best approaches always resulting in AUC above 0.8.

The main conclusion of this work is that combination of more information extracted from EHR with its semantic context as given by KG embeddings leads to improved predictive performance for ICU readmission. It is important to state that the use of KG embeddings to represent the clinical features is able to bring two advantages: it allows the representation of EHR information of different types in a common format, since it is possible to represent any number of diagnosis, tests, procedures, etc; and it allows the inclusion of scientific context through the use of the ontology annotations to create the vector representations.

6.2 Limitations

The main limitation on the method regards the data set used. The implemented method has only been applied to the MIMIC-III data set and the annotation pipeline fits its characteristics. Moreover, there is no full guarantee that the good results obtained here will generalize to other data in other clinical settings. Another limitation regards the model which is not adequately processing chart events features.The MIMIC-III data set records timestamps for every chart event feature but in an encrypted format to protect patients information. This fact makes it very difficult to handle properly. This work demonstrated that chart events are not really contributing to the prediction, but their potential is clearly untapped.

6.3 Future Work

The general methodology developed in this work can be generalized to other clinical data sets and even other predictive targets. In terms of application to other data sets, the annotation pipeline can be adapted to work over other clinical data sets, which is probably the methodological step that will require a greater amount of effort to apply in another scenario. However, the methodology can work with different selected ontologies for other domains. Once the annotations are produced, the remaining steps are directly applicable over the annotations. In terms of the target adaptability, as long as the target is available in the data set to define a training target for supervised learning, it is possible to train a model over the EHR data. In fact, the MIMIC-II data set has been used to predict other targets such as mortality, sepsis, stroke, etc., and a clear next step would be to train new models for these targets.

Moreover, this work focused on classical ML approaches, given the small size of the data. However, there is also an opportunity to explore KG embeddings using deep learning approaches, especially for larger data sets and a greater amount of features.

The methodology established in this work has a clear potential for impact, given its ability to be generalized to other clinical prediction scenarios and the ease of integration with other ML approaches. The very good performance obtained in the prediction of ICU readmission substantiates this potential.

Bibliography

Adams, D., Milton, S., Kazmierczak, E., and Lindenthal, J. (2014). Thesaurus and ontology structure: Formal and pragmatic differences and similarities: Thesaurus and ontology structure: Formal and pragmatic differences and similarities. *Journal of the Association for Information Science and Technology*, 66.

Adamusiak, T. and Bodenreider, O. (2012). Quality assurance in loinc using description logic. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012:1099–108.

Anand, R., Stey, P., Jain, S., Biron, D., Bhatt, H., Monteiro, K., Feller, E., Ranney, M., Sarkar, I., and Chen, E. (2018). Predicting mortality in diabetic icu patients using machine learning and severity indices. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:310–319.

Arbabi, A., Adams, D., Fidler, S., and Brudno, M. (2019). *Identifying Clinical Terms in Free-Text Notes Using Ontology-Guided Machine Learning*, pages 19–34.

Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., and Vidal, M.-E. (2018). Towards a knowledge graph for science. pages 1–6.

Bates, D., Saria, S., Ohno-Machado, L., and Shah, A. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health affairs (Project Hope)*, 33:1123–31.

Beaulieu-Jones, B., Orzechowski, P., and Moore, J. (2018). Mapping patient trajectories using longitudinal extraction and deep learning in the mimic-iii critical care database. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 23:123–132.

Bodenreider, O. (2008). Biomedical ontologies in action: Role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, 2008:67–79.

Bona, J., Brochhausen, M., and Hogan, W. (2019). Enhancing the drug ontology with semantically-rich representations of national drug codes and rxnorm unique concept identifiers. *BMC Bioinformatics*, 20:708.

Carey, K. and Stefos, T. (2015). The cost of hospital readmissions: Evidence from the va. *Health care management science*, 19.

Ceusters, W., Smith, B., and Goldberg, L. (2005). A terminological and ontological analysis of the nci thesaurus. *Methods of information in medicine*, 44:498–507.

Chen, J., Hu, P., Jiménez-Ruiz, E., Holter, O., Antonyrajah, D., and Horrocks, I. (2021). Owl2vec*: embedding of owl ontologies. *Machine Learning*, 110.

Chu, X., Ilyas, I., Krishnan, S., and Wang, J. (2016). Data cleaning: Overview and emerging challenges. pages 2201–2206.

Correa, T., Ponzoni, C., Rabello, R., Serpa, A., Assuncao, M., Pardini, A., and Shettino, G. (2017). Readmission to intensive care unit: incidence, risk factors, resource use and outcomes: a retrospective cohort study.

Craven, M., Gunopulos, D., and Ungar, L. (2006). Proceedings of the acm sigkdd international conference on knowledge discovery and data mining: Foreword. 2006:iii.

Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs.

Gazzotti, R., Faron-Zucker, C., Gandon, F., Lacroix-Hugues, V., and Darmon, D. (2020). Injection of automatically selected dbpedia subjects in electronic medical records to boost hospitalization prediction. pages 2013–2020.

Goldstein, B., Navar, A., Pencina, M., and Ioannidis, J. (2016). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, 24:ocw042.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. volume 2016, pages 855–864.

Gruber, T. (1993). A translational approach to portable ontologies. *Knowledge Acquisition*, 5:199–220.

Gruber, T. (1994). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43.

Hogan, W., Hanna, J., Joseph, E., and Brochhausen, M. (2013). Towards a consistent and scientifically accurate drug ontology. *CEUR Workshop Proceedings*, 1060:68–73.

Huang, L., Shea, A., Qian, H., Masurkar, A., Deng, H., and Liu, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of Biomedical Informatics*, 99:103291.

Huang, S., Cai, N., Pacheco, P., Narrandes, S., Wang, Y., and Xu, W. (2018). Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics & proteomics*, 15:41–51.

Jamei, M., Nisnevich, A., Wetchler, E., Sudat, S., and Liu, E. (2017). Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. *PLOS ONE*, 12:e0181173.

Javan, S., Sepehri, M. M., Layeghian, M., and Khatibi, T. (2019). An intelligent warning model for early prediction of cardiac arrest in sepsis patients. *Computer Methods and Programs in Biomedicine*, 178.

Jensen, P., Jensen, L., and Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature reviews. Genetics*, 13:395–405.

Johnson, A., Pollard, T., Shen, L., Lehman, L.-w., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L., and Mark, R. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Jonquet, C., Shah, N., and Musen, M. (2009). The open biomedical annotator. *Summit on translational bioinformatics*, 2009:56–60.

Jovanovic, J. and Bagheri, E. (2017). Semantic annotation in biomedicine: The current landscape. *Journal of Biomedical Semantics*, 8.

Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2:49–79.

Kononenko, O., Baysal, O., Holmes, R., and Godfrey, M. (2014). Mining modern repositories with elasticsearch.

Konopka, B. (2014). Biomedical ontologies—a review. *Biocybernetics and Biomedical Engineering*, 35.

Kulmanov, M., Smaili, F. Z., Gao, X., and Hoehndorf, R. (2020). Semantic similarity and machine learning with ontologies. *Briefings in Bioinformatics*, 22.

Lai, J.-I., Lin, H.-Y., Lai, J., Lin, P.-C., Chang, S.-C., and Tang, G.-J. (2012). Readmission to the intensive care unit: A population-based approach. *Journal of the Formosan Medical Association = Taiwan yi zhi*, 111:504–9.

Li, C. (2019). Preprocessing methods and pipelines of data mining: An overview.

Lin, Y.-W., Zhou, Y., Faghri, F., Shaw, M., and Campbell, R. (2019). Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLOS ONE*, 14:e0218942.

LOINC (2021). About loinc.

Lu, Q., Li, Y., de Silva, N., Kafle, S., Cao, J., Dou, D., Nguyen, T., Sen, P., Hailpern, B., and Reinwald, B. (2019). Learning electronic health records through hyperbolic embedding of medical ontologies. pages 338–346.

Lu, Q., Nguyen, T. H., and Dou, D. (2021). *Predicting Patient Readmission Risk from Medical Text via Knowledge Graph Enhanced Multiview Graph Convolution*, page 1990–1994. Association for Computing Machinery, New York, NY, USA.

Martínez-Romero, M., Jonquet, C., O'Connor, M., Graybeal, J., Pazos, A., and Musen, M. (2017). Ncbo ontology recommender 2.0: An enhanced approach for biomedical ontology recommendation. *Journal of Biomedical Semantics*, 8.

Miller, G. and Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6:1–28.

Min, H., Mobahi, H., Irvin, K., Avramovic, S., and Wojtusiak, J. (2017). Predicting activities of daily living for cancer patients using an ontology-guided machine learning methodology. *Journal of Biomedical Semantics*, 8.

Nichols, J., Chan, H., and Baker, M. (2018). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, 11.

Ontotext (2021). What is rdf?

Pepe, M., Janes, H., Li, C., Bossuyt, P., Feng, Z., and Hilden, J. (2016). Early-phase studies of biomarkers: What target sensitivity and specificity values might confer clinical utility? *Clinical Chemistry*, 62.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Pesquita, C., Faria, D., Falcão, A., Lord, P., and Couto, F. (2009). Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5:e1000443.

Protégé (2021). Protégé 5 documentation, ontology header.

Ristoski, P., Rosati, J., Di Noia, T., De Leone, R., and Paulheim, H. (2018). Rdf2vec: Rdf graph embeddings and their applications. *Semantic Web*, 10:1–32.

Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., and Wheeldin, B. (2007). The clef corpus: Semantic annotation of clinical text. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2007:625–9.

Sarica, A., Cerasa, A., and Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: A systematic review. *Frontiers in Aging Neuroscience*, 9:329.

Scherpf, M., Gräßer, F., Malberg, H., and Zaunseder, S. (2019). Predicting sepsis with a recurrent neural network using the mimic iii database. *Computers in Biology and Medicine*, 113:103395.

Schuurman, N. and Leszczynski, A. (2008). Ontologies for bioinformatics. *Bioinformatics and biology insights*, 2:187–200.

Shi, X., Prins, C., Pottelbergh, G., Mamouris, P., Vaes, B., and Moor, B. (2021). An automated data cleaning method for electronic health records by incorporating clinical knowledge. *BMC Medical Informatics and Decision Making*, 21.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R., Shah, N., Whetzel, P., and Lewis, S. (2007). The obo foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25:1251–5.

Storey, J., Rowland, J., Basic, D., and Conforti, D. (2001). A comparison of five clock scoring methods using roc (receiver operating characteristic) curve analysis. *International journal of geriatric psychiatry*, 16:394–9.

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. data knowl eng 25(1-2):161-197. *Data & Knowledge Engineering*, 25:161–197.

Suresh, H., Gong, J., and Guttag, J. (2018). Learning tasks for multitask learning: Heterogenous patient populations in the icu.

TABLEAU SOFTWARE, LLC, A. S. C. (2021). Guide to data cleaning: Definition, benefits, components, and how to clean your data.

Tanenblatt, M., Coden, A., and Sominsky, I. (2010). The conceptmapper approach to named entity recognition.

Tarca, A., Carey, V., Chen, X.-w., Romero, R., and Draghici, S. (2007). Machine learning and its applications to biology. *PLoS computational biology*, 3:e116.

Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1.

Webb, G. I., Keogh, E., and Miikkulainen, R. (2010). Naïve bayes. *Encyclopedia of machine learning*, 15:713–714.

Wei, W.-Q., Bastarache, L., Carroll, R., Marlo, J., Osterman, T., Gamazon, E., Cox, N., Roden, D., and Denny, J. (2017). Evaluating phecodes, clinical classification software, and icd-9-cm codes for phenome-wide association studies in the electronic health record. *PLOS ONE*, 12:e0175508.

Xu, Z., Feng, Y., Li, Y., Srivastava, A., Adekkanattu, P., Ancker, J., Jiang, G., Kiefer, R., Lee, K., Pacheco, J., Rasmussen, L., Pathak, J., Luo, Y., and Wang, F. (2019). Predictive modeling of the risk of acute kidney injury in critical care: A systematic investigation of the class imbalance problem. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2019:809–818.

Appendices

Appendix A

Folds Average Performance

Table A.1: Average performance results for the reproduction.

	Mean Fold Result	Standard Deviation
ROC-AUC-SVM	0,5916	0,0093
ROC-AUC-LR	0,5714	0,0094
ROC-AUC-RF	0,6176	0,0090
ROC-AUC-NB	0,5578	0,0047
PR-AUC-SVM	0,2576	0,0119
PR-AUC-LR	0,2378	0,0095
PR-AUC-RF	0,2634	0,0121
PR-AUC-NB	0,3196	0,0085

Table A.2: Average performance results for one ontology strategy at admission.

	Mean Fold Result	Standard Deviation
ROC-AUC-SVM	0,626	0,0140
ROC-AUC-LR	0,6102	0,0101
ROC-AUC-RF	0,6612	0,0068
ROC-AUC-NB	0,5018	0,0007
PR-AUC-SVM	0,3072	0,0171
PR-AUC-LR	0,292	0,0176
PR-AUC-RF	0,3234	0,0154
PR-AUC-NB	0,5962	0,0029

Table A.3: Average performance results for one ontology strategy at treatment.

	Mean Fold Result	Standard Deviation
ROC-AUC-SVM	0,813	0,0094
ROC-AUC-LR	0,8184	0,0130
ROC-AUC-RF	0,8274	0,0119
ROC-AUC-NB	0,6896	0,0177
PR-AUC-SVM	0,6222	0,0196
PR-AUC-LR	0,6362	0,0230
PR-AUC-RF	0,692	0,0186
PR-AUC-NB	0,5012	0,0226

Table A.4: Average performance results for one ontology strategy at diagnostic.

	Mean Fold Result	Standard Deviation
ROC-AUC-SVM	0,8004	0,0111
ROC-AUC-LR	0,817	0,0119
ROC-AUC-RF	0,8164	0,0104
ROC-AUC-NB	0,7034	0,0104
PR-AUC-SVM	0,5598	0,0141
PR-AUC-LR	0,6514	0,0168
PR-AUC-RF	0,669	0,0166
PR-AUC-NB	0,5126	0,0190

Table A.5: Average performance results for one ontology strategy at discharge.

	Mean Fold Result	Standard Deviation
ROC-AUC-SVM	0,8064	0,0102
ROC-AUC-LR	0,8204	0,0161
ROC-AUC-RF	0,826	0,0098
ROC-AUC-NB	0,6974	0,0149
PR-AUC-SVM	0,5872	0,0134
PR-AUC-LR	0,6496	0,0219
PR-AUC-RF	0,685	0,0190
PR-AUC-NB	0,5104	0,0221

Table A.6: Average performance results the multiple ontology strategy.

	Mean Fold Result	Standard Deviation
ROC-AUC-SVM	0,7676	0,0132
ROC-AUC-LR	0,6946	0,0058
ROC-AUC-RF	0,825	0,0056
ROC-AUC-NB	0,6896	0,0486
PR-AUC-SVM	0,3936	0,0093
PR-AUC-LR	0,2838	0,0090
PR-AUC-RF	0,4694	0,0139
PR-AUC-NB	0,4114	0,0770

Appendix B

Ontology coverage

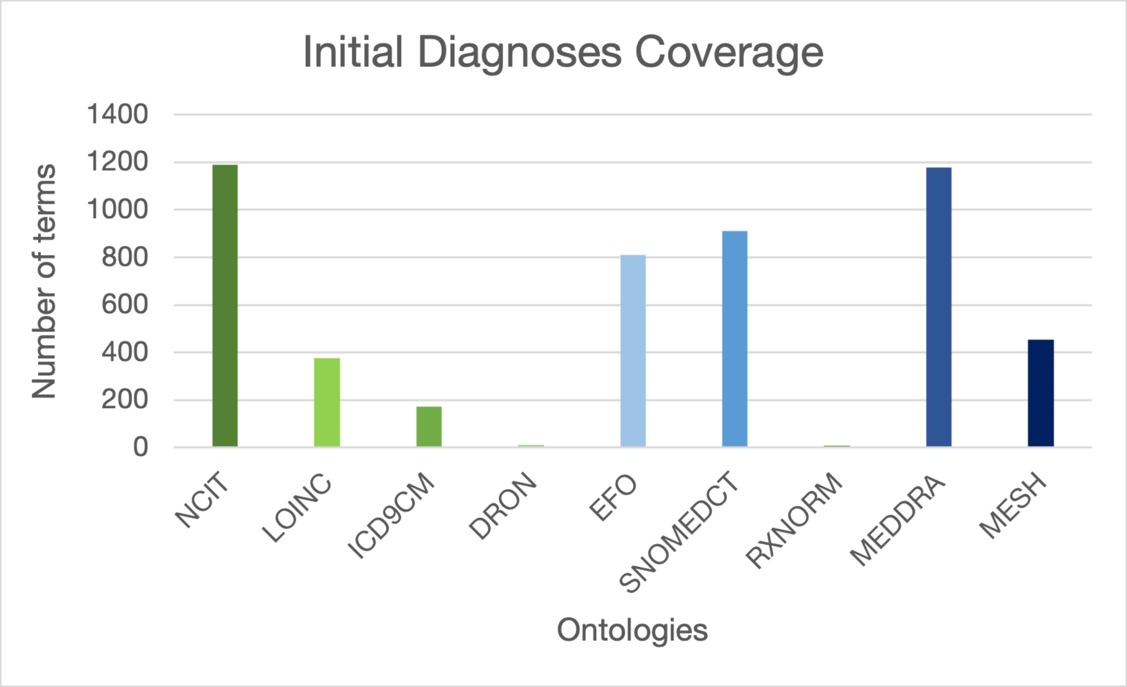


Figure B.1: Amount of initial diagnoses terms covered by each ontology.

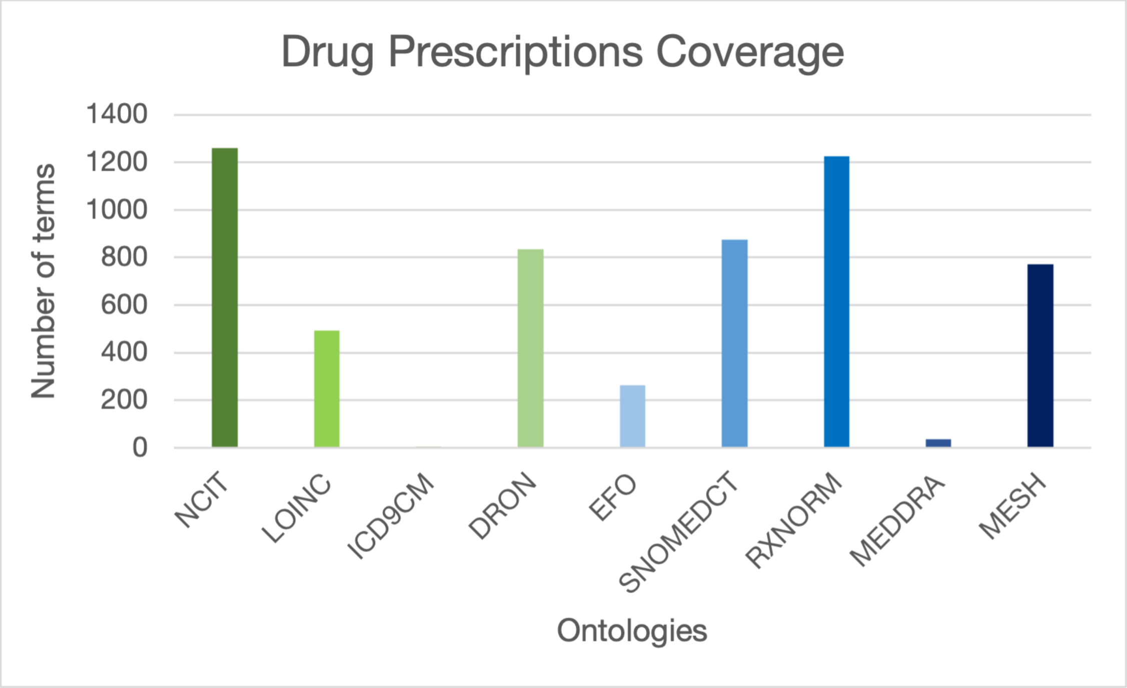


Figure B.2: Amount of drug prescriptions terms covered by each ontology.

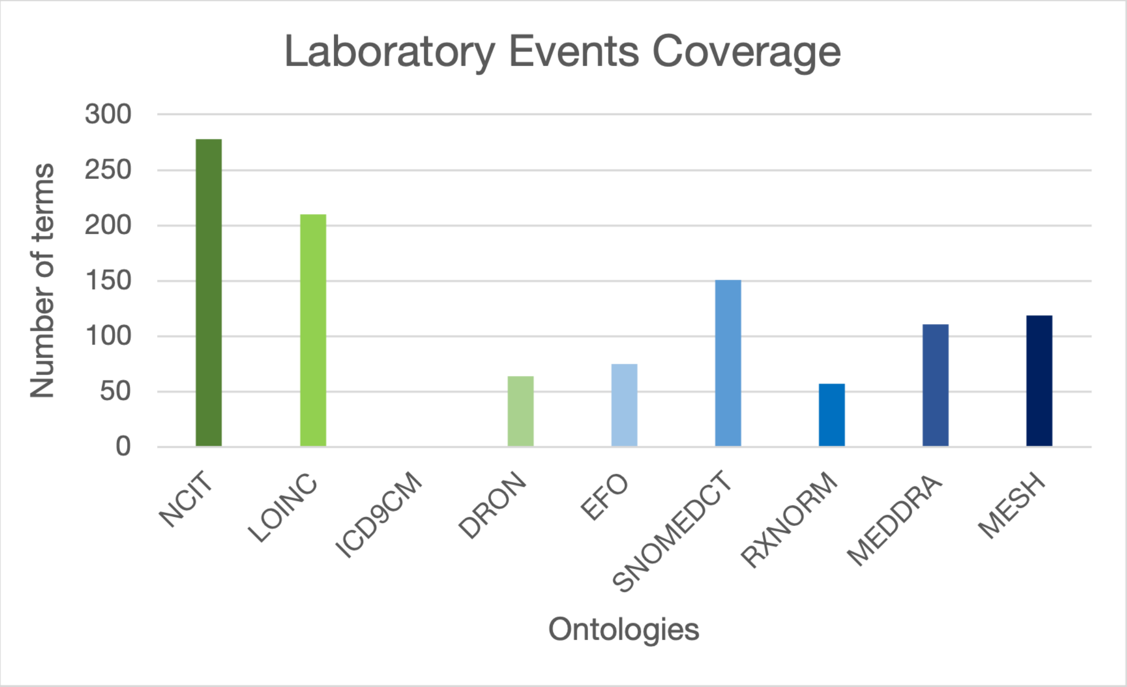


Figure B.3: Amount of laboratory events terms covered by each ontology.

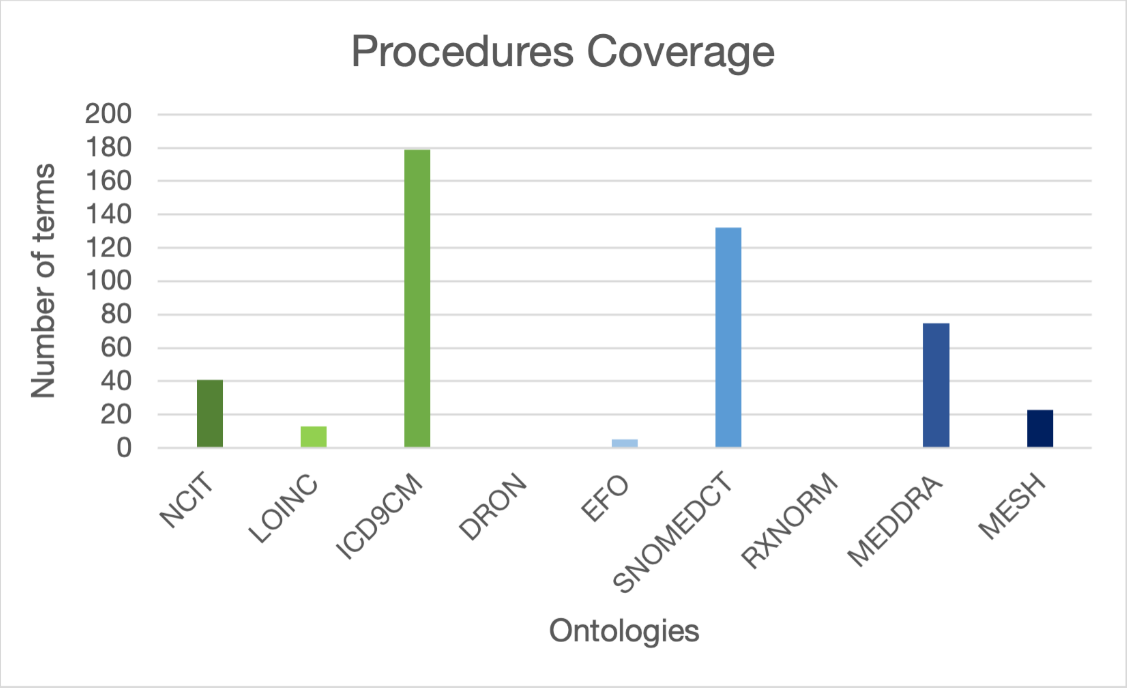


Figure B.4: Amount of procedures covered by each ontology.

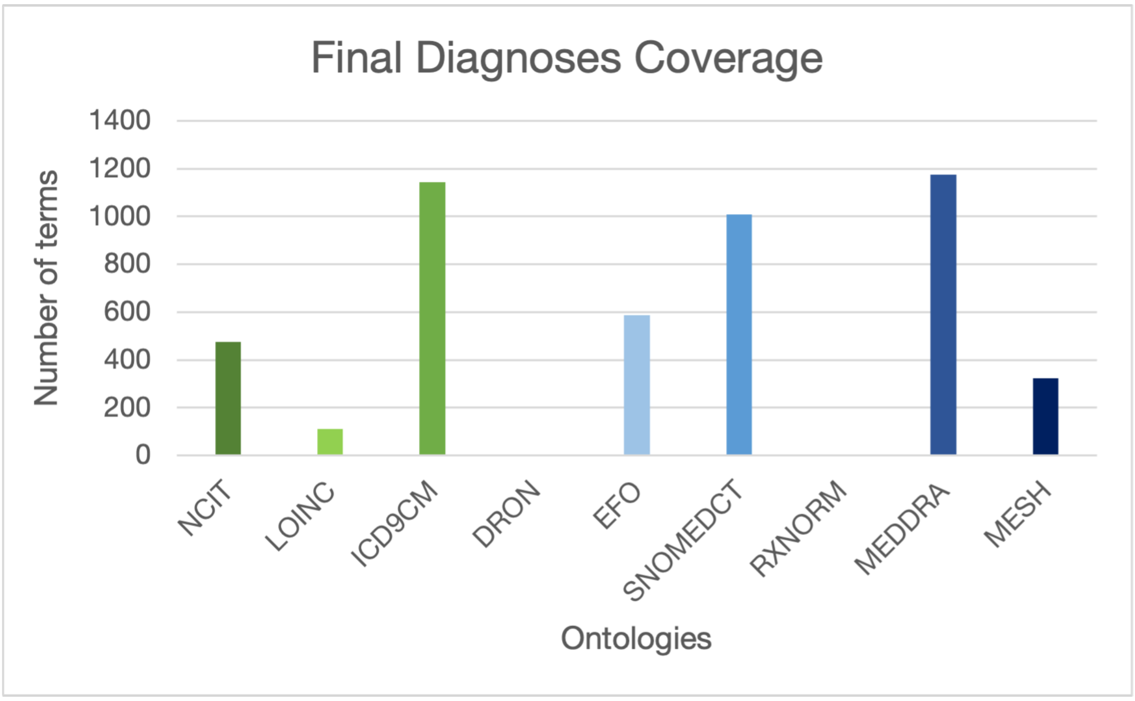


Figure B.5: Amount of final diagnoses terms covered by each ontology.

Appendix C

Embedding Combination approaches

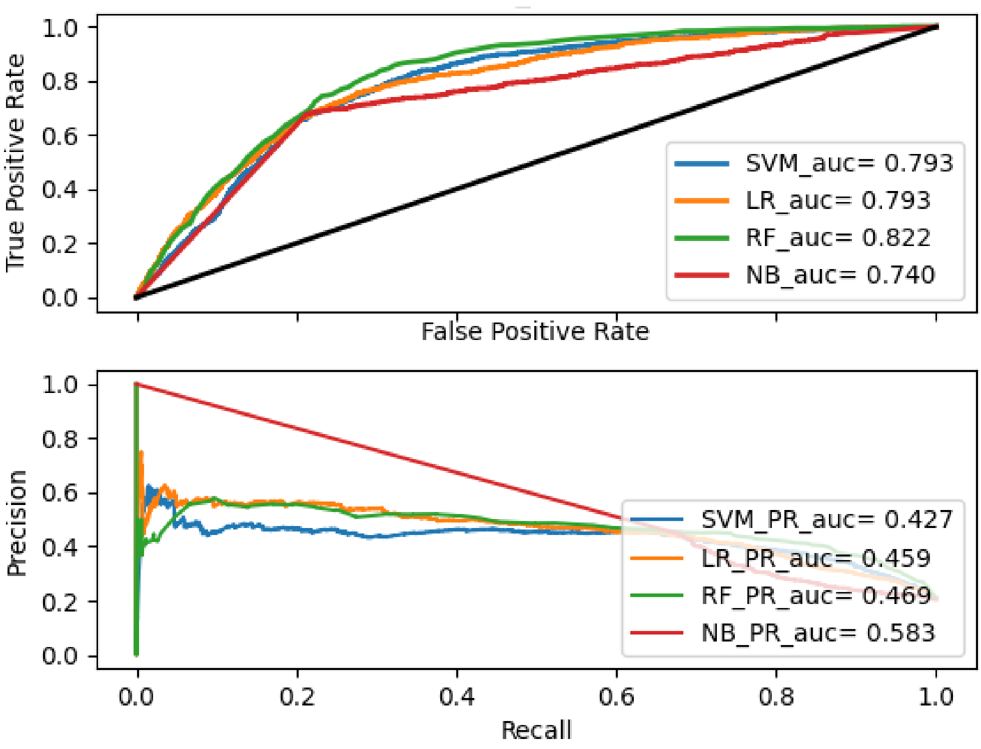


Figure C.1: Model performance values using weighted average as the embedding combination approach.

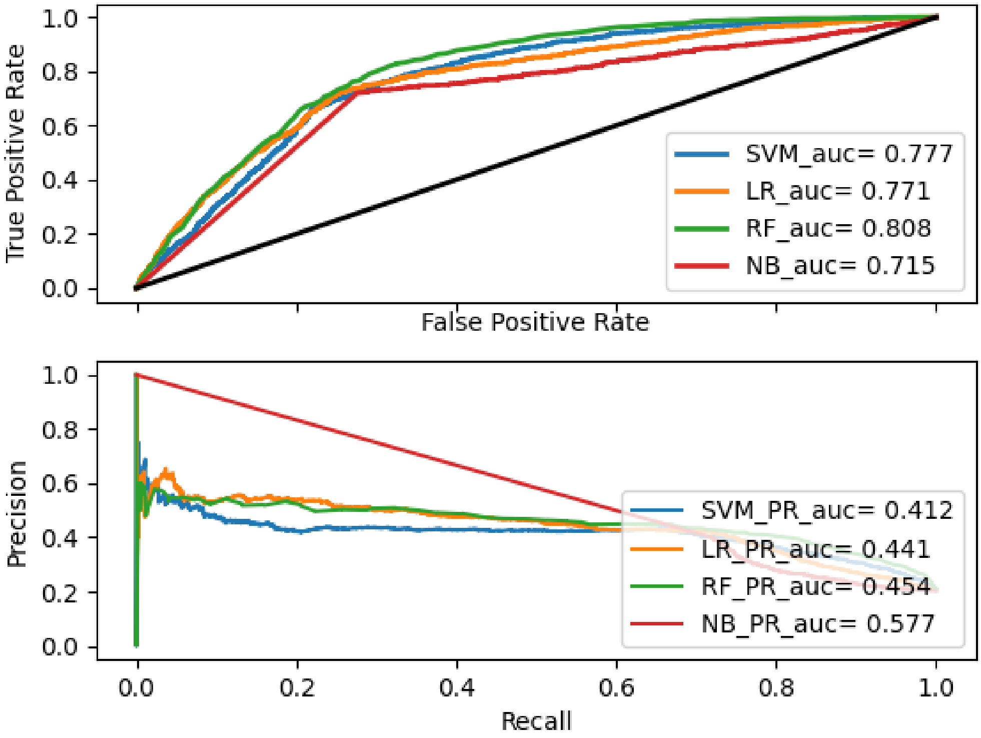


Figure C.2: Model performance values using multiplication as the embedding combination approach.

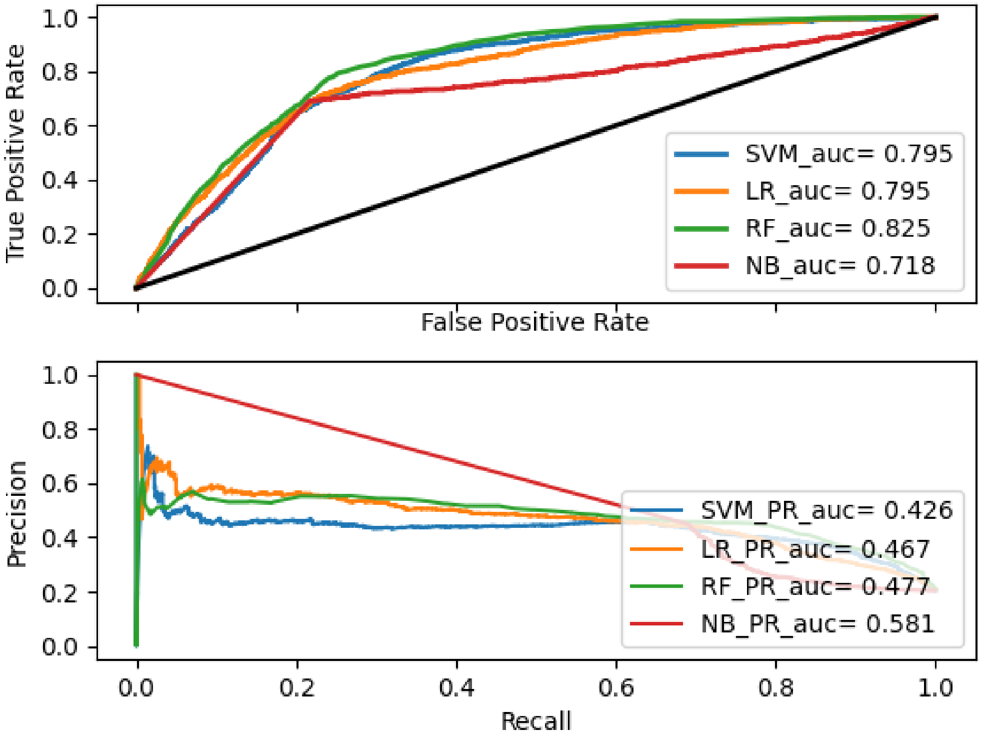


Figure C.3: Model performance values using the absolute values as the embedding combination approach.

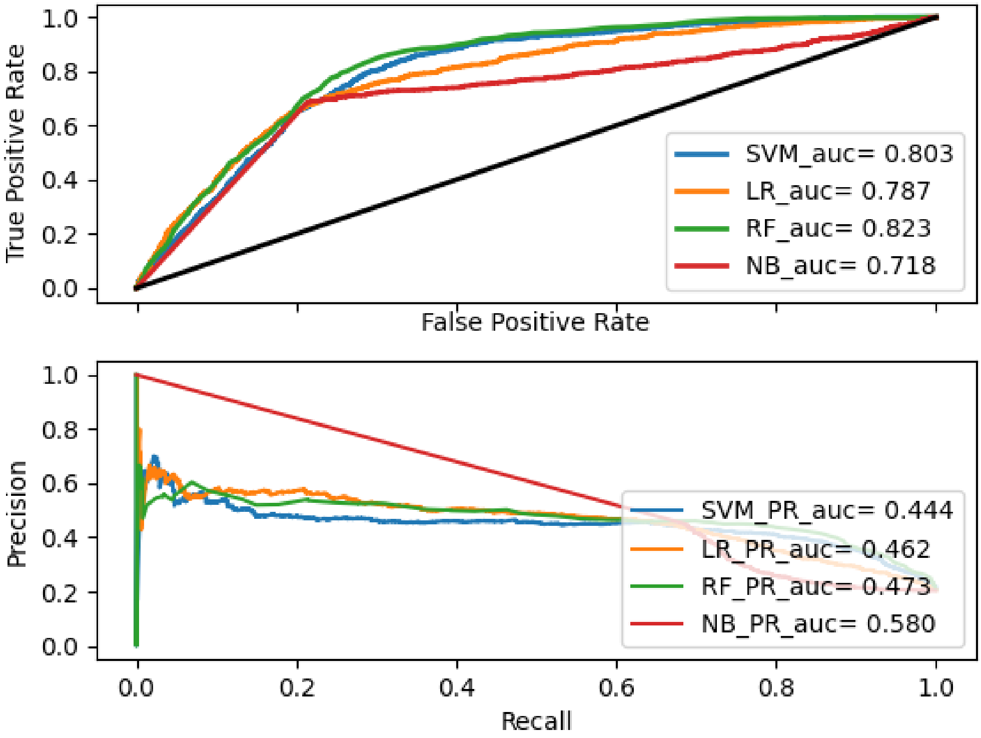


Figure C.4: Model performance values using the concatenation of the values as the embedding combination approach.