Faculdade de Engenharia da Universidade do Porto

TRIPOD - Text-based Risk Prioritisation of Dermatological Clinical Notes

Catarina Magalhães Dias

Master's Thesis

Mestrado Integrado em Bioengenharia - Engenharia Biomédica

Supervisor: Liliana Ferreira, PhD, feup | fraunhofer aicos

Junho 2021

TRIPOD - Text-based Risk Prioritisation of Dermatological Clinical Notes

Catarina Magalhães Dias

Mestrado Integrado em Bioengenharia - Engenharia Biomédica

Junho 2021

Resumo

O cancro da pele é uma doença com elevada incidência mundial. Em Portugal, um terço dos cancros diagnosticados anualmente são casos de cancro da pele. Existem limitações associadas ao diagnóstico de cancro da pele, como por exemplo o número reduzido de dermatologistas no Sistema Nacional de Saúde português disponíveis para a realização deste diagnóstico.

Recentemente, o uso de Inteligência Artificial tem vindo a ser implementado em tarefas de apoio à decisão médica, com o objetivo de melhorar os processos de diagnósticos, e assegurar que as condições dos doentes são detetadas em fases iniciais.

O presente documento tem como objetivo a apresentação do trabalho desenvolvido no Fraunhofer AICOS, no contexto de dissertação do Mestrado Integrado em Bioengenharia, no ramo de Engenharia Biomédica, na Faculdade de Engenharia da Universidade do Porto. O trabalho insere-se no projeto Derm.AI, que pretende utilizar Inteligência Artificial no rastreio dermatológico, e que pretende melhorar os processos de Teledermatologia entre Cuidados de Saúde Primários e serviços especializados em Dermatologia pertencentes ao Serviço Nacional de Saúde através da priorização de casos. Atualmente, existe um algoritmo de análise de imagem implementado. Contudo, existe informação presente nas notas clínicas que não está a ser utilizada no suporte à decisão. Assim, o trabalho desenvolvido foca-se no uso de registos textuais de notas clínicas dermatológicas para obter uma priorização dos casos de acordo com o risco.

A análise das notas clínicas é feita recorrendo a técnicas de Processamento de Linguagem Natural que permitem extrair informação relevante do texto que se relacionem com o diagnóstico e avaliação do risco do paciente. Assim, é desenvolvido um algoritmo para identificação de entidades clínicas de interesse nas notas clínicas de Dermatologia. A anotação dos dados com entidades clínicas permite que a informação extraída seja de interesse para a prática clínica de Dermatologia, assim como para o projeto Derm.AI. A validação da anotação feita por dois dermatologistas apresenta um rácio de concordância de $87.21 \pm 0.10\%$. As abordagens utilizadas para a identificação de entidades clínicas são *Conditional Random Fields*, *Bidirectional Long Short-Term Memory*, *Bidirectional Long Short-Term Memory* seguido de *Conditional Random Fields*, e modelos pré-treinados da arquitetura *Bidirectional Encoder Representations from Transformers*. Os valores de *macro average F1-score* obtidos para as abordagens mencionadas são 0.72, 0.66, 0.69, e 0.76, respetivamente.

Para provar que as entidades clínicas extraídas influenciam positivamente a previsão da priorização por risco de casos clínicos de Dermatologia, é implementada uma rede neuronal recorrente *Long Short-Term Memory*, utilizando as entidades extraídas pelo modelo supramencionado e o código *International Classification of Diseases* correspondente a cada caso clínico como input. Estratégias de *oversampling* e *undersampling* são adotadas para contrariar o desequilíbrio de classes, obtendo como melhor resultado um valor de *macro average recall* de 0.70. O peso da informação textual no modelo de priorização de risco já implementado no projeto Derm.AI é avaliado, sendo que o melhor resultado é alcançado quando se utiliza como *feature* a informação extraída das notas clínicas.

Os resultados obtidos para as tarefas de extração de informação e priorização dos casos com base no risco estão alinhados com os resultados apresentados na literatura. Por essa razão, pode ser afirmado que os principais objetivos do trabalho de dissertação são cumpridos.

Abstract

Skin cancer is a disease with a high incidence worldwide. In Portugal, one-third of the cancers diagnosed each year are cases of skin cancer. There are limitations associated with the diagnosis of skin cancer, such as the reduced number of dermatologists in the Portuguese National Health System available to perform this diagnosis.

Recently, Artificial Intelligence has been implemented in the context of medical decision support tasks, aiming to improve diagnostic processes and ensure that patient conditions are detected at early stages.

This document aims to present the work developed at Fraunhofer AICOS in the context of the dissertation for the Integrated Master's on Bioengineering, in the Biomedical Engineering field, at the Faculty of Engineering of the University of Porto. The work is part of the Derm.AI project, which intends to use Artificial Intelligence in dermatological screening and intends to improve the processes of Teledermatology between Primary Health Care and specialized services in Dermatology belonging to the National Health Service through case prioritisation. Currently, there is an image analysis algorithm implemented. However, there is information in the clinical notes which is not being used in the clinical decision support. Therefore, the implemented dissertation work utilises textual records of dermatological clinical notes to obtain a case risk prioritisation.

The clinical notes analysis is done using Natural Language Processing techniques that allow extracting relevant text characteristics that relate to the diagnosis and risk evaluation of the clinical case. Thus, an algorithm is developed to identify clinical entities of interest in Dermatology clinical notes. The data annotation with clinical entities allows the extraction of information of interest for Dermatology clinical practice, as well as for the Derm.AI project. The annotation validation by two dermatologists obtains an agreement ratio of $87.21 \pm 0.10\%$. The adopted approaches for clinical entity identification are Conditional Random Fields, Bidirectional Long Short-Term Memory, Bidirectional Long Short-Term Memory followed by Conditional Random Fields and pre-trained Bidirectional Encoder Representations from Transformers architecture models. The macro average F1-score values for the mentioned approaches are 0.72, 0.66, 0.69, and 0.76, respectively.

To prove that the extracted clinical entities can positively impact the risk prioritisation prediction, a Long Short-Term Memory recurrent neural network is implemented, using the entities extracted by the above-mentioned model, and the International Classification of Diseases code corresponding to each clinical case as input. Oversampling and undersampling strategies are adopted to counteract class imbalance, obtaining as result a macro average recall value of 0.70. The evaluation of the influence of textual data in the already implemented Derm.AI risk prioritisation model is also performed, with the best result being achieved when using information extracted from the clinical notes as feature.

The results obtained for the information extraction and clinical case risk prioritisation tasks are in line with the results presented in the literature. For this reason, it can be stated that the main goals of the dissertation work are achieved.

Acknowledgments

E, assim, chega ao fim este capítulo de cinco anos. Um final que não é de todo o que imaginava no passado, pelas experiências que quem termina agora o percurso não pôde viver e que nunca mais viverá. No entanto, se há algo que estes últimos tempos ensinaram, é que as coisas nem sempre correm como esperamos. Um ensinamento que, para mim, também esteve muito presente no decorrer do curso. Ao longo destes anos, tive cada vez mais noção de que nos cabe a nós decidir como lidar com as frustrações e percalços. Umas vezes melhor, outras pior, acabo este percurso com o sentimento que aproveitei da melhor forma quanto possível estes anos, que tantas vezes ouvimos dizer que serão os mais felizes da nossa vida. E porque, no final de contas, são as pessoas que nos acompanham que fazem a diferença e tornam este percurso nos anos mais felizes e memoráveis, sem nenhuma ordem de importância:

À minha orientadora, Prof. Liliana Ferreira, pela confiança no meu trabalho e o apoio ao longo dos últimos meses, mesmo nos momentos mais complicados. Aos envolvidos no projeto Derm.AI, agradeço a contribuição e as opiniões construtivas ao longo do desenvolvimento da dissertação. Agradeço ainda a colaboração e disponibilidade da equipa editorial da Revista da Sociedade Portuguesa de Dermatologia e Venereologia no envio de casos clínicos utilizados na elaboração do trabalho.

Aos meus pais, Anabela e António, porque se devo tudo o que alcancei a alguém, é a vocês. Vocês que sempre me deram liberdade para seguir o caminho que idealizo, que fazem os possíveis e os impossíveis para me ajudar e facilitar a minha vida ao máximo, e acreditam em mim incondicionalmente. À Bia, a personificação da irmã que vai ao armário desviar umas peças de roupa sem autorização. Embora sejas uma chatinha, a vida não tinha tanta piada se não tivesse uma irmã mais nova com quem partilhar todos os momentos.

Aos meus avós, Fernanda e Valdemar, que passados cinco anos e muitas pesquisas no tablet, já sabem que Bioengenharia dá para muitas coisas. Obrigada pelo apoio incondicional, pelo orgulho que têm nas netas, e pelas palavras reconfortantes, muitas delas eternizadas em forma de poema. Aos meus padrinhos, Helena e Jorge, e à minha prima Inês, por estarem sempre à distância de uma chamada e prontos para ajudar em tudo o que eu precisar.

À Alexandra, Carlota, Francisca, Mónica, Vera e Viviana, as amigas de Espinho, desde as que me acompanham desde que me consigo lembrar até às surpresas que o secundário trouxe. Que continuemos unidas por muito tempo e que, independentemente dos conflitos de agenda, as tradições continuem a ser mantidas.

À Xana, uma das primeiras pessoas que tive o privilégio de conhecer na FEUP, e que claramente falhou na primeira impressão que teve sobre mim. A pessoa que tem sempre uma palavra sábia para partilhar, e que é totalmente apaixonada pelas causas que defende. A minha casa tem sempre uvas verdes à tua espera, no caso de voltares a Portugal. À Mariana, companheira de playlists de lo-fi com os temas mais aleatórios e futura DJ do Paredes de Coura. Que venham mais passeios de gaivota em Amarante e serões a jogar Sims com direito a bolachinhas de manteiga. E, embora sejas a pior contadora de histórias de que há memória, que venham também muito mais histórias para contar. À Inês, a pessoa que esteve presente em todas as histórias desde o 1º ano, e melhor colega de casa com quem poderia ter dividido o 2ºN (desde que não mexa nos chás, não há problema). Desde as referências a apanhados, músicas para o Akira e gelado de menta como gelado favorito, até aos momentos mais difíceis, as revoltas com a vida,

os desabafos e os conselhos, fica uma amizade que acompanhou muitas fases diferentes e que tenho a certeza que é para sempre. E para ficar totalmente esclarecido, depois destes anos todos, eu acho que percebes português. A estas três amigas, as melhores que podia ter pedido. Que em Esmoriz sejamos sempre feliz(es), que o nosso momento Donna and the Dynamos não demore a chegar e que continuemos a criar muitas memórias juntas.

Ao Miguel, o colega de casa por afinidade e presença assídua na fisioterapia, que mostra sempre os melhores stand-ups. Ao Venâncio, o maior Portista que conheço, melhor cantor, e viciado em regras de jogos de tabuleiro (mas só quando o favorecem). Ainda estou a pensar se te devemos desculpar por não alinhares nos planos quando efetivamente podíamos sair. Ao Henrique, a pessoa mais proativa com que me cruzei nos últimos anos, melhor organizador viagens, e most likely to seguir uma carreira política, depois do sucesso na presidência no NEB.

Ao OG, porque ninguém adivinharia que de um almoço aleatório na FEP no 1º ano iria formar este grupo, que se manteve unido ao longo destes cinco anos. Espero que os comentários desportivos continuem por muito tempo, eu vou fazer o meu melhor para me manter a par da discussão.

À Cat Morgado, a amizade que Biomédica voltou a aproximar, fã número 1 de Bruno Aleixo, companheira de aulas no GoGym e de reality shows e, juntamente com a Inês, de Erasmus que virou quarentena no Z11. Única pessoa que, na Holanda, preferiu um autocarro à sua bicicleta, e que trouxe a música brasileira em força para a playlist daqueles cinco meses. Um dia voltamos para ver como anda o nosso amigo Peaky Blinder. Senão, a tão aguardada visita ao Parque Verde do Bonito não fica atrás. À Rita Barros, responsável pelas melhores prendas dos amigos secretos e que sabe sempre onde comer o melhor sushi. Espero que, quando uma porta se fechar (ou a Xana a estragar), estejas lá para fazermos os melhores TikToks. À Rita Sampaio (Ritinha para os amigos), pessoa com maior dificuldade em seguir receitas de bolos e que, juntamente com o Henrique, me convenceu a aceitar o desafio que marcou o meu ano de finalista. Olhando para trás, fico muito feliz de me ter juntado a vocês e por tudo o que conseguimos alcançar dadas as circunstâncias. Acabamos o ano com uma sensação de dever cumprido, e mais unidos que nunca.

Ao NEB, a aventura inesperada neste último ano. Em equipa, conseguimos superar todos os desafios e conquistar ainda mais do que imaginávamos ser possível num ano em que o trabalho foi feito à distância. Foi um prazer trabalhar com todas as pessoas incríveis do NEB, e tenho a certeza que estará em boas mãos daqui para a frente.

Ao Diogo, porque sem ti este percurso não estava completo. Obrigada por seres o melhor companheiro de teletrabalho nestes últimos tempos, o melhor companheiro de viagens (que estão claramente a sofrer um upgrade), e por ligares para os restaurantes por mim. Obrigada por ouvires todas as minhas queixas e por estares sempre ao meu lado, nos momentos bons e nos menos bons. Nos últimos quatro anos crescemos muito e aprendemos muito ao lado um do outro. Que venham muitos mais anos desta nossa história e que continuemos a ser esta dupla fantástica, um mais engraçado e o outro mais fotogénico (tu sabes quem é quem).

Resta a certeza de que não podia ter sido mais feliz, e que levo comigo para a vida os momentos e as pessoas incríveis que me acompanharam.

Catarina Dias

"Para ser grande, sê inteiro: nada
Teu exagera ou exclui.
Sê todo em cada coisa. Põe quanto és
No mínimo que fazes.
Assim em cada lago a lua toda
Brilha, porque alta vive."

"Odes de Ricardo Reis", Fernando Pessoa

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under Curve
B	Beginning Tag
BCC	Basal Cell Carcinoma
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BioBERT	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
CTH	Consulta a Tempo e Horas
DL	Deep Learning
Derm.AI	Usage of Artificial Intelligence to Power Teledermatological Screening
EHR	Electronic Health Records
ELMo	Embeddings from Language Models
GPT	Generative Pre-trained Transformer
I	Inside Tag
ICD	International Classification of Diseases
ICF	International Classification of Functioning, Disability and Health
ICT	Information Communication Technology
IOB	Inside-Outside-Beginning
LSTM	Long Short-Term Memory
ML	Machine Learning
MSC	Melanoma Skin Cancer
NE	Named Entity
NER	Named Entity Recognition
NHS	National Health Service
NLP	Natural Language Processing
NMSC	Non-Melanoma Skin Cancer
O	Out Tag
POS	Part of Speech
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SCC	Squamous Cell Carcinoma
SMOTE	Synthetic Minority Oversampling Technique
UMLS	Unified Medical Language System
WHO	World Health Organization

Chapter 1

Introduction

1.1 Context

Cancer is a class of diseases characterised by uncontrolled growth and spread of abnormal cells, which is a significant subject in public health worldwide [1]. In 2020, there were 19 million new cancer cases, whereas the number of deaths caused by cancer was estimated in approximately 10 million[1]. In the same year, in Portugal, around 60 thousand new cases arose, and 30 thousand individuals died due to cancer[2]. Skin cancer is one of the most common types of cancer and comprises a range of pathologies that occur from different cells of the epidermis and dermis [2]. Skin cancer can be classified in two major subtypes: Melanoma Skin Cancer (MSC) and Non-Melanoma Skin Cancer (NMSC) [3].

MSC was the 17th most prevalent type of cancer worldwide in 2020[3], and accounts for 70% of the mortality due to skin cancer [4]. The incidence of NMSC is higher than the incidence of melanoma[4], although its epidemiology is less studied and data often excluded from cancer registries [5], [6].

In Portugal, one-third of all cancers detected annually are diagnosed as skin cancer. In 2019, the Portuguese Association of Cutaneous Cancer estimated that around 13 thousand skin cancer cases were diagnosed. It is considered that 6 to 8 melanoma cases per 100 thousand habitants are registered and that the number of NMSC cases is ten times higher [7]. According to the International Agency for Research on Cancer, in 2020, there were 1,071 diagnosed melanoma cases and 289 deaths as consequence of MSC in Portugal, making it the 17th most prevalent cancer in the country that year[2].

There is an economic burden for public health services caused by the increasing skin cancer incidence. In the United States, the annual cost for NMSC medical care is around 650 million dollars [8], with an average cost per NMSC diagnosed of around 2.5 thousand dollars. The cost

[1] "All cancers - Globocan 2020" https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf. Accessed on 17-06-2021

[2] "Portugal Fact Sheet 2020" https://gco.iarc.fr/today/data/factsheets/populations/620-portugal-fact-sheets.pdf. Accessed on 17-06-2021

[3] "Melanoma of Skin - Globocan 2020" https://gco.iarc.fr/today/data/factsheets/cancers/16-Melanoma-of-skin-fact-sheet.pdf. Accessed on 07-06-2021

[4] "Non-Melanoma Skin Cancer - Globocan 2020" https://gco.iarc.fr/today/data/factsheets/cancers/17-Non-melanoma-skin-cancer-fact-sheet.pdf.Accessed on 17-06-2021

per MSC case is almost 33 thousand dollars, making the average cost ten times more expensive in melanoma cases [9].

The stage in which skin cancer is diagnosed can have an effect on the patient recovery. If detected in an early stage, the success rates for the treatment are considerably high. In melanoma cases, the estimated 5-year survival rate decreases from over 99% if detected in an early stage to about 14% if detected in a late stage [10]. In some cases, the cancer is not diagnosed in an early stage. One of the factors that can lead to a delayed diagnosis process may be the shortage of Dermatology resources in the Portuguese National Health Service (NHS). The number of dermatologists working in the Portuguese NHS represent only 60% of the estimated resources necessary for the traditional diagnosis procedure - total body scanning and skin biopsy - to work efficiently[5]. Besides, there is an uneven regional distribution of these specialists, which makes access to treatment less balanced.

That being said, it is fundamental that the diagnostics are as fast and efficient as possible, not only for skin cancer, but also for other dermatological conditions. Moreover, given the shortage in resources in the Dermatology specialty, it is important that cases of higher associated risk, such as malignant conditions, are directed to specialty follow-ups with more urgency.

In recent years, almost all healthcare facilities adopted health information systems, changing from paper records to electronic health records. Electronic health records are a fast and cheap alternative to more traditional types of medical records and allow for the development of predictive models that support a prognostic. From 2004 to 2014, the percentage of Portuguese hospitals that used electronic health records increased from 42% to 83% [11]. The shift towards electronic health records generates a large amount of data. The data increase promotes the development of Artificial Intelligence (AI) tools. In clinical applications, AI utilises the clinical data into big data analytic models that are able to extract clinically relevant information from the data. That information can assist physicians in certain clinical decisions, or replace the decision in more functional areas of medicine [12]. The application of AI in the clinical field promises to improve patient and clinical team outcomes, and consequently influencing population health. Besides, the decision assistance in the clinical practice allows for cost reductions [13]. Considering the burden in public health services, as well as the increasing amount electronic health records, there has been a growing interest in Telemedicine and Information Communications Technology (ICT) solutions to improve diagnosis efficiency.

Taking into consideration the diagnostic improvement and risk prioritisation of Dermatology clinical cases, and the potential of AI in the healthcare field, Derm.AI - Usage of Artificial Intelligence to Power Teledermatological Screening - emerges. Derm.AI is a project developed by Fraunhofer AICOS in partnership with the Portuguese NHS, aiming to enhance the existing Teledermatology processes between Primary Care Units and Dermatology Services in the NHS for skin lesion diagnosis through AI usage[6].

[5]RSE Siga Website - "RSE SIGA Derm.AI": `https://rse-siga.spms.min-saude.pt/category/rse-siga-derm-ai/`. Accessed on 25-01-2021

[6]Fraunhofer AICOS Website - "Derm.AI - Usage of Artificial Intelligence to Power Teledermatological Screening": `https://www.aicos.fraunhofer.pt/en/our_work/projects/dermai.html`. Accessed on 25-01-2021

1.2 Motivation

Nowadays, in Portugal, almost all healthcare facilities use health informatics systems and electronic health records [11]. There are specific clinical notes that are obtained as unstructured documents and contain valuable clinical information. The data unstructured nature and quantity, as well as the use of abbreviations and grammatically incorrect sentences, makes the information extraction from these documents more difficult.

Having the interoperability between healthcare facilities into account, and to make access to the first speciality appointment more manageable, expediting and facilitating appointment scheduling, the Portuguese NHS established "Consulta a Tempo e Horas" (CTH). CTH uses an electronic system and allows for appointment requests in a medical speciality[7]. There is a clinical screening in the first appointment performed by a primary care health physician. The primary care physicians can indicate if they think the speciality appointment is urgent. With that information into account, a specialist analyses the patient record and decides when the appointment is scheduled depending on the inherent risk[8]. However, the inability of the primary care physician to perform a skin cancer diagnosis is a problem in Portugal since primary care physicians do not have formation nor resources to do this kind of dermatological evaluation [7].

In Dermatology and skin cancer diagnosis, Derm.AI can be used to accelerate and support the clinical decision of risk prioritisation performed by the dermatologist, using information collected by the primary care physician. Image data is the most analysed type of data, since photographs of the dermatological condition are essential for the diagnosis phase, and the dermatologist needs to examine it and look for risk factors, such as alterations or significant growth. These factors are widely known as the ABCDE mnemonic: asymmetry, border irregularity, colour changes, diameter and evolution [14].

An image analysis algorithm is already implemented in Derm.AI. However, there is a great amount of textual data that ends not being employed in clinical decision situations because of its unstructured nature. Hence, extracting information from the unstructured clinical textual data allows to feed the risk prioritisation algorithm with textual information and improve the the prediction outcome. In the context of CTH, the risk prioritisation would be done using both image and text data, and would adapt the speciality appointment wait time to the predicted risk.

For the above mentioned reasons, it is desired that the Derm.AI algorithm comprises both dermatological imaging analysis and a textual data algorithm for the unstructured clinical notes. The presented dissertation work proposes the implementation of Natural Language Processing (NLP) methods using Dermatology clinical notes to support the risk prioritisation process.

[7] Ministério da Saúde - Portaria nº 95/2013.: `http://www.sg.min-saude.pt/NR/rdonlyres/4D921E90-4382-4E9E-B682-3FE85F261D87/34814/Portaria95%7B%5C_%7D2013%7B%5C_%7Dreferenciacao.pdf`. Accessed on 07-12-2020

[8] Consulta a Tempo e Horas: `https://www2.acss.min-saude.pt/DepartamentoseUnidades/UnidadeAcessoeContratualiza%7B%5Cc%7Bc%7D%7D%7B%5C~%7Ba%7D%7Do/ConsultaaTempoeHoras/tabid/501/language/pt-PT/Default.aspx`. Accessed on 26-01-2021

1.3 Goals

The main goal of the Master's thesis work is to develop a pipeline for risk prioritisation of unstructured dermatological clinical notes.

Considering that the textual data in the dataset used in the dissertation is unstructured, it is important to structure the data to ease information extraction. Taking that into account, clinical entities of interest in the context of dermatology and the Derm.AI project must be defined and extracted through a manual annotation process. Utilising the manually annotated data, it is desired that a clinical entity extraction model is implemented to automatically extract clinical entities from unstructured clinical notes.

The structured information extracted from the clinical notes can then be utilised as input for other algorithms. It is intended to use the clinical entities, as well as other structured information found in the dataset, to implement a risk prioritisation prediction model. This model aims to assess the impact that the extracted clinical entities from the clinical notes can have on the already implemented Derm.AI algorithm and understand how it can influence the risk prioritisation prediction outcomes.

1.4 Contributions

The contributions of this work are:

- Comprehensive analysis of the dataset of unstructured dermatological clinical notes from the Portuguese NHS
- Selection of clinical entities of interest for Dermatology clinical cases and Derm.AI project
- Annotation of part of the unstructured dermatological clinical notes dataset from the Portuguese NHS with selected clinical entities
- Implementation of word embedding model using clinical cases extracted from Dermatology journal
- Implementation of Dermatology clinical notes clinical entity extraction models
- Comparison between clinical entity extraction implemented approaches
- Implementation of Dermatology clinical notes risk prioritisation prediction model using extracted clinical entities and structured data from dataset
- Analysis of the effect of data-balancing methods in risk prioritisation prediction task
- Integration of clinical information extraction with existing Derm.AI algorithm for risk prioritisation based on image analysis

1.5 Document Structure

The document chapter organisation is further described.

Chapter 2 includes the theoretical background knowledge to understand the work developed, from skin cancer and electronic health records, to NLP tasks, namely information extraction, and Machine and Deep Learning approaches for their implementation.

Chapter 3 provides a review on the literature for information extraction and classification using clinical text. The state-of-the-art approaches for English and Portuguese clinical text are analysed.

Chapter 4 focuses on the dataset analysis regarding the structured and unstructured information available and the data distribution for certain variables. Besides, the annotation process is described, from the clinical entity selection to the manual annotation and corresponding validation. The word embedding model implementation is also outlined.

The clinical entity extraction task is reviewed in Chapter 5. Based on the literature review, four different approaches are implemented and the results are further compared.

Chapter 6 aims to analyse the effectiveness of a risk prioritisation algorithm using extracted clinical entities, along structured data from the clinical cases, as input. Due to the imbalance between priority values, data-balancing techniques, including oversampling and undersampling, are considered in this task. The addition of extracted clinical information as input in the already implemented Derm.AI prioritisation model is also assessed.

Lastly, Chapter 7 includes final remarks regarding the achieved goals in dissertation work, as well as further NLP tasks of interest to tackle in the context of the Derm.AI project.

Chapter 2

Background

This chapter includes a background on several foundation topics considered in the developed dissertation work. A context on skin cancer types, incidence, risk factors and diagnosis is presented. Health informatics systems and electronic health records are also analysed, as well as NLP and its most common methodologies and tasks, and knowledge sources utilised in clinical domain problems. An introduction to shallow Machine Learning (ML) and Deep Learning (DL) concepts and its usual pipeline is granted, also considering the models evaluation metrics more frequently used in NLP tasks.

2.1 Skin Cancer

Skin cancer includes a range of pathologies that result from the abnormal growth of different cells of the epidermis and dermis [2]. Skin cancer can be classified into two major subtypes: MSC and NMSC [3].

2.1.1 Melanoma Skin Cancer

MSC has origin in the melanocytes, a melanin-producing cell. Melanoma can occur cutaneously, as well as non-cutaneously, in ocular, gastrointestinal, genitourinary, or nasopharyngeal sites [9]. Cutaneous melanomas are the more usual, accounting for around 92% of the diagnosed cases. The four sub-types of this type of melanoma are the following:

- Superficial spreading melanoma - most usual melanoma type, in which there is a lateral growth before vertical growth happening.
- Nodular melanoma - generally present a blue or black colour, and vertical growth occurs in an early stage.
- Lentigo maligna melanoma - associated with chronic sun-exposure and, therefore, more common in older patients. The skin lesion begins as a freckle and becomes darker and more asymmetric with time.

- Acral lentiginous - melanoma type in which lesions are usual located in palms, soles, or even mucosal tissues.

The thickness of the lesion, associated with the vertical growth, is directly related to the melanoma prognosis: a thicker lesion has higher risks associated, as well as a higher mortality rate in that specific case [15].

MSC was the 17th most prevalent type of cancer worldwide in 2020[3]. Almost 70% of the deaths associated with skin cancer cases happen when the skin cancer subtype was MSC [4]. In 2020, approximately 324 thousand new MSC cases arose, while the number of MSC-related deaths worldwide in the same year was around 57 thousand[3].

2.1.2 Non-Melanoma Skin Cancer

NMSC originates in keratinocyte carcinomas, in particular, basal cell carcinomas (BCC) and squamous cell carcinomas (SCC) [16]. Both cell carcinomas originate in epidermic cells and have similar epidemiological and carcinogenic characteristics. Regarding incidence values, BCC is more frequent than SCC, existing a 4:1 standardised diagnosis ratio of BCC to SCC cases [5].

Although NMSC incidence is higher than MSC incidence, its epidemiology is less studied and data often excluded from cancer registries [5], [6]. In 2020, NMSC was the 5th most prevalent cancer worldwide, with an incidence of around 1,198 thousand cases[4].

2.1.3 Skin Cancer Diagnosis

In regards to the detection phase for skin cancer, Skin self-examination can be considered the first skin cancer detection phase. Skin self-examination is of extreme importance, since approximately 40% of MSC are self-detected by the patients [14]. A total body exam can be performed by dermatologists or primary care physicians, in case of suspicious growth or change. Besides the total skin scanning, patients also give information about their medical history and possible risk factors for skin cancer development. This exam should be periodic for patients who already had skin cancer[9]. Non-invasive optical technologies can be used to facilitate screening. Dermatoscopy is one of the technologies used, that allows an analysis of the borders of a lesion, epidermal organisation, and layer thickening, and can help improving the diagnosis accuracy [17].

A skin biopsy has to be performed in order to diagnose skin cancer. The dermatologist removes a sample of skin tissue to be further analysed in a laboratory. Information about the cancer stage can be obtained after the biopsy[9].

The stage of development in which skin cancer is diagnosed can influence the patient recovery. If detected in an early stage, the success rates for the treatment are considerably high, with a five-year survival rate of 99%. That rate decreases to around 14% if the cancer is diagnosed in a later stage [10]. Despite its importance, the diagnosis phase has several issues, such as a shortage in the number of dermatologists working in the NHS. Besides, unlike in other countries, primary care physicians in Portugal do not have necessary formation and tools to be able

[9] "How Is Skin Cancer Detected and Diagnosed?": `https://skincancer.net/diagnosis/`.Accessed on 05-01-2021

to perform the skin cancer diagnosis [7]. However, the first screenings are done by them, and using the CTH system, it is possible for primary care physicians to request appointments in a medical specialty[7]. Strategies, such as the developed dissertation work, are needed to prioritise the medical records resulting from such screenings so that those who are considered with a higher risk of developing skin cancer have an early specialty appointment.

2.2 Electronic Health Records

In recent years, there has been an increase of computational resources use in several areas, that leads to a consequent increase of data production and storage. In healthcare, because of technological improvements adopted such as electronic methods and informatics systems, there is the need to store and manage a great amount of data regarding the health and medical records from patients. This data should be accessible for clinical use in the medical environment [18].

Having health informatics systems implemented, the medical records are changed from paper to an electronic format, called Electronic Health Records (EHR). EHR intent to have data from multiple clinical episodes and providers, and become a lifelong medical record [19]. These documents allow for more efficient data management, and there is more immediate access to useful information that can be used as clinical decision support. The information should be structured according to medical standards, and can be constantly updated and can be consulted simultaneously by various health professionals [20]. It is easier to integrate the growing number of medical records because these are saved in a digital format, and several documents types are included, from laboratory tests and medical images to appointment records [21]. Besides, as the use of EHR and health informatics systems increase and information is sent and received more rapidly, public health agencies are able to process health information almost in real time [22]. These advantages lead to improved quality in the medical services, with a reduction in medical errors, and makes the patient more engaged in its healthcare decisions [23].

Over the years, the use of EHR by hospitals and physicians has grown. As of 2017, in the US, 85.9% of office-based physicians used an EHR system in their activities[10]. In Portugal, according to the Portuguese National Statistics Institute, there is an increase in computerised medical activities over the years. From 2004 to 2014, the hospitals that use EHR increased from 42% to 83% [11]. Figure 2.1 notes the proportion of computerised medical activities in 2012 and 2014. In 2014, more than 90% of the healthcare facilities have also computerised their hospital management systems and administrative activities, as well as patient databases and appointment scheduling systems [11].

Although EHR are widely implemented today, their implementation has been slow due to a number of factors. For health informatics systems to be put into practice, great initial investments are required and there are also maintenance costs to be taken into account [24]. The migration of old medical records that were previously made on paper is also a task that has to be considered in EHR functioning delays. Information extraction tasks are hindered when EHR are obtained as unstructured documents. Interoperability between systems is one of the main challenges in health informatics systems and needs to be improved before it can be put into operation, in

[10]"Electronic Medical Records/Electronic Health Records (EMRs/EHRs)": `https://www.cdc.gov/nchs/fastats/electronic-medical-records.htm`.Accessed on 21-04-2021

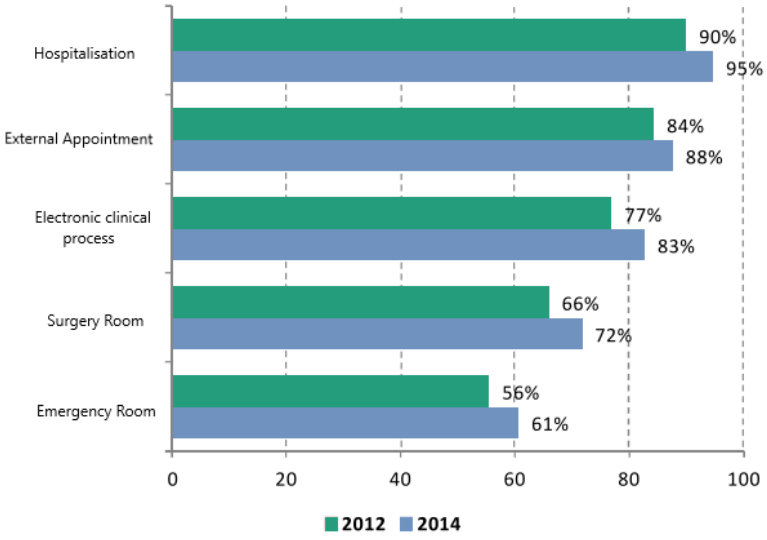


Figure 2.1: Comparison of proportion of hospitals by computerised medical activity in years 2012 and 2014, in Portugal. Adapted from [11].

regard to data duplication, heterogeneity of information systems implemented, among others. For that, social organisation and legal barriers have to be regarded, along with possible privacy issues [25].

2.3 Natural Language Processing

NLP is a field of study related to AI that aims for computers to process natural or human language. This processing is possible since human language is presented as input and converted into structured data [26]. NLP comprises several mechanisms and tasks, with different degrees of complexity.

There are some differences in natural language and computer language that lead to challenges in the processing of the information. One of these challenges is ambiguity. In natural language, there is often syntactic ambiguity in the language, as some words can belong to different parts of speech. Variability exists in various levels of language and is another challenge to consider [27]. Examples of circumstances in which variability occur are different spellings when comparing British and American English, analysis of word parts, and use of synonyms. Taking medical text into account, synonymy is a significant concept to have into account, since there are various definitions to biomedical terms. Abbreviations and acronyms also have to be considered as synonyms of a certain medical expression. Besides, in the medical context, it is common to find text that would be ungrammatical when spoken. The topics described are the main challenges to exceed by NLP for interaction between natural language and linguistic models to be possible.

2.3.1 Natural Language Processing Tasks

The NLP pipeline for natural language to be structured and understood as computational language can be adapted to the problem at hands. An illustrative example is represented in Figure 2.2. There are several NLP tasks that allow the conversion of natural language into machine interpretable data. Those tasks are further described in the following subsections.

Figure 2.2: NLP pipeline illustration.

2.3.1.1 Document and Sentence Segmentation

The first step on an NLP pipeline is the document segmentation for it to be examined into sections. This division must be reasonable in the context of the problem. The similarity between sections can be measured, so parts of the document that mention the same topic are grouped [28]. The following task is the division of the information in sentences, which is frequently considered the standard unit of analysis in NLP methods. This task takes place using models that separate the sentences in the presence of uppercase letters and punctuation [29].

In clinical notes, there are challenges associated with the tasks mentioned. The structure of the document is not standardised and can vary when comparing different hospitals and also health professionals. Besides, the notes can be ungrammatical and not be correctly punctuated, as well as not starting with an uppercase letter, raising a challenge to these techniques in this specific application [30].

2.3.1.2 Tokenisation

Tokenisation is the process of dividing a sentence into smaller elementary units called tokens. Tokens can be words, sub-words or characters, and can also include punctuation [29]. In specific cases, the dependency of words and specific domain knowledge have to be taken into account, so groups of words such as the name of a disease are not split [31]. Also, the language in which the text is written impacts the tokenisation. Besides orthographic changes, in oriental idioms, the words are not separated by spaces, so the boundaries are not as trivial as in other languages [32]. Tokenisation is possible by the implementation of either an automatic learning or a rule-based approach, in which language and grammar rules, as well as domain-specific knowledge, are required. Depending on the context, it can be advantageous to remove words that do not add information, such as stop words *"the"*, *"a"* or *"an"*.

2.3.1.3 Part of Speech

Part of Speech (POS) tagging is the task that associates each word with the corresponding category of words with similar grammatical characteristics, i.e., the POS. This task follows tokenisation in an NLP pipeline and, therefore, the POS is determined to each token [30]. There is a sequential lexical analysis that depends on the relationship with adjacent words in the sentence. POS tags are useful in many contexts, such as the medical field, because different POS can be related with different types of information [32]. It is known that adjectives represent a subjective interpretation or opinion, whereas nouns are the targets of an opinion.

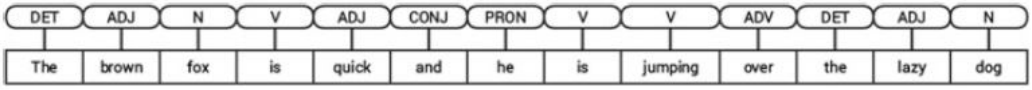


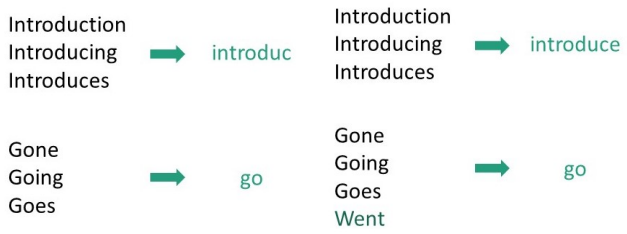
Figure 2.3: Word annotation in example sentence using POS Tagging. Adapted from [33].

It is possible to analyse the word annotation process of an example sentence in Figure 2.3. While there are some POS considered closed classes, such as the pronouns, determinants and conjunctions, in which there is a finite number of words and no additions are considered, others are open classes, admitting the addition of new words [33].

2.3.1.4 Stemming and Lemmatisation

In order for data to be processed easily, data normalisation is a task to consider. Stemming is a method that aims to reduce different forms of a word to its root form. The reduction of morphological forms to their base forms resides in the removal of affixes - suffixes and prefixes - from the index terms. The root form called the stem is the part of the word that is common to all the morphological forms [34]. Lemmatisation is another method that can be used in word normalisation scenarios. The principle of lemmatisation is similar to stemming in terms of word forms reduction. However, in this case, the context of the word in the sentence and the respective POS play a major role in finding the base form of the word, called the lemma [35], [36].

Figure 2.4 represents two stemming and lemmatisation examples. In the first example, comparing Figures 2.4a and 2.4b, it is possible to conclude that while stemming only considers the common part to the three words listed, lemmatisation finds the base word. In the second example, the stem and lemma found are the same. However, in lemmatisation, the word *"went"* is considered in the process, and the lemma is found since the base word is *"go"*. If the same word was used in the stemming process, a stem would not be found, since there would not be a common morphological form to the four words.



(a) Stemming. **(b)** Lemmatisation.

Figure 2.4: Comparison between Stemming and Lemmatisation methods. Word examples found in [34].

2.3.1.5 Parsing

Parsing is the task that performs a syntactic analysis of the text. The result of this analysis is a tree-like grammatical structure of a sentence, in which the connections between various words are described [32]. It is possible to distinguish two different approaches to this task: shallow parsing and full parsing. The first approach does not analyze the entire syntactic structure of a sentence, only focusing on a "head" word of a certain POS and finding groups of words that go together. On the other hand, full parsing obtains the full syntactic structure of the sentence, identifying both the groups of words and the relationship between them [30]. Full parsing can be further divided into dependency and constituency parsing, that focus on relationships between individual words and analyze the phrasal relationships, respectively. The dependency parsing of an example sentence can be observed in Figure 2.5.

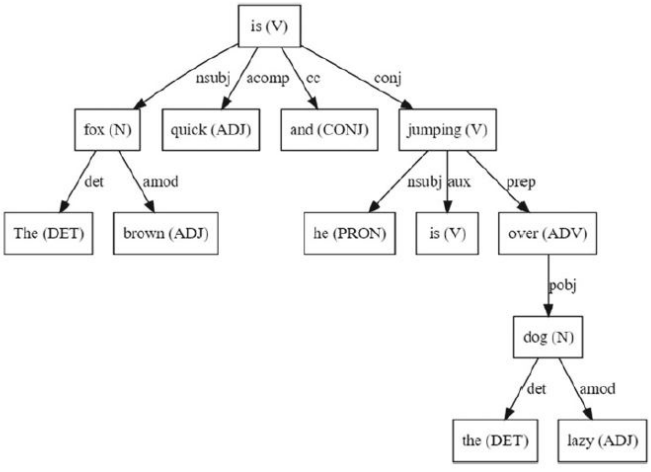


Figure 2.5: Dependency Parsing implementation in example sentence. Adapted from [33].

2.3.2 Natural Language Processing for EHR Information Extraction

The increasing use of EHR in clinical environments over the years creates the need of data interpretation and knowledge extraction. Considering that the input data in this case is clinical text, there are several possible outcomes when applying NLP to clinical notes [37], further described:

- **Representation Learning:** expansion of discrete codes (such as International Classification of Diseases (ICD) codes, further described in Section 2.4.2) into vector spaces for more detailed analysis and predictive tasks. These strategies are mainly unsupervised, clustering the represented codes into a vector space [38], [39]. The vector representations are derived so that similar concepts are close in a lower-dimension vector space.
- **Outcome Prediction:** prediction of patient outcomes. The outcome predictions can be static, when the prediction of a certain outcome does not consider temporal information [40], [41], or temporal, when the outcomes obtained are predicted to a fixed time span, or if the prediction is based on time series data [42].
- **Phenotyping:** obtainment of data-driven description of certain conditions. The two main applications are the discovery of new phenotypes and the boundary improvement in existing phenotypes.
- **De-Identification:** removal of personal information from the patient from the EHR, such as name, identification numbers, hospital names and locations. This task is mandatory for EHR to be publicly available.
- **Information Extraction:** information extraction from detailed patient documentation acquired in an unstructured manner. The outcomes include extraction of structured medical concepts from the clinical records [43], [44], clinical events and the corresponding time period [45], relational extraction between different concepts [39], and abbreviation expansion [46]. In information extraction problems, it is fundamental that the utilised unstructured corpus is preprocessed. The preprocessing steps to consider are the ones mentioned in Section 2.3.1.

The tasks included in clinical information extraction are represented in Figure 2.6.

2.3.2.1 Named Entity Recognition

Named Entity Recognition (NER) is a subtask of information extraction. It can be defined as the location and characterisation of nouns and proper nouns in text. A Named Entity (NE) is the categorisation of a token or sequence of tokens into topics. Those topics include essential information that are fundamental for several NLP applications. In systems that use general domain text, the most common NEs are Person (PER), Location (LOC), Organisation (ORG), DateTime (DT) and Miscelaneous (MISC) [47], [48]. In case the text is representative of a specific domain, the entities to be extracted from the text should be relevant for the context. For example, when considering clinical text, the classes must be adjusted to identify common information in EHR, such as the condition or disease, the exams performed for diagnostic, the provided therapeutic, among others [49].

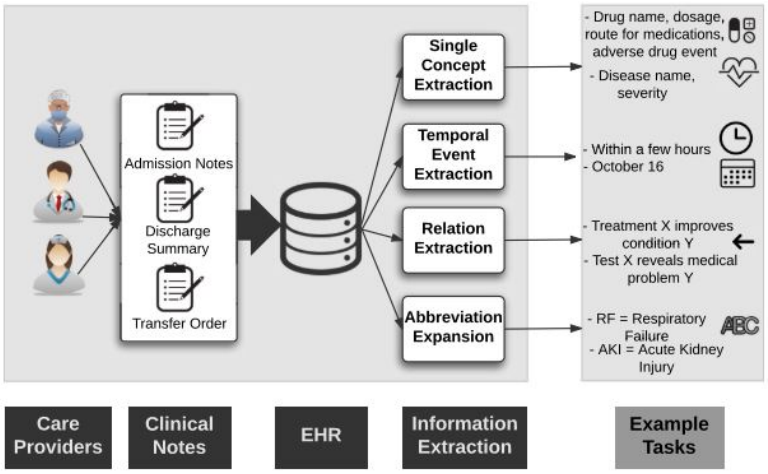


Figure 2.6: EHR Information Extraction and example tasks. Adapted from [37].

When assigning an entity to a certain token, using features not only from that word but from surrounding tokens is favorable for the task. Sequential tagging is used so that the model is able to know if a token is inside a certain NE. Inside-Outside-Beginning (IOB) tagging [50] is a common approach for sequence tagging. Besides the entity tagging, a beginning tag (B), inside tag (I) or outside tag (O) are assigned, depending on whether the entity begins in that token, the token is inside an entity (preceded by a beginning tag), or the token does not belong to any NE.

Figure 2.7 illustrates the IOB tagging of a general domain sentence, in English, with the entities previously mentioned.

Tokens	The	Symbiosis	Institute	was	established	in	2008	by	S	B	Mujumdar ...
IOB-Tags	O	B-ORG	I-ORG	O	O	O	B-DT	O	B-PER	I-PER	I-PER ...

Figure 2.7: IOB tagging example for the general domain sentence "The Symbiosis Institute was established in 2008 by S B Mujumdar...". Adapted from [48].

2.4 Medical Knowledge Sources

When representing a disease or clinical outcome, the information to perform said representation is extracted from clinical notes. A disease can be depicted by hard-code or by probabilistic rules found on the clinical records used, with the primary challenge being the selection of representative features for that disease or outcome to be detected [51]. It is possible to collect medical concepts related to the desired detected disease by using publicly available knowledge sources.

There are different categories for linguistic knowledge unstructured text treatment. A syntactic approach using lexicons and vocabulary information is the elementary linguistic knowledge using in NLP applications. However, the concepts must be organised in a hierarchical way. Thus, a typology of concepts or ontology is implemented. The models should also be able to provide semantic information, using a set of appropriate semantic rules for the NLP tasks

that return information about near words in sentences. Lastly, conceptual frames associates the different parts of the sentence [52].

There are several knowledge sources with different applications that can be used in clinical NLP approaches for feature extraction. The most widely used knowledge sources in literature are the Unified Medical Language System (UMLS) and ICD, further analysed in Sections 2.4.1 and 2.4.2, respectively.

2.4.1 Unified Medical Language System

Standardisation of clinical communication, as well as knowledge modelling, are fundamental aspects for clinical information extraction. To solve that problem, UMLS was introduced in 1990 [53]. The UMLS Knowledge Sources were established to assist biomedical systems development in information extraction from clinical text, independently of the differences in medical vocabulary and coding systems used [54]. UMLS comprises medical terminology and the respective translation, in anatomical, clinical and oncological fields [55].

Several knowledge sources are included in UMLS, such as the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon and Lexical Tools. Metathesaurus is the elementary and most recurrently used in NLP problems, and it organises the words by concept or meaning in and ontological approach. It includes approximately four million medical terminology concepts and fifteen million concept names, and comprises vocabulary in several languages, such as English, Spanish, Portuguese, German, among others. Over the years, UMLS has been used in the development, test and comparison of several NLP methods for clinical applications.

2.4.2 International Classification of Diseases

World Health Organization (WHO) established the ICD coding system aiming to globally standardize the representations of diseases, injuries, symptoms and other conditions that can be found in medical records. ICD is currently in its 11th version (ICD-11), which contains around 70 thousand medical codes composed by numbers, letters, or a combination of both, and comes into effect in 2022 [56]. ICD codes are a source of linguistic knowledge since the codes are representative of international medical procedures [52]. The standardisation attained is useful in health information management systems, health trends monitoring and medical billing and reimbursement facilitation.

Given that ICD codes contain diagnosis information and, therefore, if the patient has certain condition, this information can be used in the development of predictive models.

2.5 Machine Learning

A traditional way of approaching computer science tasks is to design an algorithm that uses input data and turns it into an output. However, in certain situations, there may be a large amount of data that difficult the knowledge-based task, and no specific algorithm can be designed. Since the algorithm that converts inputs into outputs is not known, the machine should be able to generate its own algorithm automatically.

ML is an area of AI in which computers automatically learn from the available data and perform alterations in the algorithm that are not explicitly programmed, with the aim of improving the output prediction. The algorithms are capable of extracting knowledge from data by recognising patterns in the data [57]. The trained models are able to predict new and unseen data and are efficient in dealing with a large amount of data.

ML approaches can be divided into supervised and unsupervised learning. Supervised learning algorithms make use of labeled data in order to generalise knowledge and be able to predict unseen unlabeled data [58].

Supervised ML problems can be divided into several fundamental phases for solving the problem [59], as presented in Figure 2.8 and further described:

- **Problem Definition:** it is fundamental to determine what the final goal and the knowledge to extract from the model, as well as to understand the problem domain in-depth.
- **Data Collection:** since this phase is specific for each application, the dataset collection may need the help of an expert in the area. It may require the use of specialised hardware or software tools.
- **Data Pre-Processing:** to facilitate the algorithm implementation, the data should be in a specific format that enables its processing. It should be uniformised, in case it was extracted from multiple sources. Cleaning of the data is also important, in case there are missing entries or outliers that should be estimated or discarded, respectively.
- **Feature Extraction and Selection:** the extraction of features is a task that requires knowledge about the application domain. Feature Selection extracts meaningful features and removes redundant features after the pre-processing phase, by lowering the dimension of the original feature space into a new feature space that will be further used. Feature Extraction allows the construction of new features using the initial feature set.
- **Classification Algorithms Implementation:** in this task, ML algorithms are implemented in order to extract patterns within the data and evaluate the output obtained. The ML algorithm to use can be adjusted depending on the data, as well as hyper-parameters values. Both training and testing phases are included in this task.
- **Results Interpretation:** after obtaining the results from the models implemented, these must be interpreted in the context of the problem. If the results obtained are satisfactory, the knowledge should be included in a software tool to be used in the practical context.

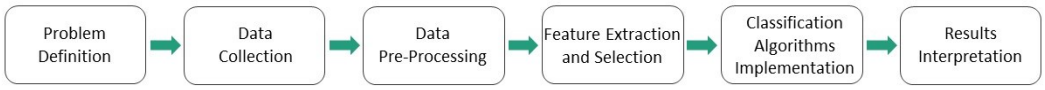


Figure 2.8: Typical ML problem pipeline.

On the other hand, unsupervised learning utilises unlabeled data and is able to group data with similar information into clusters by means of automated methods [58]. Unsupervised learning can be used as a previous step to supervised learning [60]: the clustering of data by unsupervised algorithms detects groups within the data that were not previously detected, and that can help assign a label to each data value.

Considering NLP tasks specifically, unsupervised learning approaches have the advantage of being able to utilise unlabeled data, which is more abundant than labeled data, and avoid the time-consuming task of labeling the corpus. However, the use of unlabeled corpora makes the implementation of unsupervised approaches more challenging, mainly in terms of how to evaluate the results [61]. Since there are no ground truth labels, it is impossible to apply the standard evaluation metrics further described in Section 2.7.

Both shallow ML and DL approaches can often be perceived as a black box, since there may be little insight on some aspects of the model learning process. In applications such as healthcare, it is important that humans are capable of understanding how the model performs a prediction [62]. In recent years, the legal requirements regarding AI applications and their explainability have increased [63]. The requirements are more demanding when the model prediction would be considered in a decision-making process without human supervision.

2.6 Deep Learning Approaches

As mentioned in the previous section, in order to construct a ML model, there is the need to have expertise in the application domain in order to extract relevant features and convert the raw data into useful information in which the learning system could identify a pattern. ML approaches can be divided into shallow learning and DL approaches. While shallow learning approaches utilise manually engineered features obtained by pre-processing of the input data to obtain a meaningful data representation, as mentioned in Figure 2.8 in the previous section, DL approaches target the automatic creation of a feature representation using the raw input data [64]. DL architectures are composed of artificial neural networks based on the processes that occur in the human brain [65]. Figure 2.9 represents the hierarchical relation between AI, ML and DL approaches.

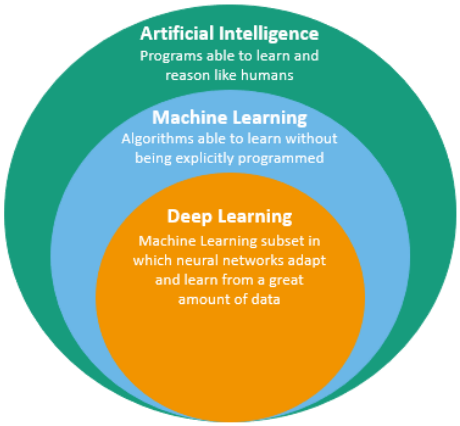


Figure 2.9: AI Venn Diagram with ML and DL Hierarchy.

Representation learning is an approach in which the machine automatically finds the best algorithm for classification, using the input data. In many applications, there are issues associated with this approach due to variation in the data, that leads to a difficulty in obtaining significant features. DL goes beyond previous limitations in representation learning, by making use of non-linear representations that, in each level, modify the representation to a higher, more abstract level [66].

The feature extraction step can be avoided in DL strategies by applying a network architecture called multilayer perceptron [67]. This architecture is based on the stacking of simple modules. This network is equivalent to the mapping of input values into output values, associated with functions. The general mathematical function of the model can be seen as a combination of several functions.

Many DL applications use Feedforward Neural Networks. This type of architecture maps a fixed-size input into a fixed-size output, computing the weighted sum of the inputs from the previous layer and passing the result through non-linear functions to move from a layer to the next. These neural networks often have hidden units that are not in the input or output. Therefore, hidden units are located in hidden layers that alter the input non-linearly, so that it is possible to linearly separate categories in the last layer [68]. An illustrative neural network architecture is shown in Figure 2.10, with four input units, one hidden layer with three hidden units, and two output units.

According to Wu *et al.* [69], the most applied DL methods in NLP applications nowadays are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) , specifically long short-term memory (LSTM) networks, as well as attention-based Transformers such as Bidirectional Encoder Representations from Transformers (BERT), that grew on popularity over recent years.

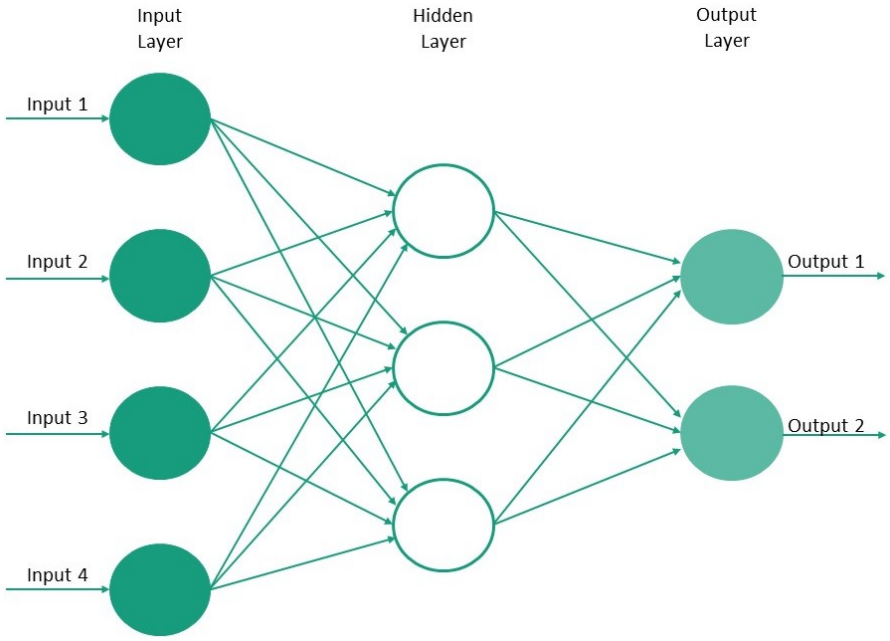


Figure 2.10: Neural Network Architecture Illustration.

2.6.1 Convolutional Neural Networks

CNN are multilayer artificial neural networks composed of three types of layers [70]. The convolutional layers learn feature representations from the inputs, computing feature maps with the use of kernels. The pooling layer aims to reduce the feature map resolution by merging similar features detected by the convolutional layers, using pooling operations such as average pooling and max pooling. Fully-connected layers generate global semantic information. This last layer is not essential and can be replaced by a 1×1 convolution layer [71]. CNN was first implemented in computer vision and later used in language processing tasks due to a high-level feature extraction need.

2.6.2 Recurrent Neural Networks

RNN are based on the sequential processing of the information [72]. The same computation is done consecutively in RNN architectures, and each step depends on previous results saved in memory. When applied to language processing tasks, RNN is able to capture the sequence that is innate to language because of the sequencial processing done by modeling units in sequence. Besides, RNN are able to model text with varied length, from long sentences to paragraphs or even entire documents [68]. Some applications in which RNN is well adapted are language modelling, machine translation and speech recognition.

Figure 2.11 represents a general RNN architecture. At time t, the artificial neurons represented by s_t receive as inputs the output from previous neurons. That way, the output o_t depends not only of the input sequence represented by x_t, but also on previous inputs, such as x_{t-1} in the figure. Matrices U, V and W represent the parameters used, that are the same in all steps [71]. At the time conventional RNN were designed, the main goal was to learn dependencies in the

long term. However, when tested, it was proven that RNN did not store the information learned for a long time. This problem is denominated vanishing gradient [73], and happens when the gradient value becomes negligible and therefore the weights of following layers do not change.

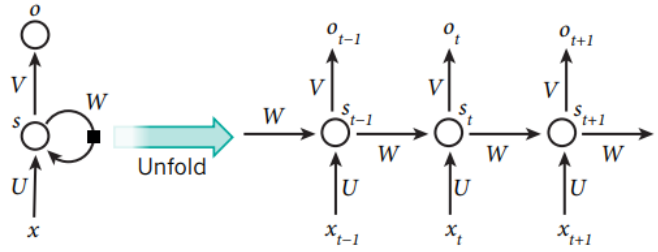


Figure 2.11: General RNN architecture. Adapted from [68].

2.6.2.1 Long Short-Term Memory Networks

In order to overcome the vanishing gradient problem [73], LSTM networks were designed [74]. LSTM networks have a hidden special unit called memory cell, C, that operates similarly to an accumulator, that accumulates the value of a certain signal, but its memory can be cleared by the decision of another cell called forget gate, f [68]. The memory saving or erasing depends on the output of the forget gate: if the output is a vector of zeros, the multiplication with the previous cell state will be zero and, therefore, the memory is erased; if the output is a vector of integer values greater than zero, the old information flows to the cell and is kept. The nearer the integer numbers are from one, the greater information quantity is passed to the cell. The control of information used in the input state, x, and hidden state, h, is done by the output state, o. In order to maintain the output of each gate between zero and one, the sigmoid function σ is used [71]. Figure 2.12 ilustrates the LSTM unit[11]. and the transformations that happen in the cell, also represented in Equations 2.1 to 2.5.

$$f_t = \sigma(W_f \times x + b_f) \tag{2.1}$$

$$i_t = \sigma(i_f \times x + b_i) \tag{2.2}$$

$$o_t = \sigma(W_o \times x + b_o) \tag{2.3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c \times x + b_c) \tag{2.4}$$

$$h_t = o_t \odot tanh(c_t) \tag{2.5}$$

[11] "Understanding LSTM Networks": `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`.Accessed on 17-06-2021

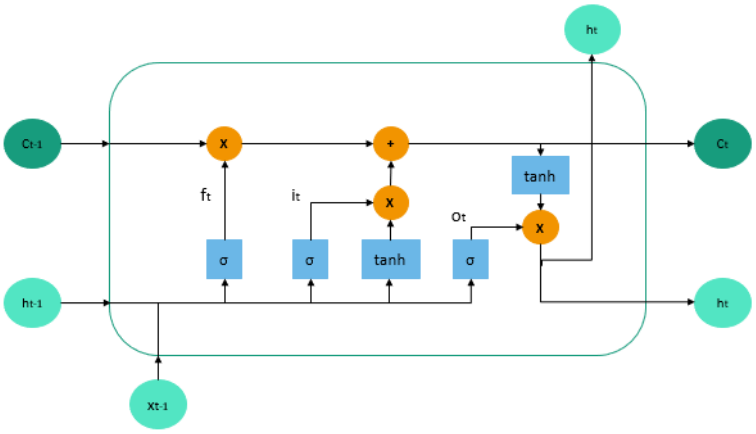


Figure 2.12: LSTM unit illustration. Adapted from[11].

2.6.3 Word Embeddings

In more recent neural network applications, the input layer to the network are word embeddings. Embeddings consist on a dense and data-driven vectorised word or concept representations in a lower dimension space [39], [69]. The vector representation is able to maintain relationships between words. Since a large data set annotated with syntactic and semantic relationships is hardly attainable, unsupervised methods are mostly used to obtain word vectors.

Word vectors are based on a distributional hypothesis. The vectors detect similarities between neighbour words, since words with similar meanings are more likely to occur in similar contexts [71]. The similarity between concepts is obtained by calculating the cosine between vectors, that varies between zero and one, depending on the degree of similarity. There are several word embedding algorithms, such as word2vec [75], GloVe [76], and FastText [77], [78].

Word2vec is a widely known algorithm for word vectorisation. This model converts a large text corpus into word vectors using a neural network architecture, and can be trained using two different approaches, represented in Figure 2.13. Whereas Continuous Bag-of-Words predicts the current word considering the context (i.e., the surrounding words), Skip-gram predicts the nearest words considering the current [75]. The two key paramethers to have into account when training a word2vec algrithm are the dimension of the embedding and the length of the context window, i.e. the words to be considered before and after the target word. The word2vec main disadvantage is the fact that the word vectors are static despite the context. Homonyms and homograph words will have the same representation, regardless of the sentence context. Besides, although it is possible to use pre-trained lexicons by using word2vec in large datasets, the results are considered generic and its application may not be meaningful in certain domains.

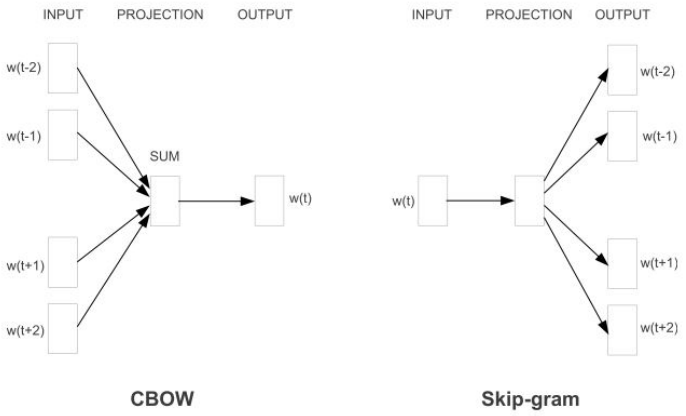


Figure 2.13: Word2vec model architectures. Adapted from [75].

GloVe is another algorithm for word embedding learning, but the main difference to the word2vec algorithm is that, instead of word prediction, this algorithm computes a term co-occurence matrix, with dimension of V$\times$V, with V being the vocabulary size. Each matrix entry represets the number of times certain words from the vocabulary occur together in a certain context window [76]. The context window usually has larger dimensions than word2vec, which makes it possible to detect long dependencies, despite the dependency order being lost. Although the resulting vectors are similar when comparing word2vec and GloVe, the first utilises a predict-base model, whereas the second utilises a count-based model [79].

A limitation of word2vec and GloVe algorithms is that only words from the vocabulary can be handled. FastText is another algorithm for word vector representation that, instead of using words for the learning process, it uses characters n-grams - sequences of adjacent characters. Each word is represented by a sum of the vector representation for each n-gram. By obtaining subword information, it is possible to handle out-of-vocabulary words, adapting the word2vec Skip-gram model. Besides the vector construction for out-of-vocabulary words, the subword components used to compute the vector representation also allow for the representation of morphology and lexical similarities between words [77]–[79].

2.6.4 Transformers

The Transformer architecture was first proposed by Vaswani *et al.* in 2017. It is a simpler architecture than RNN and CNN architectures, based solely on self-attention mechanisms that enable dependency modeling and drawing long-range dependencies between input and output [80]. In self-attention mechanisms, several positions of a sequence are associated so that a representation of that sequence is computed.

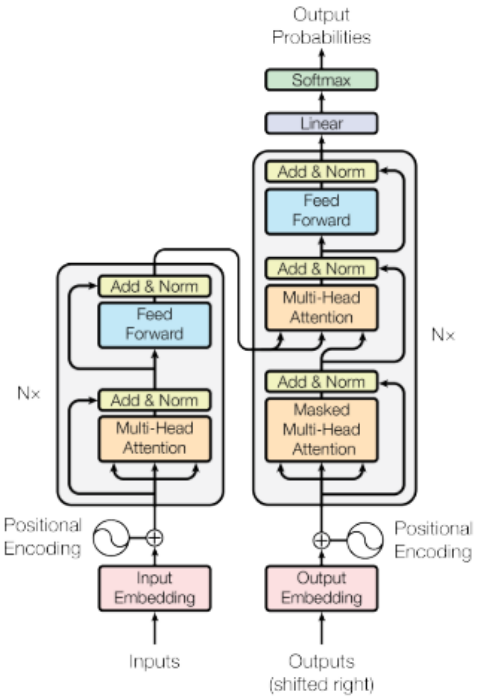


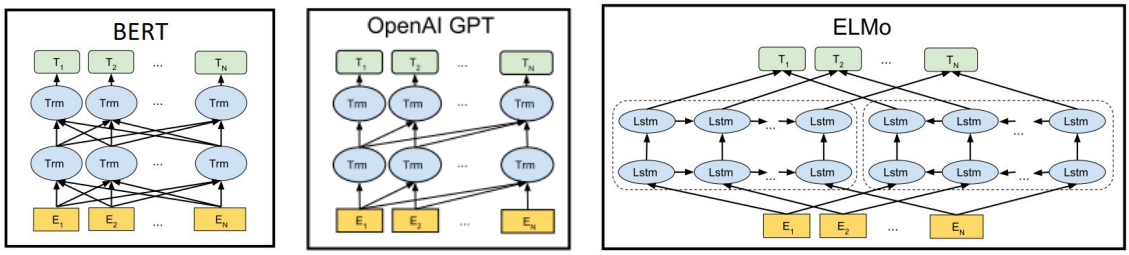
Figure 2.14: Transformer model architecture. Adapted from [80].

The Transformer model architecture is represented in Figure 2.14. It is composed by Encoder and Decoder stacks, with several modules in sequence multiple times, as represented by the *Nx* in the figure. The Encoder is the structure on the left, whereas the Decoder is on the right. Instead of using strings, the input and output are word embeddings, a representation of those strings in a vector space with a certain dimension. Each word embedding also saves information about the relative position of the word in the sequence, replacing use of RNN, that is able to remember the sequences in which words are fed to the model. The linear transformation followed by a softmax function convert the output of the decoder module into a next-token probability prediction.

In recent years, the Transformer architecture has become the state-of-the-art architecture for NLP tasks, due to computational efficiency as well as promising results that outperform other architectures [81]. Transformers have been used in recent neural networks for NLP applications, such as BERT and OpenAI Generative Pre-trained Transformer (GPT) [82], [83]. While in BERT a bidirectional Transformer is implemented, the OpenAI GPT model implements a left-to-right Transformer. Both BERT and OpenAI GPT are fine-tuning approaches: a pre-trained model for a given task is used in a similar task, after being optimised for that purpose. The model architectures can be observed in Figures 2.15a and 2.15b respectively, in comparison with the ELMo (Embeddings from Language Model) architecture in Figure 2.15c.

Contrary to the previously mentioned architectures, ELMo employs a feature-based approach, in which the word representations are computed using a two-layer bidirectional language model, that combines forward and backward LSTM [84]. Hence, the vector that is allocated to each token is dependent of the information contained in the entire input sentence, and the word embeddings that result from the pre-trained neural network are used as features in

other NLP models.



(a) BERT model architecture.

(b) GPT model architecture.

(c) ELMo model architecture.

Figure 2.15: Comparison between BERT, GPT and ELMo model architectures. Adapted from [82].

2.6.4.1 BERT

In 2019, researchers from Google AI Language introduced the BERT architecture [82]. BERT framework can be divided into two steps: the pre-training and the fine-tuning phases. The pre-training phase uses unlabeled data as input for different tasks. On the fine-tuning phase, the model is initialised with the parameters obtained in the pre-training, that will be subsequently fine-tuned to a specific downstream task, using labeled data. Each downstream task has separate models that originate from the pre-trained model.

As previously stated, BERT uses a bidirectional training of a Transformer as a masked language model that is able to predict random masked words from the input. Therefore, the architecture can be used in learning bidirectional representations based in context. Since the bidirectional mask language model is used, there is not a preferred direction to analyse the sentence, and the word context is assessed both to the left and the right of the word.

BERT pre-training not only consists in the masked token prediction, but it also performs a next sentence prediction, using sentence pairs for the pre-training. That way, the model not only understands the context of the word but also the context of the sentences in the complete document.

2.7 Evaluation Metrics

Evaluation metrics must be used in order to assess the performance of a model, as well as the progress obtained in its implementation.

To evaluate the models, it is necessary to divide the data into training and test sets. Cross-validation is a data resampling method that evaluates the generalisation ability of a predictive model. The data is divided into smaller groups designated as folds. The K in K-Folds cross-validation represents the number of folds, and is defined when implementing the model [85]. In Figure 2.16, a K-Folds cross-validation with K=10, meaning that the data is divided into 10 folds, that for their part are divided into 90% of training data and 10% for validation or test.

Regarding the evaluation metrics, the most commonly used in ML applications, including NER, are precision, recall and F1-Score. Precision measures the fraction of instances classified

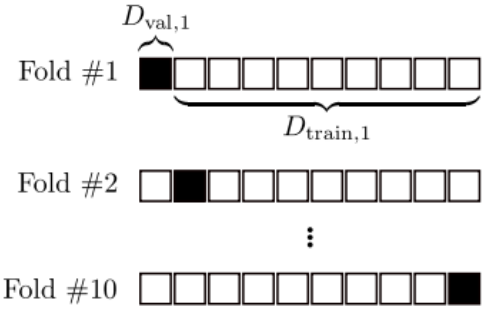


Figure 2.16: K-Fold Cross-Validation illustration with K=10. Adapted from [85].

as positive that are truly positive. On the other hand, recall measures the fraction of positive instances that are accurately labeled. The mathematical expressions for precision and recall are represented in Equations 2.6 and 2.7, respectively. F1-Score is a combination of precision and recall, defined as an harmonic mean of the previously mentioned evaluation metrics. F1-Score ranges from 0 to 1. A F1-Score closer to 1 corresponds to a prediction with a small number of false positives and false negatives, whereas a F1-Score closer to 0 means that there is a large amount of incorrect predictions. Equation 2.8 represents the expression to compute the F1-Score.

$$Precision = \frac{TruePositive(TP)}{TruePositive(TP) + FalsePositive(FP)} \tag{2.6}$$

$$Recall = \frac{TruePositive(TP)}{TruePositive(TP) + FalseNegative(FN)} \tag{2.7}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.8}$$

When considering a NER problem, there are several classes in the dataset. Therefore, each class will have their respective evaluation metrics values to assess the model prediction for that specific class. To assess the model performance for all classes, an average of all class performances should be computed. The average can be computed using different approaches. Micro average (Equations 2.9 and 2.10) considers the total amount of true positives, true negatives, false positives and false negatives regardless of the class to compute the average. Macro average (Equations 2.11 and 2.12) performs an average of the evaluation metrics for each class, without considering the proportion or weight of each class in the dataset. In contrast, weighted average (Equations 2.13 and 2.14) considers the proportion of each class in the dataset to calculate the average.

$$Micro - Precision = \frac{\sum_{i=1}^{numberofclasses} TP_i}{\sum_{i=1}^{numberofclasses} TP_i + FP_i} \tag{2.9}$$

$$Micro - Recall = \frac{\sum_{i=1}^{numberofclasses} TP_i}{\sum_{i=1}^{numberofclasses} TP_i + FN_i} \tag{2.10}$$

$$Macro - Precision = \frac{\sum_{i=1}^{numberofclasses} Precision_i}{numberofclasses} \tag{2.11}$$

$$Macro - Recall = \frac{\sum_{i=1}^{numberofclasses} Recall_i}{numberofclasses} \tag{2.12}$$

$$Weighted - Precision = \sum_{i=1}^{numberofclasses} \frac{TP_i}{TP_i + FP_i} \times \frac{TP_i + FN_i}{TotalSamples} \tag{2.13}$$

$$Weighted - Recall = \sum_{i=1}^{numberofclasses} \frac{TP_i}{TP_i + FN_i} \times \frac{TP_i + FN_i}{TotalSamples} \tag{2.14}$$

2.8 Summary

In the course of this chapter, skin cancer was introduced, namely its types, information about its global incidence, as well as how skin cancer is diagnosed. Subsequently, health informatics systems and the use of EHR were presented. EHR are an important part of the work to be performed, given that the clinical data used in the dissertation worrk is extracted from health records.

Afterwards, NLP was introduced. The fundamental tasks involved in NLP were presented, as well as the general problems considered when extracting information from electronic health records. Some core concepts regarding NER were also introduced. Knowledge sources have a collection of medical concepts available and are important for disease and condition detection tasks and extract linguistic knowledge in unstructured clinical notes. UMLS contains medical concepts in different languages, and ICD codes include diseases and conditions to standardise their representation.

Lastly, shallow ML and DL concepts were addressed, taking into account their significance in the context of the problem. The evaluation metrics that are used to evaluate ML and DL problems similar to the developed work were also explained.

Chapter 3

Clinical NLP Literature Review

After the revision of the fundamental concepts related to the dissertation topic, this chapter introduces the state-of-the-art methods in clinical NLP applications.

Nowadays, NLP applications are used for several healthcare-related tasks. The extraction of information from the medical text presented in EHR is essential to access clinical information to be used in these applications. There are several approaches in the NLP field: rule-based or knowledge-based approach, and statistical approaches, that include shallow ML and DL algorithms. An algorithm can include different approaches in simultaneous, as hybrid approaches lead to promising results in clinical applications.

3.1 Clinical Knowledge Representation in Portuguese

In order to train models for clinical NLP tasks, it is required to have medical text in a considerable amount. However, the use of clinical texts raises issues in terms of patient personal data protection and is subject to strict ethical regulation. Besides, utilising clinical text in information extraction tasks, such as NER, imply that a subset of clinical texts is annotated with entities of interest for the clinical domain. The annotated texts can afterwards be used as data for training and testing the models for NER.

Regarding annotation of Portuguese clinical text, Ferreira *et al.* [86] developed a knowledge representation with 10 classes such as Anatomical Site, Condition and Evolution, four of those with subclasses, to allow a structured representation of patient discharge letters. The annotation guidelines were made available, even though the dataset is not available. A total of 30 annotated documents were also annotated by 7 judges for validation, obtaining an F1-Score of 0.95. Lopes *et al.* [87], [88] performed a manual annotation of 280 Portuguese Neurology clinical cases, following the knowledge representation of [86], replacing the Location entity for Genetic and Additional Observations, to better adapt to Neurology cases. The Neurology dataset was made available on GitHub[12], and the annotation validation of 90 documents achieved agreement ratios between 0.86 and 1.00 to all entities. Oliveira *et al.* [89] developed SemClinBr - a corpus

[12] "Neurology dataset": `https://github.com/fabioacl/PortugueseClinicalNER`.Accessed on 31-03-2021

composed of 1 thousand brasilian clinical cases that includes several specialties and from different medical facilities. The annotation was performed using the UMLS semantic types[13] and done with a web-based annotator made from scratch for this task. The annotated corpus was not made available due to ethical regulation.

3.2 Rule-based Approaches in Clinical NLP

The first approach to consider is rule-based or knowledge-based. This method requires several sources of information about the domain. It is possible to apply information about language use, by creating linguistic pattern rules, and real-world knowledge about the topic.

Since it uses previous knowledge, rule-based approaches are manually constructed and iteratively adjusted for better text processing accuracy. Therefore, this approach requires manual effort to create the rules and knowledge bases used, which can be time-consuming. Besides, the adjustments made to obtain better results focus on the extension of previous rules, so there is no need for an extensive training set.

Chen *et al.* assessed patient eligibility in terms of selection criteria to participate in clinical trials using EHR. A rule-based clinical NLP system was developed, applying a bottom-up rule-based framework at lexical, syntactic and meta-level. Starting with the lexical level, a semantic group with relevant expressions was created for each condition, through evidence-level annotation analysis and a tailoring active learning process. Subsequently, the syntactic level rules were designed using POS tagging and parsing, and linking the relevant expressions with lab results, assertions or time mentions. Lastly, meta-level rules used document-level and patient-level information, such as the patient gender or the last timestamp on the records, in order to normalise time modules to the last record, or lab results to the correct gender [90]. Thirteen selection criteria were defined, and the rule-based NLP system must extract and evaluate annotation evidence of sub-criteria in order to give patient-level decisions on whether the patient meets all the selection criteria for the clinical trial. The rule-based algorithm was evaluated, obtaining a F1-score of 0.90.

3.3 Machine Learning Approaches in Clinical NLP

In ML or statistics-based approaches, computer models are trained using labeled or unlabeled input data, whether it performs supervised or unsupervised learning. The algorithms automatically learn how to extract knowledge from data by recognizing patterns in data. After the training phase, the models are able to make a prediction using unlabeled new data. The prediction classifies the data into classes, that should have relevant meaning in the context of the problem.

Overall, ML approaches are easier to scale and faster to develop, when compared to rule-based approaches. However, in order to have a good representation of all events that may happen, it is fundamental that a large amount of data is available.

[13] "UMLS current semantic types": `https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html`.Accessed on 25-05-2021

As previously stated in Section 2.6, ML approaches can be divided into shallow learning and DL approaches. Shallow learning approaches are considered more typical ML approaches, whilst DL approaches are more recent and related to neural networks, skipping the feature engineering task that is automatically done by the model.

3.3.1 Shallow Machine Learning

In recent years, ML approaches have been implemented in clinical NLP applications, using EHR to extract information that can facilitate and support clinical decisions. For instance, Feller *et al.* [91] developed an automated HIV risk assessment process using NLP. HIV screenings are expensive, time-consuming and can fail in high risk individuals diagnosis. The ML models used demographics, lab tests diagnosis codes and unstructured notes prior to the HIV diagnosis as input data, and three algorithms using the random forest architecture were developed. The baseline model only considered structured data. The second algorithm added NLP topic modelling to the baseline model. The authors implemented Latent Dirichlet allocation that divided the notes into 250 clusters that represent word distribution in the corpus. The third algorithm used the baseline model and an NLP method to extract clinical key words, using their respective term frequency-inverse document frequency (TF-IDF) [92] to correlate the words with higher HIV risk. From 300 words obtained, 37 were manually selected to be included in the model. Since the large amount of variables found can limit the algorithm accuracy, mutual information criteria were used to choose, for each model, the 150 variables with the highest mutual information when compared with the labels used (HIV positive or negative). The random forest predictive models obtained F-measures of 0.59, 0.63 and 0.74 for baseline, baseline and NLP topic, and baseline and NLP clinical key words, respectively. The use of clinical terms that indicated high risk of HIV diagnosis was able to improve the predictive model performance.

Similarly, Chase *et al.* [93] tried to identify signs and symptoms of multiple sclerosis in EHR before detection by health professionals, implementing a ML Naive-Bayes classifier. The gold standard for multiple sclerosis presence or absence was the respective presence or absence of an ICD-9 code for multiple sclerosis. The symptoms were mapped to UMLS terms, and it was possible to notice 1000 terms that occured more frequently in multiple sclerosis diagnosed patients than in the control group. These terms were manually clustered using synonymy criteria, so that 50 clusters are used as classification features. The ML model performance when using patients with diagnosed multiple sclerosis as input was measured using receiver operating characteristic (ROC) area under curve (AUC), sensitivity and specificity and the results were 0.94, 0.81 and 0.84, respectively. Using the clinical notes of patients with multiple sclerosis from two years prior to the diagnosis, the model prediction on whether these patients had multiple sclerosis based on those clinical notes had a performance of 0.71, 0.40 and 0.97 considering ROC AUC, sensitivity and specificity.

3.3.2 Deep Learning Approaches

The availability of clinical textual data increased as a consequence of the shift towards EHR use. Identifying hidden knowledge in data can improve the extracted information from clinical notes. In recent years, shallow ML and statistical techniques have been replaced by DL approaches [94].

Besides being less time-consuming in pre-processing and feature extraction tasks, DL techniques are advantageous in dealing with a large data volume and efficiently identifying data relationships that could be considered hidden or unknown [37].

Two applications built on the LSTM architecture were reviewed. DeepCare was able to extract previous illness histories from EHR, as well as infering current illness states and predicting future outcomes [95]. Considering diabetes and mental health as the two chronic conditions to analyse, predictions were done in terms of diagnostic of the condition and medical interventions. For diabetes, diagnostic and intervention predictions resulted in a precision of 0.66 and 0.78, whereas for mental health those values were 0.52 and 0.71, respectively. The predictions obtained with DeepCare achieved a better precision than RNN predictions. Xu *et al.*[96] introduced an approach for medical entity recognition using bidirectional LSTM (BiLSTM) and Conditional Random Fields (CRF), with three layers: a character-based BiLSTM layer, a word-based BiLSTM layer, and a CRF layer. The character-based word representation was applied to infer word and distributional sensitivity. Using a dataset of approximately 800 EHR, the results obtained with an F1-score of 0.80 for the BiLSTM-CRF approach demonstrate that this approach outperformed baseline approaches such as CRF, with a F1-Score of 0.77.

Regarding the BERT architecture, it was introduced by Devlin *et al.* [82]. BERT is based on a fine-tuning approach, as previously explained in Section 2.6.4.1. In clinical domain, there is a shift in word distribution when comparing medical corpora and general domain corpora. Lee *et al.* [97] developed BioBERT - Bidirectional Encoder Representations from Transformers for Biomedical Text Mining. The model architecture is similar to BERT, the major difference being the pre-training with biomedical text, using PubMed abstracts. BioBERT achieved a better performance than BERT in tasks such as NER and and relation extraction when using texts with medical concepts. Costa *et al.* [98] also implemented the BERT architecture to perform NER in clinical notes and automatic assignment of Spanish ICD-10 codes. BETO consists on the BERT DL architecture pre-trained in Spanish clinical text and fine-tuned on NER. Considering that there are two diferent types of ICD-10 codes, ICD10-CM (Clinical Modification, related to diseases and conditions) and ICD10-PCS (Procedure Coding System, related to sprecific procedures such as surgeries), different classifiers were developed exclusively to each code type. Afterwards, the individual predictions were combined to assess if there were relevant ICD-10 codes in a certain clinical note. The method obtained a F1-score of 0.51 for the NER task, and mean average precision of 0.51 and 0.45 for ICD10-CM and ICD10-PCS codes, respectively.

3.4 Clinical NLP Approaches in Languages other than English

The majority of clinical NLP applications utilises English clinical text as input, as assessed by Névéol *et al.* [99] when analysing the clinical NLP publications returned from language-specific queries in PubMed. As in 2017, the number of publications using clinical text in languages other than English made up only 10% of the total publications in clinical NLP. Having methodologies able to analyse the medical text in other languages is fundamental to obtaining access to valuable patient data acquired in countries where English is not the official language. The value of non-English clinical record information extraction increases when considering a rare disease since

there are few cases worldwide, and each clinical history is crucial to obtain more information about the condition.

The growth in clinical NLP publications in MEDLINE for the five most studied languages other than English during a decade is represented in Figure 3.1. It is observed that French has generated constant interest and growth, whilst more knowledge has been applied to Chinese and Spanish in recent years.

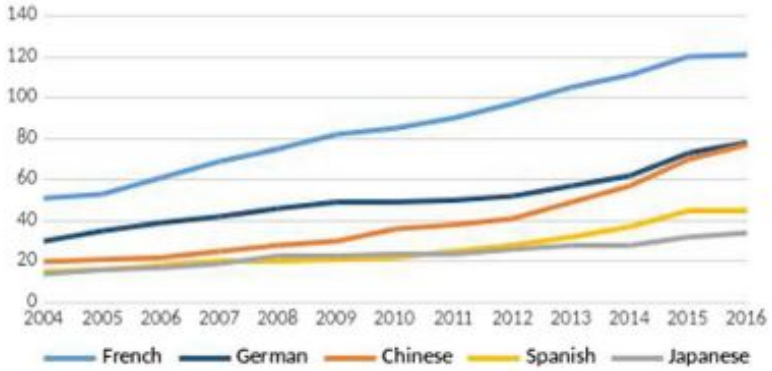


Figure 3.1: Clinical NLP publications growth in MEDLINE from 2006 to 2016, for the five most studied languages other than English. Addapted from [99].

The systems used in NLP applications in languages other than English can adapt NLP architectures previously used in English applications or build new NLP systems, considering the challenges in NLP tasks inherent to certain languages. Concerning new NLP systems or components adapted to each language, NLP tasks such as word segmentation are hindered if the language does not have clear word boundaries, for instance, Chinese or Japanese. A solution for the lack of spacing in these languages can be a probabilistic model for automatic word segmentation using dictionaries. Abbreviations are commonly used in medical text and can vary depending on the language. Term identification and normalisation strategies are important and have been implemented in languages such as Spanish, Swedish and German [100].

Depending on the NLP tasks performed, the language and the system design, adapting systems that work well for English to another language can have promising results. A modular NLP system can be adjusted more easily, as studied with cTAKES system [101]. Using Spanish corpus from annotated medical documents, the accuracy of NLP tasks such as sentence boundary detection, POS tagging and chunking was assessed, achieving a promising performance. Similarly to English applications, the strategies used in NER and information extraction tasks are rule-based, statistical or a hybrid solution. Krieger *et al.* [102] implemented a hybrid parsing strategy combining chunk parsing and deep parsing, as well as a hybrid relation extraction strategy, using both a rule-based system and a minimally-supervised ML system. The systems implemented used German clinical records as input and have delivered relatively satisfactory results. Becker *et al.* [55] performed information extraction and annotation using German clinical colorectal cancer notes, that included information about the tumour stage. The authors implemented a semi-supervised rule-based system that pre-processed the training data by pattern analysis. A set of terms related to the domain were collected and, afterwards, German UMLS terms were applied to the initial set to extract and add synonyms and common abbreviations

to the previous terms. Then, a NLP algorithm for NER was developed in order to extract the relevant UMLS terms that occur in the clinical notes analysed. Unsupervised learning methods can be used when there is a large data set available. Unsupervised methods were implemented by Moen *et al.* [103] to extract information of care episodes in Finnish clinical text. The input and hidden layers were the same as in the word2vec algorithm, and it was desired that the model output, i.e. prediction, was an ICD-10 code for each clinical record.

As in the previous application, it is sometimes possible to carry out unsupervised learning strategies. However, many clinical NLP applications depend on language-specific resources. For that reason, it is necessary to create synonym or abbreviation lexicons. There is an increased interest in expanding the terminologies and ontologies included in UMLS or SNOMED-CT for other languages. Perez *et al.* [104] departed from Spanish terms included in SNOMED-CT, implementing a semi-automatic translation algorithm from Spanish to Basque.

3.4.1 Clinical NLP Applications in Portuguese

According to a PubMed search similar to the one provided in [99], there are 33 publications from the last two decades returned when using the search query "NLP AND Portuguese". Adding "clinical" and "medical" to the search, the number of publications returned decreases to 14 and 16, respectively. Table 3.1 compares the results from similar searches obtained for other languages. When restricting the domain to clinical or medical text, the number of publications decreases in all languages. It is possible to observe that the number of publications for the English language is much higher than for the other languages. The position of the other languages regarding the number of publications in recent years supports the previous graph in Figure 3.1.

Table 3.1: Comparison of PubMed search results on NLP publications per language in general domain and clinical or medical domain.

Language	NLP AND language	NLP AND language AND clinical	NLP AND language AND medical
English	7455	2974	3750
Chinese	233	74	107
French	158	54	107
German	130	45	65
Spanish	90	30	43
Japanese	68	15	30
Portuguese	33	14	16

3.4.1.1 Rule-based Approach in Portuguese Clinical Texts

Two rule-based applications for clinical domain are considered in the scope of this review.

De Souza *et al.* applied a rule-based method in discharge summaries written in Portuguese. The purpose was to detect textual information regarding continuity of care applying NLP techniques on an annotated medical corpus. The pre-processing phase included acronym expansion, special characters and stop-word removal, and lowercase normalisation. Afterwards, the POS-tagged texts were analysed to find patterns that indicated continuity of care [105]. The rules were evaluated individually, and only the ones with more True Positives than False Positives as well as more True Negatives than False Negatives were left in the database. The rule identification was performed backwards since the information about continuity of care is usually located near the end of the discharge summary. An example of rule generation using POS-tagged is presented in Figure 3.2. In total, four rules were defined and applied to the text, reaching an overall precision, recall, specificity and F1-score of 0.84, 0.70, 0.97 and 0.76, respectively.

Sentence	Text after the POS-Tagging
RECEBE ALTA BEM, DEVENDO **SEGUIR ACOMPANHAMENTO COM** MÉDICO ASSISTENTE.	recebe_V_PR_3S_IND_VFIN alta_N_F_S bem_ADV devendo_V_GER **seguir_V_INF acompanhamento_N_M_S com_PRP** médico_N_M_S assistente_ADJ_M_S
Rule: [V_INF][N_M_S][PRP] = Infinitive Verb, Noun-Masculine-Singular, Preposition	

Figure 3.2: Example of rule generation using POS-tagged clinical text. Addapted from [105].

A rule-based approach was also implemented by Ferreira *et al.* in MedInX, a medical information extraction system utilising unstructured Portuguese discharge summaries from patients with conditions related to hypertension. MedInX extracted clinical entities from the text, converting the discharge summaries into structured documents. The developed system was utilised in two different tasks: automatic code assignment and systematic analysis of the clinical report completeness and quality. The attribution of codes to each discharge summary was automatically performed, by matching information saved in ICD-9 and International Classification of Functioning, Disability and Health (ICF), and MedInX ontologies to the information extracted from the report [86]. In the content and completeness analysis task, several rules were designed in order to identify episodes in which the number of instances of a certain entity was above or below a certain threshold. Several entities could be evaluated at once, creating more complex rules. The chosen threshold for the detection of outliers in the clinical reports was defined by the user. The definition of a rule included in the MedInX system is illustrated in Figure 3.3, in which the clinical cases with less than 17 Condition instances and more than 17 instances for the entities Medications and Active Substances were extracted. The number of instances of each entity was returned to the user, allowing for an optimised report analysis.

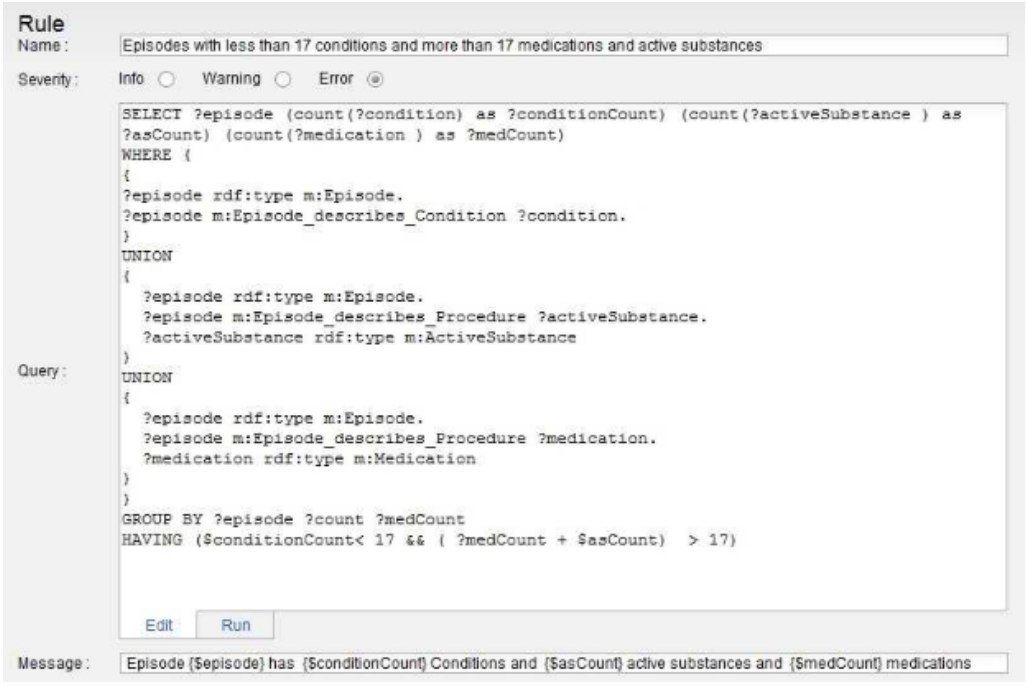


Figure 3.3: Example of rule generation using MedInX system. Addapted from [86].

3.4.1.2 Shallow Machine Learning Approach in Portuguese Clinical Texts

Fernandes *et al.* extracted information from clinical data related to admission to the Emergency Department in order to predict the admission of patients into Intensive Care Units [106], as well as to understand the risk of mortality and cardiopulmonary arrest those patients may face [107]. The ML models in both applications made use of the triage information from the Manchester Triage System (MTS). In this system, the patients are rated from level 1 to 5, with decreasing urgency.

In [107], the ML models inputs were demographics, routine clinical variables obtained at triage, and the patient main complaint. NLP methods were applied to the main complaint, such as contractions fix, punctuation removal, words set to lowercase and tokenisation, as well as abbreviation expansion, replacement of numbers by words, stop-word removal and lemmatisation. Afterwards, the data was vectorised using the TF-IDF approach [92]. Logistic regression, random forest and extreme gradient boosting were trained using stratified random sampling to split the data into training and test set, and 10-fold cross-validation to optimise the model hyperparameters. The predictions were then compared to a reference model, a regularised logistic regression that only uses triage priorities. It was possible to conclude that extreme gradient boosting presented the best results. The higher recall was achieved when using clinical variables and the main complaint for patients with MTS-3, identifying the patients with greater risk of the composite outcome. In [106], the predictors are similar to the previous study. The algorithms used were logistic regression, random forests, and a random undersampling boosting. The algorithms were then compared to the reference model, and it was possible to conclude that logistic regression provided the best results in terms of ICU admission prediction, with a higher recall for patients with MTS-3.

3.4.1.3 Deep Learning Approach in Portuguese Clinical Texts

In regards to DL approaches applied to entity recognition in Portuguese clinical text, Lopes *et al.* implement NER DL models for automatic extraction of meaningful Neurology entities. The total of 281 texts representing Neurology clinical cases are available on GitHub. A comparison between CRF, BiLSTM-CRF and a BiLSTM-CRF with residual learning connections was performed, as well as an analysis of the best features for model training [88]. A in-domain word embedding model for Portuguese clinical text was implemented, using 3 thousand clinical cases extracted from a Neurology Portuguese journal. The model performance when using in-domain and out-domain word embeddings was compared [87], and although the in-domain model was trained with a smaller corpus, it led to a higher performance. The DL algorithms using BiLSTM-CRF revealed better results, with residual learning having similar results as the single layer BiLSTM-CRF.

The increasing use of Transformers, namely the BERT architecture, has led its be application to Portuguese texts. After the implementation of BioBERT, previously mentioned in Section 3.3.2, Schneider *et al.* [108] implemented BioBERTpt to support NER in clinical and biomedical text in Portuguese. Three models were fine-tuned with data from different sources, using a BERT architecture pre-trained with general domain text in several languages (BERT multilingual cased). The clinical corpus utilised in the fine-tuning process is composed by 2 million Brasilian clinical cases with 27 million words, whereas the biomedical corpus has approximately 16 million words from titles and abstracts from Portuguese articles. Each corpus was used for fine-tuning clinical and biomedical domain BioBERTpt models, and the set of both corpora was also used to fine-tune a third model including both domains. The models were evaluated using two datasets: 1 thousand brazilian clinical records from various specialties, and the open-source dataset of Neurology texts from [88]. The model performance was compared with multilingual BERT models, and BERTpt [109], pre-trained in multilingual and Portuguese general domain text, respectively. For both evaluation datasets, BioBERTpt achieved better F1-Scores than the multilingual and Portuguese general domain models. Considering the dataset with several medical specialties, the F1-scores for BERT multilingual uncased, BERT multilingual cased, BERTpt base and BERTpt large were 0.588, 0.582, 0.585 and 0.541, respectively. Alternatively, the F1-scores resulting from BioBERTpt biomedical, clinical and with both biomedical and clinical corpora were 0.602, 0.602 and 0.604, respectively. The F1-scores obtained with the Neurology dataset using the models BERT multilingual uncased, BERT multilingual cased, BERTpt base and BERTpt large were 0.912, 0.921, 0.916 and 0.912, respectively, whereas the F1-score using BioBERTpt biomedical, clinical and with both biomedical and clinical corpora were 0.921, 0.926 and 0.920. These results show that the in-domain BioBERTpt models outperform the BERT models that are trained with general domain text [108].

3.5 Summary

The conducted review on several clinical NLP applications is fundamental to demonstrate the various approaches that can be implemented for clinical information extraction and, more specifically, NER.

Table 3.2: Summary of methods, dataset information, relevant outcomes and best results for clinical NLP applications for information extraction in English text.

Reference	Method	Data	Relevant Outcome	Best Results
Chen et al. [90], 2019	Rule-based approach	Clinical information of 288 patients, with 2 to 5 records per patient.	Rule-based clinical NLP system with good performance on cohort selection for clinical trials, using 13 evaluation criteria. Rules designed using lecical, syntactic, and meta-information.	Overall results(w/o meta information, macro avg): - Precision - 0.87 - Recall - 0.78 - F1-Score - 0-81
Feller et al. [91], 2018	Random forest	181 clinical information from HIV positive patients, and 543 controls of negative patients, including lab tests, demographics, diagnostic codes and unstructured notes.	3 ML algorithms for HIV risk detection: baseline model with structured EHR data; baseline + NLP topics (clustering); baseline + NLP clinical keywords (using 37 manually selected words related to HIV risk).	F1-Score: - baseline - 0.59 - baseline + NLP topic model - 0.63 - baseline + NLP keyword model - 0.74
Chase et al. [93], 2017	Naive-Bayes	Approx. 2,500 Clinical notes from the Columbia University Medical Center (CUMC).	Multiple sclerosis identification based on the frequency of MS-related terms. Synonymous terms were manually clustered into 50 buckets. Models trained using only MS notes or random notes with and without MS cases.	- ROC AUC - 0.90 - Sensitivity - 0.75 - Specificity - 0.91
Pham et al. [95], 2017	LSTM	EHR from a total of 18,109 patients with diabetes and mental health conditions with respective ICD-10 codes. EHR from Australian hospitals.	Memory of illness trajectories and care procedures and, present illness stages estimation, and future risk prediction. Model performs efficiently for diabetes and mental health clinical cases. Evaluation with several diagnoses prediction per case.	Precision - Diabetes: - 1 prediction per case - 0.66 - 2 predictions per case - 0.60 - 3 predictions per case - 0.54 Precision - Mental Health: - 1 prediction per case - 0.53 - 2 predictions per case - 0.46 - 3 predictions per case - 0.40
Xu et al. [96], 2017	BiLSTM-CRF	NCBI Disease corpus dataset, publicly available, with 793 PubMed abstracts and annotated PubMed citations.	Implementation of word-based and character-based LSTM layer. Parameter setting test with several WE vector dimensions using BiLSTM model.	BiLSTM-CRF: Precision - 0.85 Recall - 0.76 F1-Score - 0.80
Lee et al. [97], 2020	BioBERT	PubMed corpus with 4.5 thousand million words. PMC corpus from full arcticles with 13.5 thousand million words. 8 datasets for specific entity extraction task evaluation (disease, drugs, genes and species).	Pre-training 3 BERT models on clinical domain text (PubMed, PMC arcticles and both). Evaluation on 3 NLP tasks: NER, relation extraction and question answering. BioBERT outperformed BERT in all three tasks, but the model trained with PubMed texts is more efficient. Evaluation o several datasets, to extract drugs, diseases, genes and species.	Disease: - Precision - 0.88 - Recall - 0.91 - F1-Score - 0.89 Drug/Chemicals: - Precision -0.92 - Recall - 0.91 - F1-Score - 0.92

Table 3.3: Summary of methods, dataset information, relevant outcomes and best results for clinical NLP applications for information extraction in Portuguese text.

Reference	Method	Data	Relevant Outcome	Best Results
De Souza et al. [105], 2013	Rule-based approach	110 discharge summaries with care continuity information, morphologically tagged.	4 rules defined and applied to 200 discharge summaries. Relevant information identified in the summaries with success. Possibility of editing the rules for identification of other information type in medical text (diagnosis or procedures, for example)	Precision - 0.84; Recall - 0.70; F1-Score - 0.76
Ferreira et al. [86], 2013	Rule-based approach	950 discharge letters from patients with hypertension	Automatic coding for each clinical case and systematic analysis of quality and completeness of clinical notes using extracted entities and creating rules to interpret the clinical note content.	Precision - 0.95; Recall - 0.95; F1-Score - 0.95
Fernandes et al. [106], 2020	Random Forest, Logistic Regression Gradient Boosting	599,276 triage clinical notes, including structured information such as temperature, comma scale, priority, and main complaint.	Risk assessment of mortality and cardiac arrest from triage notes. Predictions with and without main complaint as feature.	Recall (w/main complaint): - Random Forest - 0.94 - Logistic Regression - 0.95 - Gradient Boosting - 0.96
Lopes et al. [87], 2020	CRF, BiLSTM-CRF	281 Neurology clinical texts from Neurology jounal and 20 clinical texts from Coimbra Hospital and University Center for testing. 3377 clinical cases from Neurology journal for in-domain WE training	Annotated dataset with Neurology clinical cases written in Portuguese. CRF implementation as baseline approach. WE model for Portuguese clinical text using FastText algorithm proven to have better performance than general domain WE when used as input feature in BiLSTM-CRF model.	CRF: - Precision: 0.70 - Recall: 0.59 - F1-score: 0.57 BiLSTM-CRF: In-domain WE(macro avg): - Precision - 0.67 - Recall - 0.68 - F1-score - 0.65 Out-of-domain WE (macro avg): - Precision - 0.65 - Recall - 0.63 - F1-Score - 0.62
Schneider et al. [108], 2020	BioBERTpt	Two corpora for BERT model evaluation: 1,000 clinical notes from various Brasilian hospitals with several specialties (SemClinBr); 281 Neurology clinical texts from ... (CLINpt)	3 different pre-trained models with 3 million clinical cases (BioBERTpt(clin)), with paper abstracts and titles from Scielo (BioBERTpt(bio)) and with both corpora (BioBERTpt(all)). Pre-training with clinical domain texts lead to a better performance in NER task when comparing with other BERT based models.	F1-Score (SemClinBr): - BERTpt - 0.58 - BioBERTpt(clin) - 0.60 F1-Score (CLINpt): - BiLSTM-CRF - 0.75 - BERTpt - 0.91 - BioBERTpt(clin) - 0.93

The majority of the NER problems require an annotated corpus, so that clinical entities of interest are identified and later extracted by the model. The models can be divided into three main approaches: rule-based approaches, shallow ML and DL. Tables 3.2 and 3.3 summarise the methods, data, and relevant outcomes from clinical NLP applications for information extraction tasks in English and Portuguese, respectively. The approaches for clinical NER with Portuguese clinical text have special relevance since, as shown in Section 3.4, there is a much smaller percentage of applications in languages other than English, and the works presented dealt with similar challenges characteristic of the Portuguese language.

It is important to emphasise that there are limitations in terms of comparing the presented results, as well as the results obtained in the context of this work, due to differences between the utilised datasets. Observing the evolution in approaches for clinical information extraction, the approaches to implement are CRF, BiLSTM, BiLSTM-CRF and BERT finetuning.

Chapter 4

Dataset Analysis and Experimental Setup

It is fundamental to have a comprehensive understanding of the data used in the course of the dissertation work to adopt the most relevant approaches. This chapter introduces the Dermatology dataset with NHS clinical cases. The dataset analysis comprises the distributions of structured data, such as case priority, gender, appointment type and ICD codes. Besides, the unstructured text data is analysed. The annotation process is described, having into consideration which clinical entities are relevant to extract in the scope of the work. Lastly, the word embedding model implementation is outlined.

4.1 NHS Dermatology Dataset

The NHS Dermatology corpus consists of information on 13,058 dermatology clinical cases from the Portuguese NHS. Each case includes a case ID, the patient age and gender, the appointment type, the priority assigned to the case at screening and the ICD-9 codes assigned to each case by the dermatologist. Each clinical case is also composed of unstructured clinical notes. These notes are divided into four different sections.

The priority assigned to each case is represented by an integer number that can vary from 1 to 3. The higher priority cases are represented by the number 1, whereas the lower priority cases are assigned with the number 3. Figure 4.1 represents the dataset distribution considering the priority assigned to the clinical cases. The number of instances with a low priority are 10,516, representing the majority of cases. These are followed by 2,205 intermediate priority cases and 337 high priority cases. There is a notorious class imbalance when dividing the clinical cases per priority.

Figure 4.2 represents the patient population gender distribution, as well as the Dermatology screening appointment type, that indicates whether the appointment was in-person or a tele-appointment. Figure 4.2a shows that there is a balance regarding the patient gender in the dataset. Concerning the appointment type, depicted in Figure 4.2b, it is observable that the great majority of appointments are tele-appointments, representing approximately 86% of the

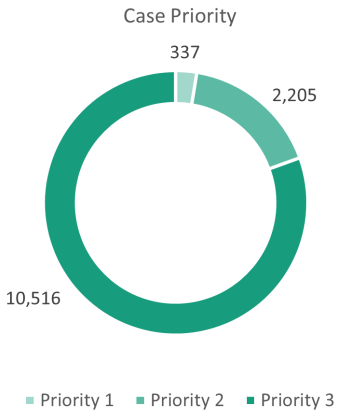
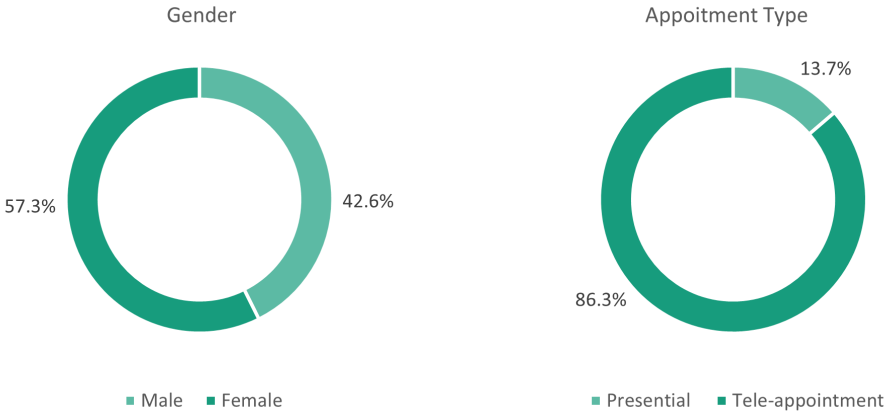


Figure 4.1: Dataset distribution according to case priority.

total of appointment instances. Cases in which there is a referral to in-person appointments are still considered tele-appointments in the dataset, leading to the discrepancy in the appointment types.



(a) Patient Gender. **(b)** Appointment Type.

Figure 4.2: Dataset distribution according to patient gender and appointment type.

The ICD codes, previously described in Section 2.4.2, are a systematic representation for conditions or procedures. The codes used in the Derm.AI dataset correspond to the version 9 of ICD. There are two ICD-9 code representations on the dataset: a 3-digit representation that only considers the category information, and the most specific code with one or two decimal places, representing additional information such as the site, lateralisation or aetiology. The ICD-9 codes are assigned to the case by the dermatologist responsible for the appointment. In total, 329 unique ICD-9 codes occur when considering the decimal places, a number that decreases to 111 when only considering the more general first three digits. Table 4.1 displays the ten most frequent ICD-9 codes in the dataset, along with their frequency and meaning.

Concerning the unstructured part of the clinical notes, as previously stated, the information is divided into four sections. The sections are Summary, Therapeutics, Proposed Exams and

Table 4.1: Ten most frequent ICD-9 codes in the dataset, with the corresponding frequency and meaning.

ICD-9 Code	Frequency	Meaning
702.1	2761	Other dermatoses - Seborrheic keratosis
448.1	2212	Nevus, non-neoplastic
702.0	1439	Other dermatoses - Actinic keratosis
238.2	739	Neoplasm of uncertain behaviour of other and unspecified sites and tissues - Neoplasm of uncertain behaviour of skin
696	483	Psoriasis and similar disorders
702.19	292	Other dermatoses - Other seborrheic keratosis
706.1	275	Diseases of sebaceous glands - Other acne
078.1	254	Other diseases due to viruses and clamydiae - Viral warts
173	243	Other and unspecified malignant neoplasm of skin
695.3	241	Erythematous conditions - Rosacea

Conclusion, and each one should contain information about the respective field. However, the structure is not always followed by the physician that is taking notes in each appointment. For that reason, the content in each section can often differ from case to case, and some can be left empty when taking notes. The bar chart in Figure 4.3 presents the number of instances with and without textual information divided into the sections previously mentioned. While the Referral Summary and Therapeutics sections are more balanced in terms of textual content, the other two sections exhibit a significant imbalance in that matter. The Proposed Exams section has the lowest number of cases with text, whilst the Conclusion section contains, by a wide margin, the highest number of instances with text. This discrepancy can be justified by the fact that, in some clinical cases comprised in the dataset, there are no proposed exams nor a defined therapeutic, which lead to the lower number of instances with text in the respective sections. Besides, in several instances, the information sections are not respected by the physicians, who end up filling them in a non-organised manner. For example, there is a significant set of cases in which all the clinical information is written in the Conclusion, rather than divided by the other dataset sections.

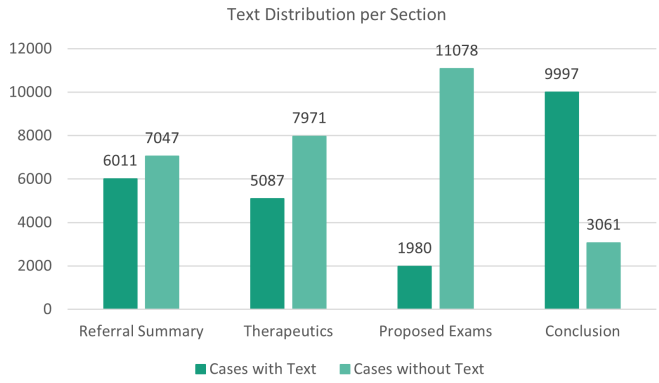


Figure 4.3: Dataset distribution according to cases with and without text for each section.

The last statement is supported with the data analysis in Figure 4.4, which illustrates the number of clinical cases with a certain amount of text sections filled in the dataset. Out of the total 13,058 clinical cases, 7,486 instances only have text in one of the four possible sections of the clinical note. In contrast, only 1,585 clinical cases have text information in all four sections of the document.

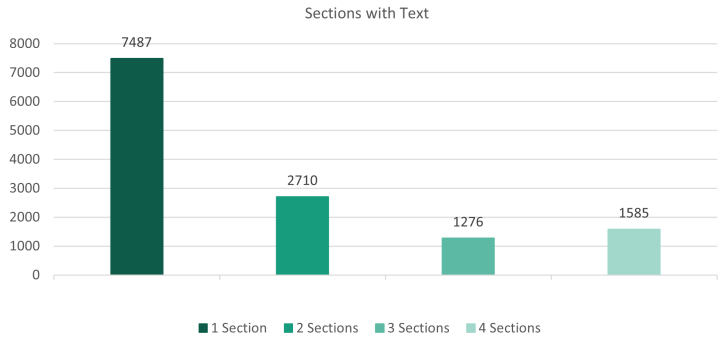


Figure 4.4: Dataset distribution according to text sections filled per clinical case.

4.2 Annotation

When training a supervised model to perform an entity extraction task, it is necessary to have a corpus labelled with the entities of interest for that specific implementation. The Dermatology corpus is available in a raw, unstructured format. Therefore, it was necessary to carry out manual annotation of these texts with clinical entities of interest. The annotation was executed using the annotation tool Prodigy[14]. The Prodigy interface for manual annotation is represented in Figure A.0.1, found in the Appendix. Besides the manual annotation to create training and evaluation data for entity extraction models, Prodigy includes several methods to inspect the data and support the annotation task. For instance, the model performance when increasing the training data tended to increase as well, suggesting that a greater number of annotated notes would be beneficial to the results. However, since the annotation process is time-consuming and there was limited time available for the annotation process, only 5 thousand sentences belonging to 980 clinical cases were annotated by the last year Integrated Masters student who carried out the dissertation work.

4.2.1 Clinical Entities

The baseline entities used for the annotation process were the entities defined by Ferreira *et al.* [86]. The entities defined in that previous work are the following: Administration Route, Anatomical Site, Condition, DateTime, Evolution, Lateralisation, Modifier, Negation, Procedure and Value. Alterations were made to the mentioned entities, considering the scope of the dissertation work and the Dermatology corpus.

Analysing the methods and procedures performed in Dermatology appointments present in the corpus, the Route of Administration entity was not considered. Moreover, in [86], the

[14] "Prodigy": `https://prodi.gy/`.Accessed on 19-05-2021

Procedure entity was composed of four subclasses - Therapeutic or Preventive, Laboratory, Diagnostic and Chemical. However, having the information included in the Dermatology dataset into account, only the Therapeutic subclass is used as an equivalent entity to those previously mentioned.

In the context of the Derm.AI project, there is interest in extracting additional information to the previously mentioned entities. One of the main objectives in Derm.AI is the junction of information from dermatological imaging analysis and clinical notes. An automatic assessment of image quality based on the clinical notes is desired. An additional entity named Exam is added to the base set of entities to extract information about clinical photographs and other exams. Additionally, information about the referral to an in-person appointment or procedure is important to define the risk associated with the case, and therefore its priority. The referral can also happen when the images have low quality and the diagnostic cannot be precise. Hence, the entity Appointment is also added to the previously mentioned entities. Its purpose is to extract information about appointments and procedures to which the patient already has been or scheduled for a future date. This results in eleven entities represented in Figure 4.5, the description of which can be found in Table 4.2.

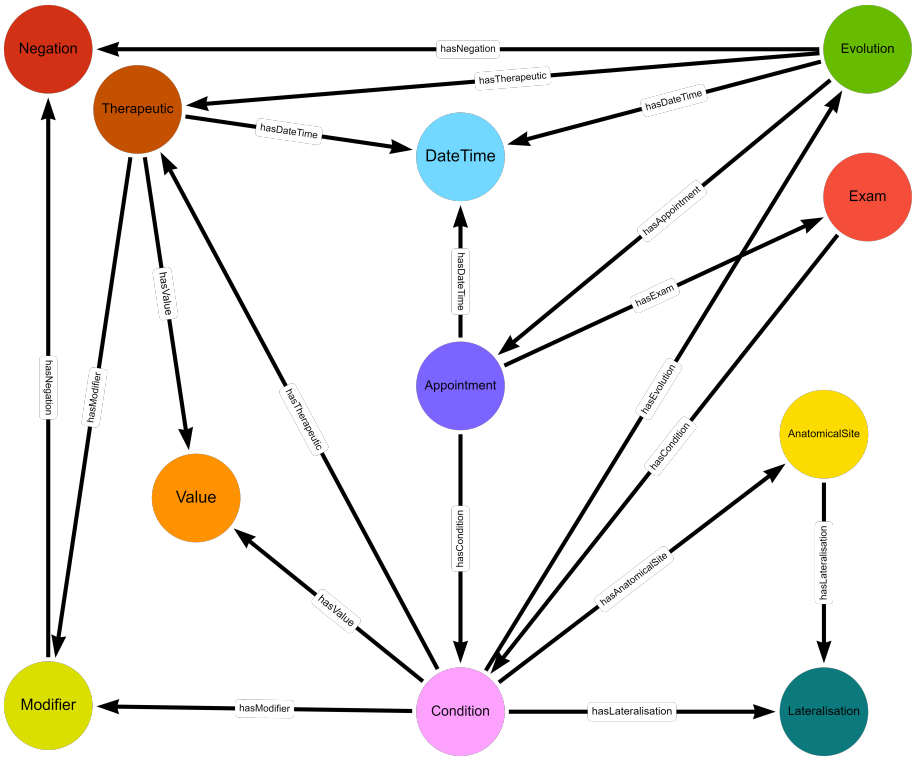


Figure 4.5: Representation of the entities of interest in TRIPOD.

Table 4.2: Entities considered in the NER problem and respective description.

Entity	Description
Anatomica Site	References to an anatomical structure or location, normally the site in which Condition is observable;
Appointment	References to appointments and in-person procedures scheduled for a future date;
Condition	References to a pathology, disease, symptom or complication of a patient;
DateTime	References to temporal expressions, such as frequencies, dates and times and durations;
Evolution	References to clinical evolution of a patient with certain Condition;
Exam	References to clinical photographs and exams used in the diagnosis;
Lateralisation	References to the lateralisation of an Anatomical Site or Condition;
Modifier	References that modify an entity, such as Condition or Therapeutic;
Negation	References to expressions denoting negation of an entity, being considered apart from said entity;
Therapeutic	References to treatments previously carried out or to be carried out to handle a certain Condition;
Value	Numerical values not included in DateTime entities, including the unit that characterises the values.

The IOB tagging format, mentioned in Section 2.3.2.1, is employed in the annotations, to differentiate between tokens in the beginning and inside of a certain entity. This results in twenty-three different tags, two for each entity, representing its beginning and inside tokens, and the O tag, attributed to tokens that do not belong to any of the entities. A representative example of an annotated sentence from the dataset is presented in Table 4.3.

Table 4.3: Example of annotated sentence using IOB tagging

Token	IOB Tag
Antecedentes	B-Evolution
de	I-Evolution
seguimento	I-Evolution
em	O
consultas	B-Appointment
de	I-Appointment
Dermatologia	I-Appointment
e	O
Cirurgia	B-Appointment
Plástica	I-Appointment
por	O
carcinoma	B-Condition
basocelular	I-Condition
temporal	B-AnatomicalSite
esquerdo	B-Lateralisation

Table 4.4 provides a corpus analysis in terms of the number of tokens labelled with a certain tag, from the total of 73,906 tokens in the annotated corpus, as well as the number of unique tokens per tag, from the total of 10,087 unique tokens. The unique token analysis includes stopwords and punctuation, and the corpus was not lowercase. Thus, there is a difference between two tokens if one of them has an uppercase character. Table 4.4 also includes the percentage of total and unique tokens for each tag. Interpreting the information, as expected, the O tag is the most frequent, accounting for 61.58% of the corpus tokens. In contrast, the tags I-Lateralisation and I-Negation account for 0.06 and 0.01% of the corpus tokens, respectively. The residual percentages for I-Negation are justified as errors in the annotation process. The Negation entity only considers single word expressions, such as "Não" (*No*), but two instances were annotated as Negation when the correct entity was Value, generating inside tags for the entity under consideration. On the other hand, the I-Lateralisation tag attribution happened in rare cases with more than one lateralisation mention. Those cases can be illustrated in the expression "face posterior esquerda" (*posterior left face*), the last two tokens belong to the Lateralisation entity. The B-Lateralisation tag corresponds to the token "posterior" (*posterior*), whereas the I-Lateralisation tag corresponds to the token "esquerda" (*left*).

4.2.2 Annotation Validation

The manual annotation using Prodigy was performed by one person (the last year Integrated Masters student who carried out the dissertation work). Therefore, it is essential to validate the annotation. The validation was carried out by two physicians whose specialty is Dermatology, who collaborate in the Derm.AI project. The validation was done manually and, for each physician, the agreement ratio is calculated.

The results from the annotation validation are presented in Table 4.5. As a matter of simplifying the manual annotation process, the IOB tagging was not considered in the validation process. Instead, the general clinical entities defined in Table 4.2 were considered, as well as the Out tag. Both physicians validated the same 400 sentences, so that it is possible to evaluate the agreement between the two annotators.

An analysis of the agreement ratios leads to the conclusion that, considering the validation performed by Physician 1, the entities Evolution and Modifier have the lowest agreement ratio value. This situation was expected, since the two entities encompass a larger amount of text, as shown in Table 4.2. Out of all the inside tags, I-Modifier and I-Evolution have the largest amount of unique tokens. The unique token quantity demonstrates that the expressions annotated with these entities are lengthy with a greater token variety. The Out tag presents the highest agreement ratio, which means that the physician agreed that most of the tokens without any associated entity do not belong to any of the clinical entities. On the contrary, taking into account the validation performed by Physician 2, the Out tag exhibits the lower agreement ratio out of all considered entities. This happens because the physician considered that tags without an associated entity should in fact belong to a clinical entity. On the other hand, the agreement ratio for the clinical entities presents satisfactory results, with agreement ratios above 90% for all entities except Exam. This means that Physician 2 agreed with most of the identified entities, and would consider more tokens as clinical entities.

Table 4.4: Number of tokens annotated with each entity in the annotated corpus

Entity	**Total tokens**	**Total token percentage (%)**	**Unique tokens**	**Unique token percentage (%)**
B-AnatomicalSite	1,961	2.65	366	3.63
I-AnatomicalSite	657	0.90	150	1.49
B-Appointment	1,001	1.35	137	1.36
I-Appointment	1,237	1.68	187	1.86
B-Condition	2,395	3.24	458	4.54
I-Condition	1,223	1.65	288	2.86
B-DateTime	1,519	2.06	319	3.16
I-DateTime	2,168	2.93	223	2.21
B-Evolution	1,200	1.62	356	3.53
I-Evolution	3,639	4.92	789	7.82
B-Exam	432	0.59	92	0.91
I-Exam	585	0.79	166	1.64
B-Lateralisation	612	0.83	64	0.63
I-Lateralisation	48	0.06	20	0.20
B-Modifier	2,918	3.95	834	8.27
I-Modifier	2,116	2.86	650	6.44
B-Negation	427	0.58	15	0.15
I-Negation	11	0.01	6	0.06
B-Therapeutic	1,838	2.49	520	5.15
I-Therapeutic	713	0.96	169	1.68
B-Value	994	1.35	213	2.11
I-Value	701	0.95	46	0.46
O	45,511	61.58	4,019	39.84
Total	**73,906**	**100**	**10,087**	**100**

Table 4.5: Agreement analysis using annotation validations from two physicians and average of both validations. Percentage agreement ratio (AR) per entity and for the total agreed tokens (AT) and non-agreed tokens (NAT).

		Physician 1			Physician 2			Mean ± STD
Entity	**Total Tokens**	**AT**	**NAT**	**AR (%)**	**AT**	**NAT**	**AR (%)**	**AR (%)**
Appointment	220	195	25	88.64	219	1	99.54	94.09±5.45
Anatomical Site	152	114	38	75.00	142	10	93.42	84.21±9.21
Condition	230	178	52	77.39	225	5	97.83	87.61±10.22
DateTime	251	193	58	76.89	249	2	99.20	88.05±11.16
Evolution	416	265	151	63.70	385	31	92.55	78.12±14.42
Exam	98	83	15	84.69	86	12	87.76	86.22±1.53
Lateralisation	48	37	11	77.08	48	0	100.00	88.54±11.46
Modifier	427	257	170	60.19	394	33	92.27	76.23±16.04
Negation	44	34	10	77.27	44	0	100.00	88.64±11.36
Therapeutic	170	155	15	91.17	189	2	98.95	90.05±8.90
Value	134	116	18	86.57	126	8	94.03	90.30±3.73
Out	3749	3565	184	95.09	3097	652	82.61	88.85±6.24
Total	**5939**	**5192**	**747**	**87.42**	**5204**	**756**	**87.32**	**87.21±0.10**

Considering the total agreed and non-agreed tokens, and calculating the agreement ratio with that information, the results of both annotation validations are very similar. The mean value of both validations is 87.21±0.10%, a satisfactory result having in mind the considerable amount of clinical entities extracted from the clinical notes, as well as the results found in the literature both for clinical Portuguese corpora [86], [87] and clinical corpora in other languages [110]–[112].

4.3 Word Embeddings

As previously stated in Section 2.6.3, Word Embeddings (WE) are vectorised word or concept representations. To obtain WE, it is necessary to train the model using textual data. Having into consideration the literature that stated that in-domain WE originated from clinical text perform better than out-of-domain WE obtained from general domain text [87], [113], [114], only clinical domain WE were utilised in the course of the work.

To obtain WE within the Dermatology domain and train the model using similar text to the NHS Dermatology dataset previously described, 626 clinical cases from the Journal of the Portuguese Society of Dermatology and Venereology were used. The clinical cases belong to all fifty-seven available editions of the Journal, from March of 2005 to March of 2021, making a total of 163,828 tokens. However, the number of Dermatology cases is small since the journal mostly publishes articles in English in more recent editions. The reduced amount of words could compromise the WE model performance. For that reason, the publicly available clinical cases from a Neurology journal used in [87] and [88] to train word embeddings, composed by 3,377

cases and 686,762 tokens, were added to the Dermatology clinical cases for the WE training. The total of tokens used in the WE training is 850,590, 15,620 of them being unique tokens.

From the algorithms described in Section 2.6.3, the chosen algorithm for the WE model was FastText. Having clinical text into consideration, the ability of not only representing out-of-vocabulary words but also learning morphological characteristics of the words, such as prefixes or suffixes, because of the use of character n-words, makes FastText the most suitable algorithm. The FastText algorithm used for training the model is made available in the Gensim library [115].

The hyperparameters used for the FastText training are presented in Table 4.6. Since only words that occur more than five times in the corpus are learned by the word embedding model, the final vocabulary contains 9,444 unique tokens. The WE dimension is set to 300, and the minimum n-gram is 1 so that all the characters can be utilised for training and therefore allowing for unknown words recognition. The selected training algorithm is Skip-Gram that, as explained in Section 2.6.3, predicts the most similar words given a certain word. Lastly, negative sampling is utilised so that the number of neuron weight updating is reduced and, therefore, the training time is reduced as well. The chosen value of 10 defines the number of noise words that are drawn as a negative sample.

Table 4.6: Hyperparameters used in FastText training.

Hyperparameter	**Value**
Vector Dimension	300
Training Algorithm	Skip-gram
Minimum Word Count	5
Minimum n-gram length	1
Negative Sampling	10

With the purpose of analysing the FastText model performance, several words belonging to a Dermatology specific vocabulary were chosen and the five words with the most similar WE were selected. The words from the Dermatological specific terminology and the five most similar words are presented in Table 4.7. For the most part, the most similar words belong to the same context, as observed for the words "Melanoma" (*Melanoma*), "Queratose" (*Keratosis*) and "Cutânea" (*Cutaneous*). However, in other cases such as "Eczema" (*Eczema*), "Nevo" (*Nevus*) or "Verruga" (*Wart*), some of the similar words are not related to the clinical domain, such as "esquema" (*scheme*) "levo" (*take*), "novo" (*new*) and "vera". The fact that words outside the Dermatology specific terminologies are included in the most similar words may be justified by the use of Neurology clinical cases to train the model. Even if it is clinical text, the terminologies are quite different. Nevertheless, being clinical domain text, the use of Neurology clinical cases is better than using general domain text. It is possible to conclude that the FastText model has satisfactory performance when considering specific Dermatology domain words. The results could be improved by using more dermatological domain cases for the FastText model training.

Table 4.7: Analysis of five most similar words using FastText word embedding model, for frequent conditions and expressions in Dermatology clinical notes.

Word	5 Most Similar Words
Melanoma	melanocítica, melanodérmica, melanina, melanocíticos, adenoma
Eczema	edema, empiema, enfisema, esquema, linfedema
Nevo	nevos, levo, nervo, novo, relevo
Queratose	paraqueratose, ceratose, hiperqueratose, aretomatose, queratodermia
Seborreica	seborreicas, verborreia, diarreicas, rinorreia, sialorreia
Verruga	verrocosa, verrucoso, vera, ver, vermelha
Cutânea	cutâneas, percutâneas, subcutânea, mucocutânea, cutâneo

The FastText model described is further employed to obtain the vector representation of each token in the NHS dataset in both clinical entity extraction and risk prioritisation tasks, described in Chapters 5 and 6, respectively. The WE for each token and the corresponding lemma are calculated. Afterwards, both WE are used as inputs for the neural networks implemented in each task.

4.4 Summary

Throughout this chapter, several key insights of the dissertation work implementation are analysed and discussed. Starting with the NHS Dermatology dataset analysis, it encompasses both structured and unstructured data. Concerning the case priority distribution, there is a significant imbalance among the priority values, since cases with the lowest priority account for the majority of the dataset. The appointment type analysis reveals that most clinical cases relate to tele-appointments. All clinical cases have a corresponding ICD-9 code, which directly relates to the diagnosed condition. As regards the unstructured clinical notes, these are divided into four sections. However, as concluded in Section 4.1, the physicians frequently do not respect the delimited text sections. Therefore, for each case, the clinical notes comprised in the fours sections are merged into one single clinical record.

The manual annotation process included 980 clinical cases from the analysed Dermatology dataset, corresponding to 5 thousand sentences. The selected clinical entities for the annotation include typical entities used in clinical entity extraction problems. In addition, information of interest for the Derm.AI project was also considered. The Appointment and Exam entities represent the appointment type and in-person referrals and the existence and quality of the clinical photographs, respectively. The validation performed by two physicians reveal a mean agreement ratio of $87.21 \pm 0.10\%$, a result in accordance with the literature that allows to conclude that the extracted entities in the manual annotation process are indeed of interest for dermatology clinical practice. The annotated texts will be used in the training of the clinical entity extraction models. Clinical notes and the corresponding clinical entities will also be used as input for the risk prioritisation model.

Regarding the WE model, the FastText algorithm is chosen because of its capability of recognising out-of-vocabulary words and considering morphological characteristics in the vector space. The fact that Neurology texts were used in the WE model training increases the vocabulary size, but decreases the specificity of the medical terms in the vocabulary. Nevertheless, since it is clinical text, it is adequate for the WE purpose in this work. Training the WE model using only Dermatology texts in a substantial amount may potentially improve the model performance.

Chapter 5

Clinical Entity Extraction

This chapter addresses the clinical entity extraction task, with the aim of evaluating the extraction of relevant information in the context of Dermatology clinical practice and the Derm.AI project, as described in Section 4.2.1. The chapter starts by considering the common characteristics that exist in all model implementations for this task. Afterwards, the information is divided by implemented approach, including theoretical and practical considerations, followed by the respective results and discussion of that same model. A comprehensive comparative analysis is done at the end of the chapter, having the results for all implemented approaches into account.

Considering the literature for clinical entity extraction tasks, four different approaches are selected to assess this task. CRF is a shallow ML approach relevant to sequence prediction, using additional contextual information in model prediction. The sequence prediction allows the model to understand the dependencies between labels and possible transitions between labels. Adopting the IOB tagging, CRF can learn that an I tag should not follow an O tag, because the first token of an entity must be labelled with a B tag. BiLSTM is a bidirectional neural network architecture in which one LSTM layer takes the input in a forward direction whereas the other LSTM layer receives the input in a backward direction. The bidirectionality increases the available context information available to the algorithm. However, each model prediction is independent, so the logic behind IOB tagging cannot be learned. The BiLSTM-CRF model combines the previously mentioned approaches. By adding a CRF layer to the BiLSTM model, it can learn the logic regarding the possible label transitions, increasing the model performance. Lastly, BERT based models apply the bidirectional training of the Transformer attention model, allowing for the model to have a deeper sense of language context than single-direction approaches.

5.1 Experiment Design

There are several characteristics that are common to all the models implemented in the clinical entity extraction task. The data employed in this task results from the manual annotation process described in Section 4.2. A total of 5 thousand sentences corresponding to 980 clinical cases were annotated with the entities of interest.

Since the number of annotated sentences is scarce, an independent test set was not defined.

Instead, in all approaches, K-fold cross-validation with K equal to 5 was applied. In each fold, 1 thousand sentences are set as test data. From the remaining sentences, 1 thousand are saved as the validation set, and the training data in each fold is composed of 3 thousand sentences. The training data is used to train a model in each fold. The model is then evaluated using the validation set, by analysing the corresponding macro F1-score, that is considered in the model optimisation phases and in the evaluation of the model generalisation capacity. Lastly, a prediction is made in each fold using the respective test set, and both the ground truth values and the predicted values for the 5 trained models are saved. Therefore, the evaluation of each approach is done using all the manually annotated data.

The evaluation metric utilised in the clinical entity extraction task to assess model performance is F1-score, although precision and recall outcomes are also presented. The expressions for the mentioned evaluation metrics are presented in Equations 2.8, 2.6 and 2.7, respectively.

Taking into consideration the token percentage for each tag analysed in Table 4.4, the percentage of tokens annotated with the O tag is substantially higher than the percentages for the remaining tags. On the contrary, the I-Negation and I-Lateralisation tags percentages are negligible comparing with the remaining tags. For this reason, in an attempt to minimise the class imbalance with the other entities, the above-stated tags are left out from the clinical entity extraction task evaluation.

5.2 Classification with CRF

ML architectures, such as CRF, require a selection of relevant features for the addressed task. In NER problems, relevant features have been analysed, such as lemma, POS, word morphological and orthographic characteristics [110].

The token classification task can benefit from extracting features from the said token, as well as neighbour tokens. In the developed CRF implementation, 3-token feature extraction was employed, considering the current token, the previous and the subsequent tokens, to which the following features were extracted:

- Linguistic features:
  - Token, e.g. "Delimitados" (*Delimited*);
  - Lowercased token, e.g. "delimitados" (*delimited*);
  - POS tag, e.g. "VERBO" (*VERB*);
  - Lemma, e.g. "delimitar" (*delimit*);
- Morphological and Orthographic features:
  - Token begins with uppercase word, e.g. "Aerius". Possible values: True and False;
  - Token only has uppercase characters, e.g."QA". Possible values: True and False;
  - Token is numerical value, e.g. "42". Possible values: True and False;
  - Token prefixes using 3-character window: considering the token "granuloma" (*granuloma*), the suffixes are "gra", "gr" and "g";

- Token suffixes using 3-character window: considering the token "granuloma" (*granuloma*), the prefixes are "oma", "ma" and "a";
- Token length, e.g. "granuloma" (*granuloma*) has 9 characters.

In the model training process, it is possible to assess the weight that each feature has in the model prediction outcome. The feature weight analysis allows a better understanding of what features have more influence in the model prediction, with positive weights corresponding to a greater influence in the model outcome. Summing all the weights for each feature, it is possible to analyse which ones have a greater influence in the model prediction.

The CRF model implementation utilised the *sklearn-crfsuite* package [15] in Python. The hyperparameters used in the CRF training are presented in Table 5.1. The number of maximum iterations is set to 100, and the possible transitions are set to false so that transitions between entities that are not observable in the training set are not considered by the model. The L1 and L2 coefficient values for regularisation result from a hyperparameter optimisation using a gradient descend for the Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [116] algorithm.

Table 5.1: Hyperparameters used in CRF training.

Hyperparameter	**Value**
Training Algorithm	Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)
Maximum Iterations	100
All Possible Transitions	False
L1 Coefficient	2^{-1}
L2 Coefficient	2^{-5}

The L1 and L2 coefficients combination considered in the grid search varied between 2^{-5} and 2^{5}, with the exponent varying 1 unit at the time. The values were selected having macro average F1-score as the metric to maximise. The grid search results for the L1 and L2 coefficients variation is represented in Figure 5.1. It can be concluded that higher F1-scores are obtained when using lower values for L1 and L2 coefficients. The best performance is achieved for the L1 coefficient equal to 2^{-1} and the L2 coefficient equal to 2^{-5}, resulting in an F1-score of 0.645. The results are in accordance with the literature since lower coefficients cause the most important features to have a higher weight. In contrast, higher coefficients give equal weights to all features [88].

[15] "sklearn-crfsuite package": `https://sklearn-crfsuite.readthedocs.io/en/latest//`.Accessed on 02-06-2021

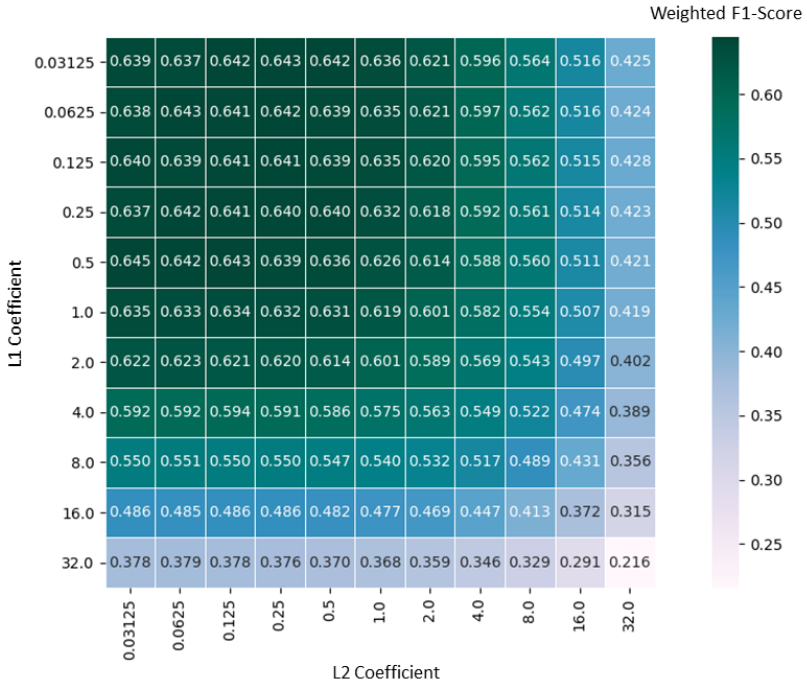


Figure 5.1: Grid Search for L1 and L2 coefficients optimisation with Macro F1-score maximisation.

5.2.1 Results and Discussion

Figure 5.2 comprises the macro average F1-scores both for the validation and test set. The validation results of the 5-fold cross-validation are also reported in Figure 5.2 as it allows discussing the generalisation power of the model and the existence, or not, of overfitting.

It is possible to conclude that, considering the macro F1-scores, the results are similar for the test and validation set. So, it can be assumed that overfitting was avoided, and the model is able to generalise its performance to unseen data.

The results for the clinical entity extraction are presented in Table 5.2. For each entity resulting from the IOB tagging, the precision, recall and F1-score are displayed. The same evaluation metrics are calculated using macro and weighted averages to evaluate the model performance.

Similarly to the annotation validation results, the B and I tags referring to the Modifier and Evolution entities show the worst F1-score values. This is expected because these entities refer to parts of the text that are too general and variable, with a high unique token percentage, as confirmed in Table 4.4. Other entities in which the text is more specific, such as Value, Lateralisation or Negation, are more easily correctly predicted by the CRF model, achieving better F1-score values. Regarding the Anatomical Site and Condition entities, that incorporate relevant information for clinical cases risk prioritisation, the B tags (B-AnatomicalSite and B-Condition) reveal F1-scores of 0.78 and 0.79, respectively. The high F1-score values for these tags allow to conclude that the vast majority of instances are being correctly identified by the model. However, the F1-score values for the I tags (I-AnatomicalSite and I-Condition) are considerably lower, decreasing to 0.62 and 0.64, respectively. These tags have a lower amount of tokens than

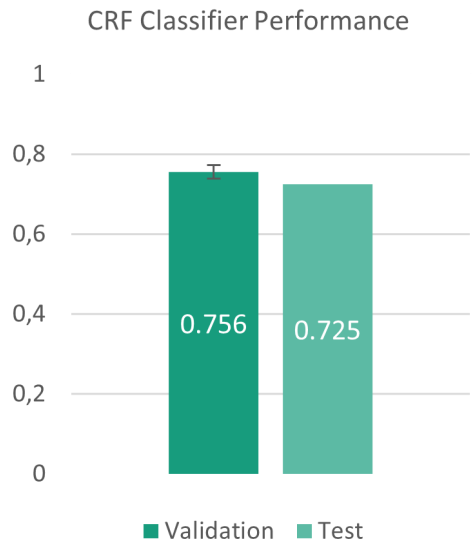


Figure 5.2: CRF model performance assessment using macro F1-score for test and validation sets.

the B tags, which influences the ability of the model to be able to make a correct prediction.

Comparing with the results obtained in the literature using CRF architectures for entity extraction, Lopes *et al.* [88] achieved an average macro F1-score for 10-fold cross-validation of 0.765±0.023%, using 13 clinical entities. Thus, it can be stated that the results obtained with the CRF architecture, using the annotated Dermatology clinical notes, are similar to the results found in the literature.

5.2.1.1 Entity Transition Analysis

The CRF algorithm performs a prediction based on sequence information. The model can learn what are the most and least likely tag transitions. Each transition is assigned a weight, either positive or negative, representing the transition likelihood.

Table 5.3 displays the ten most likely transitions learned by the CRF model. Analysing the values, seven out of the ten most likely transitions represent a transition from the B tag to the I tag of a certain clinical entity, e.g. from B-DateTime to I-DateTime. From these results, it can be concluded, to a certain extent, that the logic behind the IOB tagging is understood by the model. Being able to consider this logic is detrimental to making the model prediction as precise as possible. The fact that DateTime, Lateralisation and Exam are the entities in which the transition from the B tag to the I tag have the highest weights may happen because the majority of instances belonging to the mentioned entities are similar. Therefore, it is expected that the prediction is more precise in these entities than, for example, the entity Modifier, that comprises all tokens that modify other entities and has a higher number of unique tokens, as confirmed in Table 4.4. The three remaining most likely transitions are between two tokens of the same entity with an I tag, e.g. from I-Exam to I-Exam. The three entities with a likely I tag to I tag transition comprised in Table 5.3 (Exam, Appointment and DateTime) are entities in which each

Table 5.2: Results for clinical entity extraction task using CRF model. For each evaluation metric, best results in bold and worst results underlined.

	Precision	**Recall**	**F1-Score**
B-AnatomicalSite	0.81	0.75	0.78
I-AnatomicalSite	0.66	0.59	0.62
B-Appointment	0.79	0.73	0.76
I-Appointment	0.68	0.67	0.67
B-Condition	0.82	0.76	0.79
I-Condition	0.67	0.62	0.64
B-DateTime	0.81	0.77	0.79
I-DateTime	0.84	0.82	0.83
B-Evolution	0.56	0.43	0.49
I-Evolution	0.56	0.52	0.54
B-Exam	0.87	0.78	0.82
I-Exam	0.71	0.54	0.62
B-Lateralisation	0.87	0.80	0.83
B-Modifier	0.61	0.53	0.57
I-Modifier	0.57	0.52	0.54
B-Negation	**0.88**	**0.94**	**0.91**
B-Therapeutic	0.86	0.78	0.82
I-Therapeutic	0.82	0.68	0.74
B-Value	**0.88**	0.82	0.85
I-Value	0.81	0.74	0.77
macro avg	0.75	0.69	0.72
weighted avg	0.72	0.66	0.69

occurrence has typically more than one token. For that reason, it becomes easier for the model to learn these transitions, becoming part of the most likely in the CRF model prediction.

In contrast, Table 5.4 presents the ten least likely transitions in the CRF model prediction. Given that the model only considers transitions that are observable in the training set, the presented transitions are found in the annotated corpus. In the annotation process, there were instances in which two consecutive tokens from the same entity were separated. Considering the example "Aplicar aldara ou imicare 3x semana durante 4 a 6 semanas" (*Apply aldara or imicare 3x week during 4 to 6 weeks*), extracted from the annotated corpus, the expressions "3x semana" (*3x week*) and "durante 4 a 6 semanas" (*during 4 to 6 weeks*) are considered two separate instances from the DateTime entity. That generates an unusual I-DateTime $\rightarrow$ B-DateTime transition, that is unlikely to be predicted by the CRF model. The same situation happens with the Modifier

Table 5.3: Ten most likely tag transitions in CRF model prediction. Outline of involved tags in transition (From → To) and respective weight.

From	→	To	Weight
B-DateTime	→	I-DateTime	9.334
I-Exam	→	I-Exam	8.403
B-Lateralisation	→	I-Lateralisation	8.079
B-Exam	→	I-Exam	8.016
I-Appointment	→	I-Appointment	7.951
I-DateTime	→	I-DateTime	7.835
B-AnatomicalSite	→	I-AnatomicalSite	7.712
B-Therapeutic	→	I-Therapeutic	7.490
B-Appointment	→	I-Appointment	7.378
B-Modifier	→	I-Modifier	7.147

and Evolution entities, as concluded by analysing the weights of the transitions I-Modifier → B-Modifier, B-Modifier → B-Modifier, and I-Evolution → B-Evolution. Besides, from the values presented, it can be stated that the Evolution entity is rarely preceded by the Modifier entity or followed by text that does not belong to any entity (and, therefore, assigned with the O tag). The transition B-Value → B-Value is most likely to result from an annotation error, since the Value entity is commonly composed by more than one token, corresponding to the numerical value and the unit.

Table 5.4: Ten least likely tag transitions in CRF model prediction. Outline of involved tags in transition (From → To) and respective weight.

From	→	To	Weight
I-DateTime	→	B-DateTime	-3.082
I-Modifier	→	B-Modifier	-2.599
I-Evolution	→	B-Evolution	-1.891
B-Modifier	→	B-Modifier	-1.849
B-Value	→	B-Value	-1.791
B-Evolution	→	O	-1.607
B-Modifier	→	B-Lateralisation	-1.573
I-Evolution	→	B-Modifier	-1.512
I-Evolution	→	O	-1.298
I-Modifier	→	B-Evolution	-1.257

5.2.1.2 Feature Weight Analysis

As stated in Section 5.2, it is possible to examine how each feature influences the clinical entity prediction process. A positive weight implies that the considered feature is strongly correlated to the prediction of the corresponding tag. Contrarily, a negative weight reveals the features that least impacted the prediction. The weights for each feature were summed, so that it is possible to assess which features have more influence in the CRF model prediction.

The features with more weight in the CRF model prediction are presented in Table 5.5.

Table 5.5: Ten most meaningful features in CRF model prediction.

Feature	**Weight**
Token lowercase	749.39
Lemma	695.34
Token	597.74
Token 3-character prefix	562.03
Token 3-character suffix	377.49
Previous token lowercase	324.93
Previous token	318.16
Subsequent token	261.68
Subsequent token lowercase	256.77
Previous token prefix	236.15

Besides the token itself, the lowercase version of the token, the lemma and the prefix and suffix analysis considering 3 characters are the features with more influence on the model predictions. The results imply that features related to the specific token being analysed, mostly linguistic, have a greater weight in the prediction. The previous and subsequent tokens and lowercase tokens also belong to the features with more weight in the CRF model, with the previous token having slight higher weight values for both features than the subsequent word. Lastly, the previous word prefix is also included in the top ten features that more strongly influence the model prediction.

On the other hand, Table 5.6 indicates which ten features have less weight and, therefore, less influence in the CRF model prediction. Analysing the results, it can be concluded that the previous and subsequent token length do not add meaningful information to the model. The same can be concluded to the boolean feature that returns if the first character in a token is uppercase, which has low weights for the current, the previous and subsequent tokens. Considering the subsequent token, POS, lemma and the boolean variable that shows if the token is digit are part of the ten lower feature weights. This indicates that the subsequent token is less considered in the model prediction than the previous and current tokens. However, there are two boolean features in the ten features with less weight relative to the current token - the features that show if all characters in the token are uppercase or if the token is a digit, having a minimal influence on the CRF model prediction.

Table 5.6: Ten least meaningful features in CRF model prediction.

Feature	Weight
Subsequent token length	-4.91
Token is uppercase	-2.79
Previous token length	-2.72
Previous token first character is uppercase	-2.44
Token first character is uppercase	-0.77
Subsequent token POS	-0.72
Subsequent token lemma	-0.70
Subsequent token is digit	-0.64
Subsequent token first character is uppercase	-0.20
Token is digit	-0.04

5.3 Classification with BiLSTM

The Bidirectional LSTM implementation was carried out using Keras[16], an API for high-level neural networks, in version 2.3.1. Keras has a Tensorflow backend, and it is an intuitive way to create the neural networks, easily selecting the desired layers to include.

As previously stated, the annotated data is divided into 5 thousand sentences. For the implementation, it is fundamental to have sentence length into account, since all model inputs may have the same length. Analysing the data, it was concluded that the maximum sequence length is 80. Afterwards, each sentence was preprocessed using a padding strategy, so that all sentences have the maximum sentence length. The tokens used for the padding are at the end of the sentence. The padding process is performed both in the model inputs and the output. The value used in each input and in the output is carefully selected, because it cannot be the same as any token found in the sentences. The value *"ENDPAD"* is used for padding since it cannot be found in the vocabulary.

An illustration of the implemented BiLSTM model architecture with the corresponding inputs and outputs for the clinical entity extraction task is exhibited in Figure 5.3. The BiLSTM model used three different inputs, with information related to each token and padded to the maximum sentence length. The inputs are the token, the corresponding lemma, and the POS tag. The last two inputs were computed using the lemmatiser ans POS tagger methods from the Spacy Portuguese model, which is trained in general domain text from news.

Both the input data and the labels are converted from a categorical format to numerical data, by performing integer encoding in all variables. Integer encoding results in a dictionary, in which each unique token has one corresponding integer value. Subsequently, one-hot encoding is applied to the labels [117], resulting in binary class matrixes for each sentence. The input integer encodings are inputs to the embedding layers in the model. For the token and lemma inputs, a matrix of weights is generated using the FastText model described in Section 4.3. The

[16]"Keras API": `https://keras.io/s`.Accessed on 20-06-2021

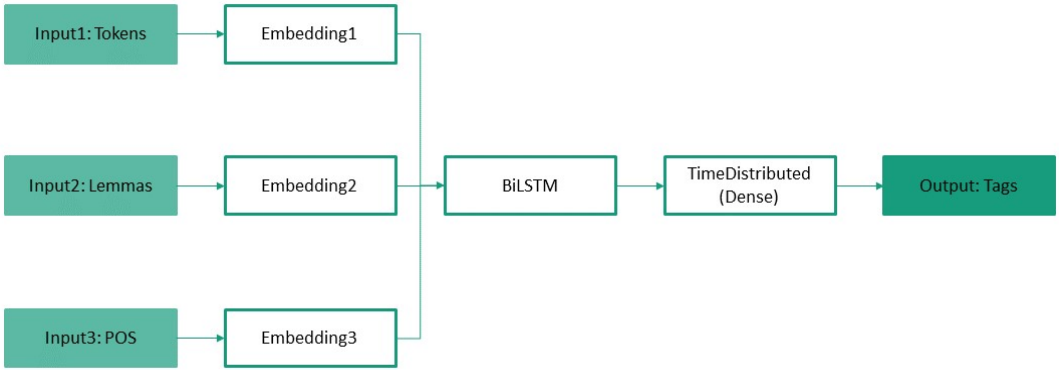


Figure 5.3: BiLSTM model architecture illustration.

corresponding embedding layers is seeded with the FastText embedding weights. On the other hand, the embedding layer for the POS tag input does not use pre-trained WE, but updates the weights during the training.

The hyperparameters used in the BiLSTM model are summarised in Table 5.7. The number of LSTM units is the same as the WE vector size. The dropout is a technique that aims to regularise the training and avoid overfitting. It refers to the number of randomly neural network units that are left out of the training process. The dropout percentage is set in 50%, so that overfitting is avoided. The Adam optimisation function was the selected optimiser [118]. Adam adapts the learning rate to the average first model, as well as to the second moments of the gradient descend, i.e., considering the mean and variance. The learning rate is set to 0.0005. A low learning rate is preferred since it favours gradient convergence, assuring neither any local minimums are missed, nor gradient descent overshoots.

Table 5.7: Hyperparameters used in BiLSTM training

Hyperparameter	**Value**
LSTM Units	300
Dropout Percentage	50%
Optimiser	Adam
Learning Rate	0.0005

The number of epochs was not determined in advance of model training. Instead, an early stopping method was implemented. For that, a considerably large number of epochs was defined, and the model training stopped after 20 epochs with no validation F1-score variance greater than 0.5%. Early stopping helps avoid model overfitting to the training set, such as when an arbitrary substantial number of epochs is fixed, or the opposite problem of underfitting for a reduced epochs number.

5.3.1 Results and Discussion

The comparison between the mean of the macro F1-scores obtained for the validation set in each 5 folds and the macro F1-score for the test set is exhibited in Figure 5.4. There is a decay on the F1-score for the test set. However, the difference between the two F1-score values is not considerable, since it is still in the range of the validation set standard deviation. Therefore, it can be stated that there is no overfitting in the model training process.

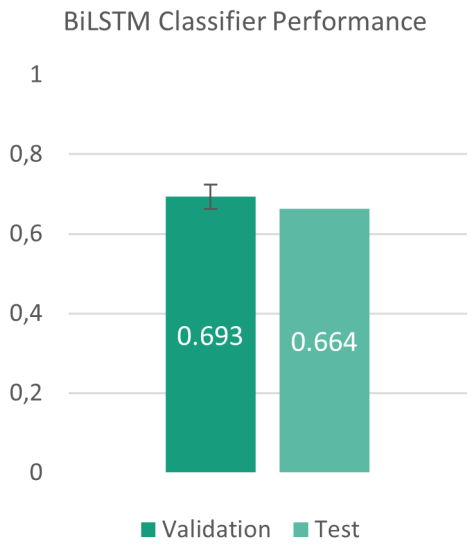


Figure 5.4: BiLSTM model performance assessment using macro F1-score for test and validation sets.

The results per tag, as well as the average results, considering precision, recall and F1-score as evaluation metrics are presented in Table 5.8.

Similarly to the previous results, the Modifier and Evolution entities have the worst F1-scores out of all entities analysed, in accordance with the validation annotations. The fact that the vocabulary and sentence construction for these entities varies more than in other entities compromises the model performance.

The F1-score values have almost entirely decreased, with only one exception that maintained the F1-score. Considering the I tags for all entities, and comparing them with the results presented in the previous section, the respective F1-scores decreased more considerably than the B tags evaluation metrics. These results reinforce the inability of the BiLSTM model to understand the logic behind IOB tagging. The Negation entity also suffered a considerable decrease for a F1-score of 0.78, comparing with 0.91 for the CRF model. Oppositely, B-Condition F1-score only had a slight decrease to 0.76, and B-AnatomicalSite maintained the 0.78 F1-score.

Overall, the results are lower than those obtained using the CRF model. Comparing the macro F1-score, it decreased from 0.72 for CRF to 0.66 in the BiLSTM implementation. BiLSTM is composed by two LSTM layers, each interpreting the input in a forward and backward direction. The additional context resultant from the bidirectional architecture can improve model

Table 5.8: Results for clinical entity extraction task using BiLSTM model. For each evaluation metric, best results in bold and worst results underlined.

	Precision	**Recall**	**F1-Score**
B-AnatomicalSite	0.80	0.77	0.78
I-AnatomicalSite	0.68	0.44	0.53
B-Appointment	0.73	0.69	0.71
I-Appointment	0.76	0.57	0.65
B-Condition	0.80	0.72	0.76
I-Condition	0.65	0.54	0.59
B-DateTime	0.78	0.57	0.66
I-DateTime	0.84	0.73	0.78
B-Evolution	0.52	0.22	0.31
I-Evolution	0.52	0.48	0.50
B-Exam	0.82	0.73	0.77
I-Exam	0.74	0.49	0.59
B-Lateralisation	0.83	**0.85**	**0.84**
B-Modifier	0.57	0.49	0.53
I-Modifier	0.55	0.40	0.46
B-Negation	0.83	0.73	0.78
B-Therapeutic	**0.88**	0.70	0.78
I-Therapeutic	0.72	0.63	0.67
B-Value	0.82	0.79	0.81
I-Value	0.83	0.73	0.77
macro avg	0.73	0.61	0.66
weighted avg	0.70	0.59	0.64

performance in sequence classification tasks. However, the predictions are still independent from each other. For that reason, the results do not outperform the CRF model.

5.4 Classification with BiLSTM-CRF

Throughout this chapter, the capacity of CRF models to perform a sequence prediction has been described as important in the clinical entity extraction task. The results presented for CRF and BiLSTM, in Sections 5.2 and 5.3, respectively, corroborate the CRF layer importance. The fact CRF obtained better results than the BiLSTM neural network model can be associated to the ability to understand the sequential logic behind IOB tagging.

For that reason, based on the neural network developed in the previous section, a BiLSTM-CRF model was implemented. The added CRF layer was implemented using the TensorFlow Addons repository[17], in version 0.13.0.

The hyperparameters used for training the BiLSTM layer are the same as presented in Table 5.7 of the previous section. Besides, the early stopping method is also implemented. That way, in each fold, the model training ends when the validation macro F1-score does not vary more than 0.5% during 20 epochs. The number of units in the CRF layer is equal to the number of IOB tags representing the clinical entities. The described neural network is illustrated in Figure 5.5.

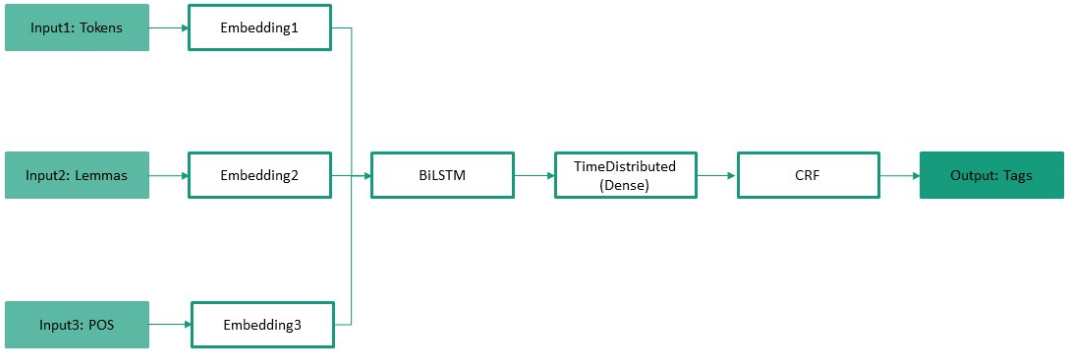


Figure 5.5: BiLSTM-CRF model architecture illustration.

5.4.1 Results and Discussion

The bar chart in Figure 5.6 shows the macro average F1-scores for the validation and test sets. The similarities between values, and the fact that the test F1-score value is inside the standard deviation range corroborates the conclusion that there is no overfitting in the BiLSTM-CRF model training.

[17] "Tensorflow Addons repository": `https://www.tensorflow.org/addons`.Accessed on 09-06-2021

Figure 5.6: BiLSTM-CRF model performance assessment using macro F1-score for test and validation sets.

Table 5.9 summarises the results obtained for the clinical entity extraction task using the BiLSTM-CRF implementation.

The macro and weighted F1-scores prediction results for the the BiLSTM-CRF model have increased when compared to the results of the BiLSTM architecture. Considering the evaluation scores of the I tags, the majority achieved a better F1-score. This excludes I-Condition and I-Appointment, that maintained the F1-score for both approaches, and I-Therapeutic, that decreased from 0.67 to 0.64. The tags with highest F1-score increase are B-Negation, with an increase of 0.14 over the CRF F1-score value, and I-Value, with a 0.13 F1-score increase.

Similarly to the results obtained for the previous analysed approaches, and in concordance with the annotation validation, the Modifier and Evolution entities obtain the worst F1-scores, both for B and I tags.

It is important to address that, although the results increased when comparing with the BiLSTM architecture implementation, a decrease in the evaluation metrics values is observed when comparing with the CRF model. It would be expected that the hybrid approach, with a CRF layer on top of the BiLSTM architecture, would have better results, which does not occur. The F1-scores discrepancy may be due to the different inputs used in the two models. In the CRF model, all the features listed in Section 5.2 were considered in the CRF training, characterizing each token and the neighbour tokens as well, in terms of linguistic and morphological features. On the other hand, in the BiLSTM-CRF architecture implementation, only linguistic features as tokens, lemma and POS were used as input to the model. Additionally, the prefix and suffix information, that had a high weight in the CRF architecture, are in a way considered in the BiLSTM-CRF prediction because of the FastText model used for WE. Nevertheless, a direct comparison of the results obtained for both approaches is not feasible, because of the difference in inputs utilised for both approaches.

Comparing the BiLSTM-CRF implementation outcomes with the results found in the literature for Portuguese clinical entity extraction, Lopes *et al.* [88] achieved a macro F1-score of 0.794±0.021. Once again, the discrepancy in performance when compared with the obtained results can be justified with the fact that the input data in [88] included morphological information from each token, that was not considered in the BiLSTM-CRF architecture implemented in this dissertation work.

Table 5.9: Results for clinical entity extraction task using BiLSTM-CRF model. For each evaluation metric, best results in bold and worst results underlined.

	Precision	**Recall**	**F1-Score**
B-AnatomicalSite	0.78	0.72	0.75
I-AnatomicalSite	0.74	0.58	0.65
B-Appointment	0.83	0.84	0.83
I-Appointment	0.67	0.46	0.55
B-Condition	0.82	0.73	0.77
I-Condition	0.69	0.51	0.59
B-DateTime	0.77	0.57	0.66
I-DateTime	0.79	0.76	0.77
B-Evolution	0.55	<u>0.15</u>	<u>0.24</u>
I-Evolution	<u>0.54</u>	0.57	0.55
B-Exam	0.82	0.72	0.77
I-Exam	0.74	0.52	0.61
B-Lateralisation	0.79	0.93	0.85
B-Modifier	0.58	0.53	0.55
I-Modifier	0.56	0.40	0.47
B-Negation	0.89	**0.96**	**0.92**
B-Therapeutic	**0.91**	0.80	0.85
I-Therapeutic	0.80	0.53	0.64
B-Value	0.89	0.85	0.87
I-Value	0.83	0.89	0.86
macro avg	0.75	0.65	0.69
weighted avg	0.72	0.63	0.66

5.5 Classification with BERT

BERT models are Transformer-based large neural network architectures, with a great amount of trainable parameters. The model pre-training requires a large dataset. Otherwise, the pre-training in a small dataset would most likely result in overfitting. Considering that the annotated

Dermatology corpus only contains 5 thousand sentences, it would not be sufficient for the pre-training process. Therefore, the adopted approach was model fine-tuning.

The fine-tuning of a pre-trained BERT-based model to a certain task results from adding an additional output layer to a pre-trained model. That way, the number of learned parameters are minimal when comparing with model pre-training [82]. In NER tasks, the additional output layer is a projection of the token hidden state size to the number of entities. A succeeding softmax operation converts the outputs into probabilities. Figure 5.7 illustrates the BERT fine-tuning process to a NER task. The [CLS] token included in the beginning of each sentence, and the [SEP] token at the end of each sentence (not included in the illustration because a single sentence is given as input) are not particularly relevant for the NER taks, since no tag is associated to these tokens. Nevertheless, these two particular tokens should be included in the fine-tuning input, so that it does not differ from the pre-training input.

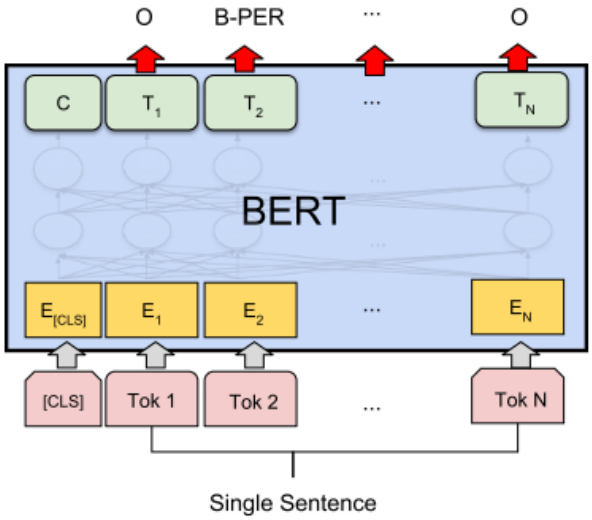


Figure 5.7: BERT fine-tuning for NER task using single sentence as input. Adapted from [82].

The BERT base models fine-tuning was performed using the NERDA package[18]. NERDA supports the fine-tuning of transformers for NER tasks specifically. The fine-tuning can be carried out utilising an already pre-trained transformer model available in the HuggingFace[19] repository.

In the context of the developed work, three pre-trained BERT based models were fine-tuned. The models choice takes into account the language in which the corpus is written. In this case, as the clinical notes are written in Portuguese, the appropriate models for the clinical corpus should be pre-trained in Portuguese, or in multiple languages, resulting in a multilingual BERT model. For those reasons, the three selected architectures are the following:

- **BERT Base Multilingual Cased (BERTmulti cased) -** pre-trained BERT architecture using a large multilingual general domain corpus composed by the top 104 languages with

[18] "NERDA Python package": `https://pypi.org/project/NERDA/`.Accessed on 02-06-2021
[19] "HuggingFace repository": `https://huggingface.co/`.Accessed on 20-06-2021

largest Wikipedia repositories. In cased models, the tokens uppercase and lowercase characters are maintained in the model training[20].

- **BERT Base Multilingual Uncased (BERTmulti uncased) -** pre-trained BERT architecture using a large multilingual general domain corpus composed by the top 102 languages with largest Wikipedia repositories. In uncased models, the tokens are lowercased before the model pre-training[21].
- **BERT Base Portuguese Cased (BERTpt) -** pre-trained BERT architecture for Brasilian Portuguese trained on the general domain Brasilian Web as Corpus dataset (BrWaC) [109].

The BERT model size can be Base or Large. As mentioned, all three chosen BERT based models are defined as Base models. BERT Base models have 12 layers, 12 self-attention heads and the hidden state size is 768 per token. In this architectures, the total trainable parameters are 110 million. In contrast, BERT Large architectures are composed by 24 layers, 16 self-attention heads and hidden state size of 1024, resulting in 340 million trainable parameters. As stated in Section 3.4.1.3, in [108], the BERTpt Large model achieved worse F1-score results for both datasets in the NER task. Considering this factor, as well as the computational cost associated with the BERT Large architecture, only BERT Base architectures were utilised in the dissertation work.

The same hyperparameters were used in the fine-tuning of the three selected BERT-based models, and are dispalyed in Table 5.10. According to Devlin *et al.* [82], the recommended number of epochs for model fine-tuning is 4. The training warmup steps are defined to 500, meaning that for 500 steps in the beginning of the fine-tuning process the learning rate is lower. This allows for the network to gradually adapt to the training data. The training batch size hyperparameter defines the number of samples that is propagated through the network at the time. The recommended values for training batch size in [82] are 16 and 32. The advantages of smaller training batch sizes are less memory used for the training process and less training time. For that reason, the training batch size of 16 was selected. The learning rate hyperparameter controls the rate at which the model adapts to the training data. Lower learning rate values lead to less weight changes in each update, and therefore more training epochs, that favour gradient convergence. The selected learning rate of 0.0001 was the default value for the NERDA fine-tuning model.

Table 5.10: Hyperparameters used in BERT-based models fine-tuning.

Hyperparameter	**Value**
Epochs	4
Warmup Steps	500
Train Batch Size	16
Learning Rate	0.0001

[20] "BERT multilingual cased model": `https://huggingface.co/bert-base-multilingual-cased`.Accessed on 20-06-2021

[21] "BERT multilingual uncased model": `https://huggingface.co/bert-base-multilingual-uncased`.Accessed on 20-06-2021

5.5.1 Results and Discussion

A comparison between the macro F1-scores for the validation and test sets is an important step to assess the model ability to generalise the results to any set used as input. Figure 5.8 depicts that comparison for BERTmulti uncased, BERTmulti cased and BERTpt, respectively. No considerable discrepancy is observed between the validation and test F1-scores. Hence, it can be concluded that no overfitting occurred in the fine-tuning of any of the models.

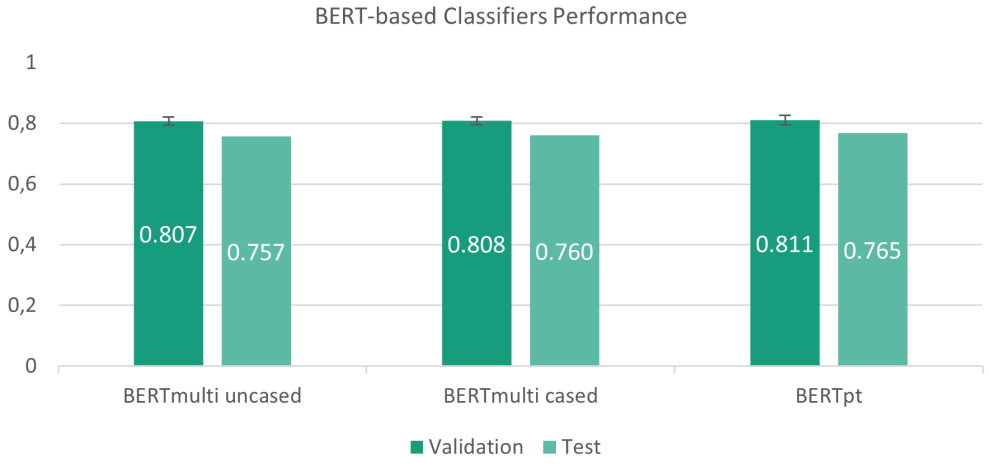


Figure 5.8: BERT-based models fine-tuning assessment using macro F1-score for test and validation sets.

As previously stated, a training and test set is extracted for each five folds computed. For each fold, a model is trained using the training set and afterwards, a tag prediction is performed for the test set. Hence, the entire annotated dataset is used as test set as consequence of the K-Fold cross-validation performed, and the evaluation metrics comprise the predictions of the five trained models.

The classification reports obtained for the fine-tuning of BERT multilingual uncased, BERT multilingual cased, and BERTpt, using the annotated clinical corpus are presented in Tables 5.11, 5.12, and 5.13, respectively. An analysis of the results obtained for the three BERT-based model allows to conclude that, although there are slight variations in the evaluation metrics for each tag, the average F1-scores computed for the three models are similar. Along with Figure 5.8, it is concluded that BERTpt holds the greater F1-score value of 0.765, followed by BERTmulti cased with 0.760, and lastly, BERTmulti uncased with an F1-score of 0.757.

Although the Modifier and Evolution entities maintain the tendency of previously analysed approaches and are the entities with the lowest values in the evaluation metrics, there is a considerable increase in the evaluation metrics values for these entities, which positively influences the overall performance of the fine-tuned models.

Analysing the evaluation metrics for certain entities that will subsequently have a greater influence on the risk prioritisation task helps assessing and comparing the performance of each BERT-based model. For both B and I tags of the Anatomical Site entity, the highest F1-score values of 0.86 and 0.71, respectively, are obtained in the fine-tuning of BERTpt. Higher values

Table 5.11: Results for clinical entity extraction task using BERT multilingual uncased model. For each evaluation metric, best results in bold and worst results underlined.

	Precision	**Recall**	**F1-Score**
B-AnatomicalSite	0.83	0.87	0.85
I-AnatomicalSite	0.83	0.58	0.68
B-Appointment	0.78	0.82	0.80
I-Appointment	**0.90**	0.64	0.75
B-Condition	0.80	0.79	0.80
I-Condition	0.71	0.62	0.66
B-DateTime	0.81	0.84	0.83
I-DateTime	0.88	0.88	0.88
B-Evolution	0.53	0.54	0.54
I-Evolution	0.65	0.60	0.62
B-Exam	0.84	0.81	0.82
I-Exam	0.80	0.57	0.67
B-Lateralisation	0.82	0.91	0.87
B-Modifier	0.59	0.65	0.62
I-Modifier	0.63	0.54	0.58
B-Negation	0.87	**0.96**	**0.92**
B-Therapeutic	0.86	0.90	0.88
I-Therapeutic	0.83	0.69	0.76
B-Value	0.85	0.90	0.88
I-Value	0.76	0.73	0.75
macro avg	0.78	0.74	0.76
weighted avg	0.75	0.73	0.74

are also obtained for the tags B-Condition and I-Exam, reaching F1-score values of 0.81 and 0.69. The results for the Evolution entity are the equal for the three BERT models. Still, when compared with the other implemented approaches, a substancial increase is observed, from B-Evolution and I-Evolution F1-score values of 0.31 and 0.50 in the BiLSTM implementation, to 0.54 and 0.62 in all fine-tuned BERT models.

Concerning the results found in the literature for the fine-tuning of the three chosen BERT models, Schneider *et al.* [108] analysed the BERT based models fine-tuning outcomes in two different clinical domain text datasets. One of the datasets included 1 thousand notes from several medical specialties, while the other dataset was composed by 281 Neurology notes annotated in [87], [88] and made publicly available. For the multiple specialty corpus, the macro average F1-scores obtained with BERTmulti uncased, BERTmulti cased and BERTpt were 0.588, 0.582 and 0.585, respectively. The Neurology corpus achieved macro F1-scores of 0.912, 0.921 and 0.916, re-

Table 5.12: Results for clinical entity extraction task using BERT multilingual cased model. For each evaluation metric, best results in bold and worst results underlined.

	Precision	**Recall**	**F1-Score**
B-AnatomicalSite	0.82	0.86	0.84
I-AnatomicalSite	0.81	0.60	0.69
B-Appointment	0.79	0.80	0.80
I-Appointment	0.88	0.65	0.74
B-Condition	0.81	0.79	0.80
I-Condition	0.71	0.63	0.67
B-DateTime	0.83	0.84	0.84
I-DateTime	**0.89**	0.88	0.88
B-Evolution	0.53	0.55	0.54
I-Evolution	0.64	0.61	0.62
B-Exam	0.84	0.82	0.83
I-Exam	0.80	0.59	0.68
B-Lateralisation	0.82	0.91	0.86
B-Modifier	0.60	0.67	0.63
I-Modifier	0.64	0.55	0.59
B-Negation	0.88	**0.97**	**0.92**
B-Therapeutic	0.85	0.90	0.88
I-Therapeutic	0.81	0.72	0.76
B-Value	0.84	0.90	0.87
I-Value	0.78	0.72	0.75
macro avg	0.78	0.75	0.76
weighted avg	0.75	0.73	0.74

spectively. Comparing these results with the ones obtained in the scope of the dissertation, there is an inverse relationship between the number of clinical cases used in the fine-tuning process and the obtained results: the greater the number of clinical cases used, the lower the evaluation metrics obtained. In addition, the inclusion of texts from various medical specialties lead to the presence of terms and expressions from different medical domains in the vocabulary, and may undermine the model prediction. Besides, the Neurology corpus is not representative of EHR clinical notes because the cases were extracted from a journal. Therefore, the corpus does not include spelling errors that are very common in EHR notes.

The results obtained for the Dermatology annotated corpus for the three implemented BERT models are similar to the results obtained in [108]. The results are intermediate values when compared to the results for both the Neurology and multiple specialty datasets. When compared with the results obtained for the other implemented approaches, the BERT-based models achieve

Table 5.13: Results for clinical entity extraction task using BERTpt base model. For each evaluation metric, best results in bold and worst results underlined.

	Precision	**Recall**	**F1-Score**
B-AnatomicalSite	0.85	0.87	0.86
I-AnatomicalSite	0.84	0.62	0.71
B-Appointment	0.77	0.81	0.79
I-Appointment	**0.89**	0.62	0.73
B-Condition	0.81	0.80	0.81
I-Condition	0.70	0.62	0.66
B-DateTime	0.83	0.83	0.83
I-DateTime	0.87	0.88	0.88
B-Evolution	<u>0.54</u>	<u>0.54</u>	<u>0.54</u>
I-Evolution	0.64	0.60	0.62
B-Exam	0.84	0.82	0.83
I-Exam	0.79	0.61	0.69
B-Lateralisation	0.85	0.87	0.86
B-Modifier	0.59	0.66	0.62
I-Modifier	0.64	0.56	0.60
B-Negation	0.88	**0.97**	**0.92**
B-Therapeutic	0.87	0.91	0.89
I-Therapeutic	0.85	0.72	0.78
B-Value	0.85	0.90	0.87
I-Value	0.77	0.75	0.76
macro avg	0.78	0.75	0.76
weighted avg	0.76	0.73	0.74

the highest evaluation metrics values. The results are, therefore, congruent with the literature, that affirm that BERT models are the state-of-the-art approach for many NLP tasks, such as entity extraction.

Although the performance of the three analysed BERT models is similar, BERTpt is more precise in extracting certain tags that strongly correlate with the risk prioritisation, such as Anatomical Site and Condition, besides having a slight higher macro F1-score. Hence, the fine-tuned BERTpt model is selected to extract the clinical entities from the remaining 12,058 texts from the NHS dataset. The extracted information is used as input in the risk prioritisation model described in the following chapter.

It is important to mention that there is a BERT-based model pre-trained with clinical domain text in Portuguese, named BioBERTpt, introduced by Schneider *et al.* [108]. As explained in Section 3.4.1.3, BioBERTpt is pre-trained with Portuguese clinical domain text, being the most

adequate BERT based model in the scope of the developed work. Although it is available in the HuggingFace repository, the BioBERTpt model pre-trained with the 3 million clinical notes is not fine-tuned for the NER task. Therefore, it is not possible to use BioBERTpt as a model in NERDA.

5.6 Comparison of Clinical Entity Extraction Approaches

In order to assess which one of the approaches revealed a better performance in the clinical entity extraction task, it is essential to compare the results obtained for each implemented approach. This comparison is important, since it is necessary to select the model with the best performance to extract the clinical entities from the clinical notes that were not manually annotated, that will be utilised in the risk prioritisation task described in Chapter 6. Table 5.14 summarises the test macro average scores for each of the implemented models.

Table 5.14: Macro average scores for clinical entity extraction task for the implemented models.

Model	Precision	Recall	F1-Score
CRF	0.75	0.69	0.72
BiLSTM	0.73	0.61	0.66
BiLSTM-CRF	0.75	0.65	0.69
BERTmulti uncased	0.78	0.74	0.76
BERTmulti cased	0.78	0.75	0.76
BERTpt	0.78	0.75	0.76

Being a shallow ML approach, in which there is a feature selection step, CRF can be considered as the baseline model, achieving a macro F1-score of 0.72.

The BiLSTM architecture, on the other hand, is a DL approach. It was expected that the performance for this model would surpass the CRF model. One reason for the lower F1-score of 0.66 for this implemented model, when compared to the CRF, could be the fact that, although BiLSTM is able to extract context information bidirectionally in the sentence, the tag predictions are independent of each other, as the logic of IOB tag sequences is not learned, unlike in the CRF model. However, when a CRF layer was added on top of the BiLSTM architecture, the evaluation metrics increased, achieving a F1-score value of 0.69, but did not surpass the CRF outcome. Comparing the inputs for the CRF and BiLSTM models, it was concluded that the difference in the inputs may explain why better results are obtained using the shallow ML model. The BiLSTM and BiLSTM-CRF architectures only considered the token, lemma and POS as inputs. However, when analysing the most meaningful features for the CRF model in Table 5.5, the POS is not one of the most meaningful features, and although the FastText model extracts information about prefix and suffix to a certain extent, that information is not directly used as input for the models. Besides, features related to previous and subsequent tokens are simply not considered. Therefore, in order to directly compare the results for CRF, BiLSTM and

BiLSTM-CRF, the features used in the CRF implementation should also be used as input for the BiLSTM architectures.

The BERT-based architectures reveal the best outcomes in the clinical entity extraction task. Analysing the results for the three BERT models fine-tuning, the macro F1-scores are similar. There is not a considerable difference in results when comparing the models trained multilingual or Portuguese corpora. Observing Figure 5.8, it can be concluded that BERTpt outperforms BERTmulti cased by a margin of 0.005 i the F1-score value. Although the difference is residual, it was expected that BERTpt revealed the best outcome out of the three BERT-based fine-tuned models, since it is pre-trained only with general domain Portuguese text. Given the fact that all BERT architectures fine-tuned in this work were pre-trained with general domain corpora, there is scope for improvement if using a BERT model pre-trained with clinical text, such as BioBERTpt [108], which was pre-trained with Portuguese clinical notes. Although it is available on HuggingFace, it was not possible to use BioBERTpt because it is not fine-tuned for the NER task.

5.7 Summary

In the course of this chapter, the adopted approaches for the clinical entity extraction task were outlined. The choice of which algorithms to implement was based on the literature review presented in Chapter 3, with particular focus on the clinical NLP applications for information extraction in Portuguese text.

The beginning of the chapter starts with some considerations regarding the workflow of the implementations. In order to compare the results, four ML and DL architectures were selected as the approaches to implement. The chosen architectures were CRF, BiLSTM, BiLSTM-CRF, and three BERT architectures, two of them pre-trained with multilingual corpora and one pre-trained with general domain Portuguese text. For each approach, theoretical considerations and the chosen hyperparameters utilised in the implementation were presented, followed by the results and discussion regarding that approach, including the obtained outcomes and the expected outcomes considering the results found on the literature.

BERTpt achieved the best macro F1-score of 0.765, an outcome that is in accordance to the literature. BERTpt will be used in for the entity extraction of the remaining non-annotated clinical notes, that will be used as input for the risk prioritisation model, which is addressed in the next chapter.

Chapter 6

Dermatology Cases Risk Prioritisation

Throughout this chapter, two strategies for risk prioritisation of Dermatology cases are introduced, using both structured and unstructured information included in the Dermatology clinical records. This task can be interpreted as a proof of concept, which aims to demonstrate the influence that textual information extracted from clinical notes has in the case risk prioritisation algorithm developed in the Derm.AI project.

In the first strategy, the data utilised in the implemented risk prioritisation model are 12,058 clinical notes that were not manually annotated. The data clinical entities are extracted utilising the BERTpt model finetuned in Chapter 5. The risk prediction is carried out using the tokenised clinical text, the corresponding clinical entities and the ICD-9 code assigned to the case as inputs to an LSTM neural network. The data imbalance in terms of priority values may cause issues in the risk prioritisation prediction. For that reason, data-balancing techniques are applied to the training data, and the results are compared with the imbalanced dataset results.

In the second strategy, a separate dataset with 3,428 clinical cases that include additional information from clinical images is utilised, and the cases are divided into 13 differential diagnoses with a corresponding associated risk. For each clinical case, the clinical entities are extracted using the fine-tuned BERTpt model, and the parts of the notes tagged with the entities Anatomical Site, Evolution and Value are used as input to the risk prioritisation model implemented previously in the Derm.AI project to assess if the clinical entity extraction has a positive impact on the risk prioritisation task.

6.1 Risk Prioritisation Algorithm

Considering the information included in the NHS Dermatology dataset, previously described in Section 4.1, the prioritisation of a clinical case depends on numerous factors. Taking into account the structured information, the ICD-9 code assigned to each clinical case, representing a given diagnosis, includes significant information about a certain condition in a simpler representation. Besides, the entities from the clinical entity extraction task described in the previous chapter are

fundamental to classify each case and understand its risk and, therefore, assign a value in the priority scale.

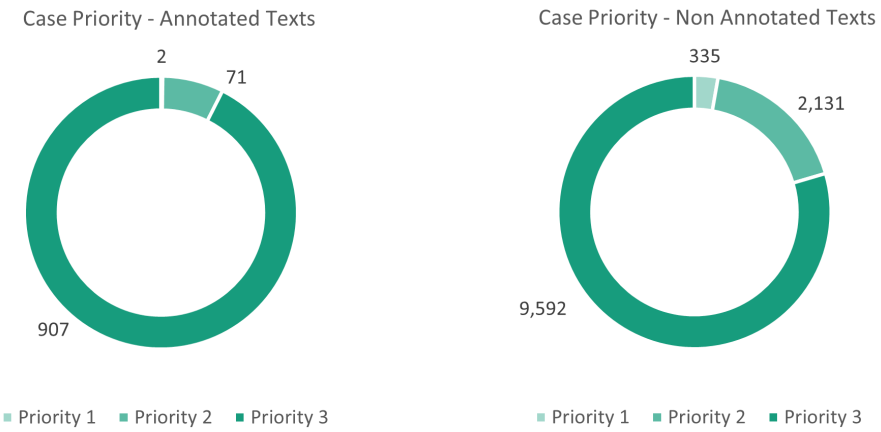
The different lengths of the textual information for each clinical case are a limitation, since the model inputs must have the same length. For that reason, and similarly to the previous task, each case is divided into corresponding sentences. Each sentence belonging to the same clinical case has the corresponding ICD-9 code and priority value. This problem can therefore be interpreted as a multiclass sentence classification task, in which the model predicts the risk prioritisation for each sentence.

In Chapter 4, the risk priority values distribution was analysed. On a scale from 1 to 3, a priority of 1 represents high risk priority clinical cases, whereas 3 represents low priority cases. It was observed that the case priority distribution was extremely unbalanced, with 337 extreme priority cases, 2205 intermediate priority cases, and 10,516 low priority cases, as shown in Figure 4.1.

Analysing the case priority distribution for the 980 annotated Dermatology clinical cases, approximately 93% of the cases are assigned with the lower priority value represented by number 3, and only 2 cases have the higher risk priority value of 1, which is a residual value, as observed in Figure 6.1a. Hence, the annotated dataset exhibits a larger class imbalance when compared with the entire dataset.

The strategy adopted to surpass the class imbalance problem in the annotated cases was to use 12,058 non-annotated clinical notes.The clinical entities used as input for the prioritisation model are extracted from the notes using a clinical entity extraction model from the ones implemented in the previous chapter. In accordance with the results obtained in Chapter 5, the BERTpt fine-tuned model was chosen, because it obtained a macro F1-score of 0.765, the best result for the clinical entity extraction task. The BERTpt model was fine-tuned using the 5 thousand annotated sentences, and afterwards, a clinical entity prediction was performed for a total of 127,564 sentences comprised in the 12,058 clinical cases. In Figure 6.1b, it is observable that the general class distribution in the non-annotated clinical cases is still quite imbalanced. The same analysis in terms of priority value per sentence reveal 117,262 sentences with lower priority, 9,515 with intermediate priority, and 787 sentences corresponding to cases with higher priority. Nonetheless, the distribution is less imbalanced in terms of case prioritisation than the annotated corpus and has includes more clinical cases as data. For these reasons, it is favourable to use this dataset after extracting the clinical entity as input for the risk prioritisation model.

According to the literature, the most common approaches for sentence and document classification are LSTM, CNN and BiLSTM-CNN [119]–[121]. For the purpose of the dissertation work, only the LSTM architecture was implemented in the context of the risk prioritisation task. Figure 6.2 illustrates the model architecture, with the tokens, tags and ICD-9 codes as inputs, and the risk priority value as output. Similarly to the DL architectures in the previous chapter, the inputs are converted to a numerical representation using integer encoding. The conversion results in dictionaries in which each unique instance for tokens, tags and ICD-9 codes have a corresponding integer value. Additionally, the output, composed of the priority values, was converted to a binary matrix of classes through one-hot encoding [117]. The embedding layer for the sentence tokens utilises once again the matrix of weights resulting from the FastText algorithm. Besides, all sentences have undergone a padding process so that the length is always equal to the



(a) Annotated clinical cases. **(b)** Non-annotated clinical cases.

Figure 6.1: Dataset distribution according to case priority for annotated and non-annotated NHS dermatology clinical cases.

maximum length of a sentence, which is 90 for the non-annotated corpus. The hyperparameters defined for the LSTM layer are the same as the ones used in the BiLSTM layer in the previous task, displayed in Table 5.7.

Similarly to the BiLSTM and BiLSTM-CRF approaches carried out for the clinical entity extraction task, an early stopping algorithm was implemented. For this task, it was defined that the model training stops after 10 epochs without an increase on the validation macro F1-score greater than 0.5%.

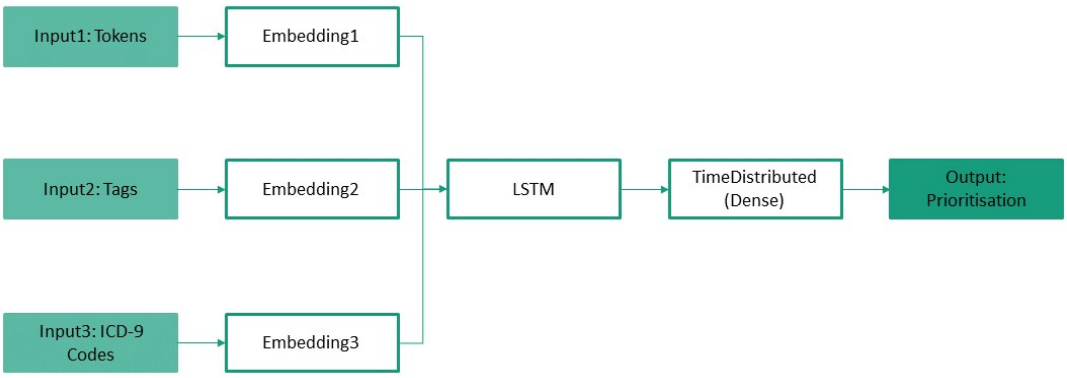


Figure 6.2: LSTM model architecture illustration for risk prioritisation task.

The data available for this task is much larger than the 5 thousand annotated sentences used in the clinical entity extraction task. In total, there are 127,564 sentences available for the risk prioritisation task. For computational efficiency manners, the K-fold cross-validation technique was not implemented, and a separate test set was defined. In order to split the data into training, validation and test sets, a stratified shuffle split cross-validation was applied. This way, the proportion of samples of each class is maintained in each set, making sure that even with the observed class imbalance there are instances of all classes in all sets. The training set

accounts for 80% of the total sentences in the dataset, whereas both validation and test sets are each composed by 10% of the total number of sentences.

6.1.1 Oversampling and Undersampling Methods

As examined in the previous section, there is a significant class imbalance between the prioritisation classes. This discrepancy can negatively affect the model prediction. If the class imbalance is verified in the training set, the classes with less instances in the training set are less likely to be predicted by the model because are harder for the model to learn.

Data-balancing approaches are used to modify the data distribution, in an effort to achieve a balanced distribution between classes. Oversampling increase the number of instances of minority classes to come close or equal to the number of instances from the majority class. On the other hand, undersampling removes instances from the majority classes. Balancing the dataset distribution is a simple and general approach, since it can be implemented independently of the chosen classifier, and has improved the prediction results in numerous studies [122].

By generating new data examples from the minority classes, oversampling increases the dataset size at the same time it balances the class distribution, without rejecting relevant information. The first oversampling approach to consider is random data replication. In the scope of the dissertation work, this approach replicates specific textual instances of one or more minority classes. The main drawback of random data replication is the overfitting possibility, since only specific parts of the dataset are replicated.

An alternative oversampling method is SMOTE - Synthetic Minority Oversampling Technique [123]. SMOTE generates new data instances from the minority class in the feature space, instead of creating real textual instances. The SMOTE algorithm is illustrated in Figure 6.3, considering two different classes. The majority class is represented by dark green points, whereas the minority class is represented by light blue points. Choosing a data point from the minority class, as the black point in Figure 6.3 b), the k-nearest neighbours from that point are selected [124]. Low k numbers would lead to a greater similarity between the original and the synthetic data points. On the other hand, extending the neighbour points to be considered to a large k values could lead to the synthetic data points being influenced by distant data points. K values between 5 and 10 are considered adequate values for a correct application of the SMOTE algorithm. In this illustration, for simplicity matters, k equals to 3, and the data points corresponding to the nearest neighbours are represented by orange points. For each nearest neighbour, a new artificial data point is created in the line segment that unites the initial data point and the neighbour data points in the feature space, represented by the bright blue data point in Figure 6.3 c).

In contrast to the described oversampling approaches, undersampling achieves a balance in class distribution by decreasing the data instances of majority classes. There are several undersampling algorithms, such as NearMiss methods [125], that implement distance-based algorithms to see what instances in the majority class are closer in the feature space to the minority class instances, so that information loss is not a problem by removing data with meaningful

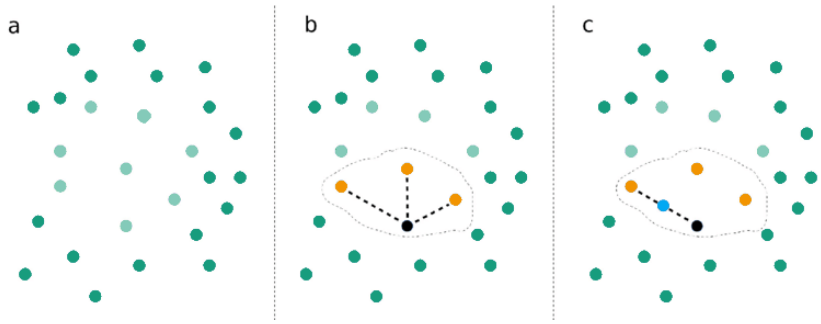


Figure 6.3: SMOTE algorithm illustration. Adapted from [124]

information. A clustering approach can also be used, in which the number of clusters in the majority class is close or identical to the number of data points belonging the minority class [126]. The majority cluster can be represented by the cluster center or using a k-nearest neighbour strategy. The simpler strategy to apply is random undersampling, in which the data to remove is randomly chosen. There is a risk of loss of information, because any data instance can be deleted. However, textual information in the majority class data can be similar because there is common information in most clinical notes. Therefore, due to simplicity and, consequently, less computational cost, random undersampling is the chosen approach.

Having the described data-balancing approaches into account, the adopted strategy in the develop work was using both undersampling and oversampling. There is a considerable discrepancy between the majority class, representing the low priority cases, and the minority class, representing the higher priority cases. The intermediate priority also reveals an intermediate number of instances of 9,515. Therefore, when applying the data-balancing approaches to the training set, the oversampling was only performed in the minority class, and the undersampling was performed in the majority class, so that in the end both had equal data instances to the intermediate class. Therefore, the LSTM architecture for risk prioritisation classification was implemented using both the totality of the 12,058 cases with the case priority distribution illustrated in Figure 6.1b, and using the same data after the implementation of data-balancing methods: the undersampling approach was random undersampling, and both SMOTE and random oversampling were implemented for further comparison.

6.1.2 Results and Discussion

The analysis of the generalisation capacity of the risk prioritisation models is important since undersampling and oversampling methods make the model more prone to overfitting. The bar chart in Figure 6.4 compares the macro F1-scores for the validation and test sets in the three risk prioritisation model implementations. All implemented models reveal a higher F1-score for the validation set. The differences between the validation and test sets for each model are 0.075, 0.187, and 0.117, respectively. The difference is more evident in the models in which oversampling and undersampling were performed, in which the capacity of the models to generalise the predictions is compromised. Since the early stopping method was adopted for the training

to stop when the validation results did not improve, a reason for the discrepancy between validation and test set can be overfitting, as a consequence of the data-balancing techniques that replicate instances from the minority class in textual format or in the feature space, depending on the adopted approach.

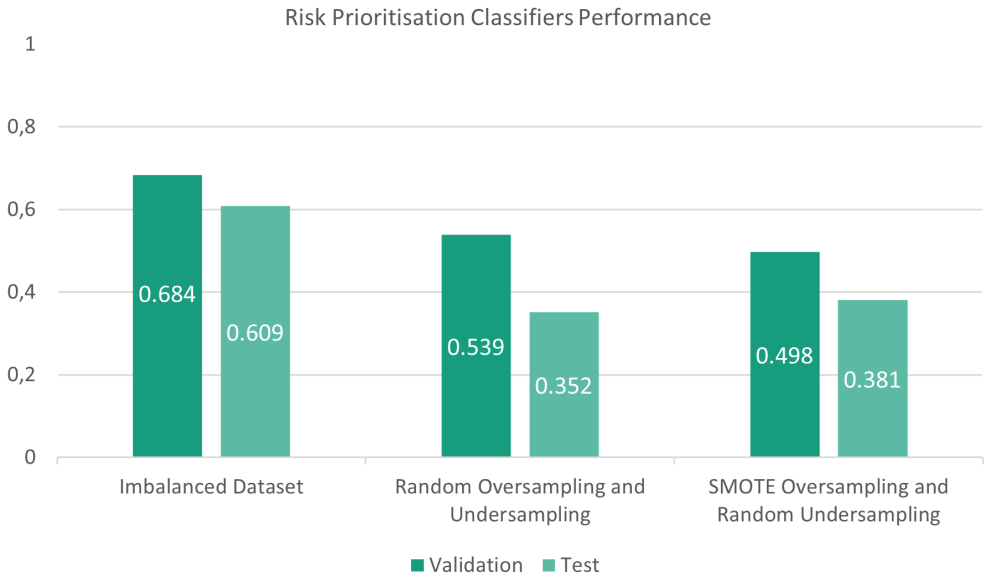


Figure 6.4: LSTM risk prioritisation models assessment using macro F1-score for test and validation sets.

The stratified shuffle split used in the dataset division into training, validation and test sets maintained the proportion of the classes in all sets. However, since the risk prioritisation task did not apply K-fold cross validation, only one validation set is being compared, contrary to what happened in the clinical entity extraction task,in which the presented validation F1-score resulted on a mean of the F1-scores obtained in each fold. If a K-fold cross-validation was employed and a mean of the validation macro F1-scores was calculated, the difference between validation and test sets could potentially decrease.

Starting by analysing the results for the unbalanced risk-prioritisation model, the classification report is presented in Table 6.1, with the precision, recall and F1-score values for each priority value, as well as macro and weighted average. The model performance is represented in a visual form in the confusion matrix seen in Figure 6.5.

It was expected that the lower priority class would present better results when compared to the other classes since it is by far the most frequent class and, for that reason, it is easier for the model to learn how to predict this class. This affirmation can be supported by Figure 6.5 since the instances in which the priority value 3 is predicted are the majority for all true classes. Consequently, this results in lower recall values for the other two priority values, since a significant amount of instances has a wrong priority value as predicted label, being mainly predicted as lower priority cases.

Table 6.1: Results for risk prioritisation task using imbalanced data distribution. For each evaluation metric, best results in bold and worst results underlined.

	Precision	**Recall**	**F1-score**
1	0.82	<u>0.23</u>	<u>0.36</u>
2	<u>0.62</u>	0.42	0.50
3	**0.95**	**0.98**	**0.97**
macro avg	0.80	0.54	0.61
weighted avg	0.93	0.93	0.93

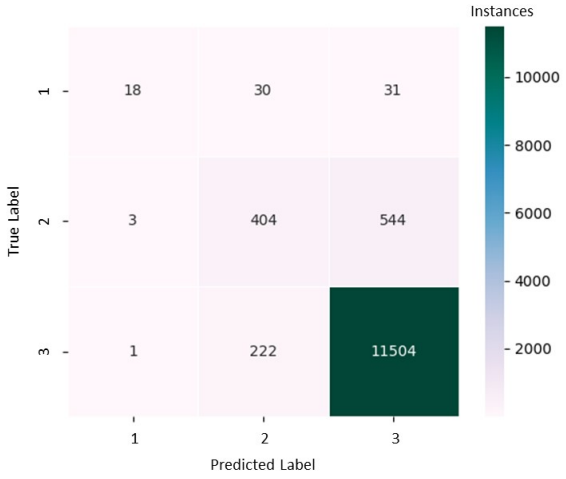


Figure 6.5: Confusion matrix for risk prioritisation classification using imbalanced data distribution.

On the contrary, for the priority values 1 and 2, the precision values are significantly higher when compared to the recall values. Given that precision represents the number of instances from a predicted label which the ground truth label is equal to the predicted label, observing the confusion matrix for the higher priority value, it is concluded that, from 22 instances predicted with the label 1, 18 instances are correctly labeled. The results for the priority label 2 are similar, although a greater fraction of cases with ground truth priority value 3 are predicted as having priority 2, which decreases the precision value for this class.

The macro average F1-score is 0.61. However, the high F1-score values for the three priority values are a consequence of the high precision values, that do not add significant value in this specific case. As previously explained and observed in the confusion matrix in Figure 6.5, the number of predictions for the high and intermediate priorities are considerably low, although there is an high fraction of cases whose ground truth label is equal to the predicted one, leading to high precision values. However, it is easily concluded that the model is not able to learn how to predict the higher priority classes, as a consequence of few instances from this priority in the training set. The high values of weighted average F1-score are justified by the fact that most of

the cases on the test set have priority value equal to 3, which obtained more favourable results. By doing a weighted average, the proportion of each priority value is taken into account, so the two lower priorities will have negligible weight in this result, since they have a smaller number of instances.

To balance the number of instances of each priority value and try to tackle the class imbalance issue, data-balancing strategies are also evaluated. In both strategies, random undersampling was utilised. The difference resides in the implemented overampling strategy.

Starting by analysing the random oversampling strategy, the results are presented in Table 6.2. These results are supported by the information found in the confusion matrix, presented in Figure 6.6. There is a clear difference between these results and the ones obtained for the class imbalanced dataset. Focusing on the priority classes 1 and 2, the precision values significantly decreased. This is a consequence of the balancing of classes in the training set, which results in the model predicting the priority values 1 and 2 more often in low priority instances. Comparing the number of predicted labels per priority class, it is observable that the predictions are more balanced than in the data-imbalanced model, in which the great majority of predictions were for the lower priority class. However, the predictions obtained for priority values 1 and 2 are not always correct, leading to an increase in false positive instances, that decrease the precision value as observed in the results.

Reversely, the recall values for priority values 1 and 2 increased substantially. Considering the ground truth labels for each one of these priority values, the number of instances that is accurately labeled in the prediction increased when random oversampling and undersampling methods were implemented. Consequently, the number of false negatives, represented by ground truth instances being predicted with another priority value, decrease, leading to a recall increase for the priority values 1 and 2. Contrarily, the recall for the priority value 3 decreased, because ground truth instances assigned with the lower risk priority are more likely to be predicted as having one of the higher risk priorities, as observed in the confusion matrix. The instances that are wrongly predicted as having a lower priority value increase the false negative instances for this class and, therefore, decrease the recall value.

Considering the macro average evaluation metrics, the F1-score decreased to 0.35. The precision also decreased from 0.80 to 0.39. On the other hand, the recall value increased from 0.54 to 0.70. This increase in recall can be interpreted as less instances being incorrectly labeled with a predicted priority value, which is an important aspect for the risk prioritisation prediction. It is important that high priority cases are not predicted as low priority cases, so that the wait time for an appointment is in accordance with the risk of the diagnosed condition.

Table 6.2: Results for risk prioritisation task using random oversampling and undersampling methods. For each evaluation metric, best results in bold and worst results underlined.

	Precision	**Recall**	**F1-score**
1	0.06	**0.81**	0.12
2	0.13	0.72	0.22
3	**0.99**	0.56	**0.71**
macro avg	0.39	0.70	0.35
weighted avg	0.92	0.57	0.67

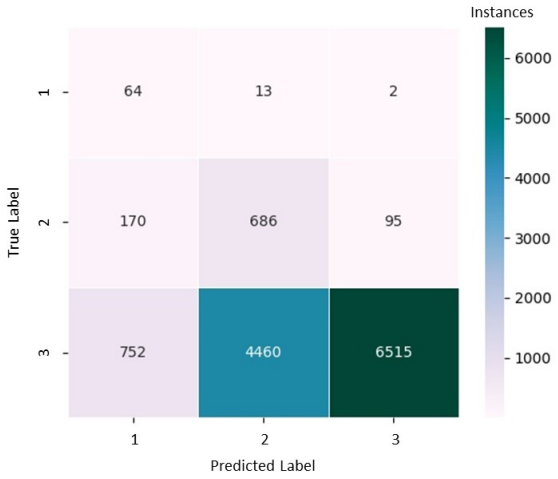


Figure 6.6: Confusion matrix for risk prioritisation classification using random oversampling and undersampling methods.

The second method chosen for oversampling of the test set was SMOTE. In conjunction with random undersampling, SMOTE was implemented and the results are presented in Table 6.3. The confusion matrix illustrated in Figure 6.7 represent the correct and incorrect predictions that result in the prediction and recall values for this model.

Similarly to the random oversampling results, the precision values for high and intermediate risk priority are lower than the ones obtained for the model trained with imbalanced data, because the number of predictions for labels 1 and 2 increase as a consequence of the increase instances used in the model training phase. The recall for the higher priority class is 0.38, also lower than the corresponding value for the random oversampling model, meaning that cases with a ground truth value of 1 were more frequently incorrectly labeled in the model prediction with a lower priority class, represented by 2 or 3. Besides, the overall prediction of instances with the higher priority value decreased for the SMOTE oversampling approach, when comparing with the random oversampling results. This may translate the inability of the model to be able to learn how to predict high priority cases.

The recall for the intermediate priority class increased from 0.72 for random oversampling to 0.85 for SMOTE oversampling. The number of correctly labeled cases with intermediate priority increased, as observed in the confusion matrix. Analysing the evaluation metrics for the lower priority class, the precision and recall values are 0.98 and 0.60, and are similar to the random oversampling outcome.

Table 6.3: Results for risk prioritisation task using SMOTE oversampling and random undersampling methods. For each evaluation metric, best results in bold and worst results underlined.

	Precision	**Recall**	**F1-score**
1	0.09	0.38	0.14
2	0.15	**0.85**	0.26
3	**0.98**	0.60	**0.74**
macro avg	0.41	0.61	0.38
weighted avg	0.91	0.61	0.70

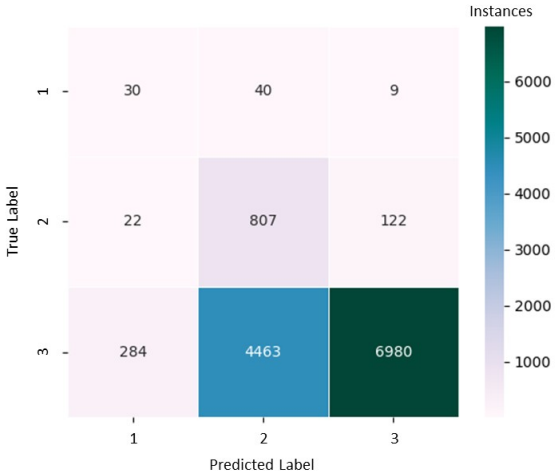


Figure 6.7: Confusion matrix for risk prioritisation classification using SMOTE oversampling and random undersampling methods.

The macro average values for precision, recall and F1-score are 0.41, 0.61 and 0.38, respectively. Although the F1-score value increased, it is still has margin for improvement. Besides, it is more appropriate to evaluate the performance considering the recall value, in order to understand the fraction of clinical cases that were labeled with the correct priority value, leading to an adequate waiting time for the specialty appointment according to the condition risk. The drop in the higher priority class recall affects the macro average value, that also decreases. This means that the capacity of the model to correctly predict a clinical case priority decreases.

6.1.3 Comparison of Text-based Risk Prioritisation Approaches

A comparison of the macro average evaluation metrics can be found in Table 6.4, aiming to compare the three implemented approaches using the LSTM architecture for risk prioritisation and evaluate how data-balancing techniques may improve the model outcome.

Table 6.4: Macro average scores for risk prioritisation task for the implemented models.

Model	Precision	Recall	F1-Score
Imbalanced Dataset	0.80	0.54	0.61
Random Oversampling and Random Undersampling	0.39	0.70	0.35
SMOTE Oversampling and Random Undersampling	0.41	0.61	0.38

In a medical context, such as the scope of this dissertation work, it is important that the amount of false negatives is minimal. In this context, this means that the model prediction must be able to identify a certain priority value when the ground truth label for the case is that same value. This is translated by the recall evaluation metric. In this context, a low recall on the higher priority cases means that patients with more severe conditions and an urgent need of a referral to Dermatology appointments would be given a lower priority, and would have to wait for a longer period of time to be followed up to the specialty appointment. Contrarily, if low priority cases were assigned with a higher priority, non-urgent cases would have access to a Dermatology appointment with a shorter waiting time. Therefore, a higher overall recall value is desired. It is specially important to maximise the correct predictions for the higher priority class, even if at the expense of some non-urgent cases being predicted as high priority cases and being referred to a specialty appointment more quickly.

Having this in mind, as well as the results presented in both Table 6.4 and Section 6.1.2, the approach that employed random oversampling and random undersampling achieved the best recall values, both for macro average recall and for the higher priority class. The recall results obtained for SMOTE oversampling and random undersampling are lower than for the previous mentioned approach, showing that, in this work, generating artificial data points in the feature space reveals a worse outcome than when textual data is replicated. Nevertheless, the results are better than when the imbalanced data was used in model training, that obtained a macro recall value of 0.54, and the recall for the higher priority class was 0.23. It can be concluded that data-balancing approaches are important to achieve better outcomes in the risk prioritisation task.

6.2 Derm.AI Risk Prioritisation Algorithm

The dissertation work is part of the Derm.AI, previously introduced in Chapter 1. In the present moment, an image analysis algorithm is implemented. To achieve a bimodal solution that also utilses information from the clinical notes, the information obtained from the clinical entity extraction task discussed in Chapter 5 must be used.

The utilised dataset in the Derm.AI algorithm is not the same as the one utilised in the scope of this work, being composed by 3,428 dermatology cases selected by a group of dermatologists cooperating on the project. The clinical cases are divided into 13 differential diagnoses, and only a single lesion is diagnosed in each case. Each differential diagnosis has an associated risk, and therefore, the prediction will assign one of the 13 classes to the clinical case. Table 6.5 presents the case distribution in relation to the differential diagnosis provided by dermatologists, as well as the image types observed for each diagnosis. Each case in the DermAI dataset has an associated clinical image, that can be macroscopic or anatomical. Although teledermatology guidelines recommend the acquisition of macroscopic images, in practice this is not always the case, as proved in the dataset, that contains 296 anatomical images. Given the small dataset size, as well as the difficulty to differentiate between macroscopic and anatomic images in anatomical sites such as hands, arms, feet, or the face, both image modalities were merged for the training of a deep neural network model.

Table 6.5: Derm.AI differential diagnosis dataset distribution

Class	**Differential Diagnosis**	**Macroscopic**	**Anatomical**	**Total**
1	Seborrheic Keratosis	1,125	61	1,186
2	Actinic Keratosis	442	77	519
3	Non-neoplastic Nevus	561	57	618
4	Molluscum contagiosum	50	21	71
5	Haemangioma	66	4	70
6	Neoplasm of Uncertain Behaviour	233	13	246
7	Dermatofibroma	134	6	140
8	Solar Lentigo	45	3	48
9	Pendulum Fibroma	98	16	114
10	Viral Warts	167	25	192
11	Other Malignant Neoplasm	108	8	116
12	Basal Cell Carcinoma	53	3	56
13	Malignant Melanoma	50	2	52
	Total	**3,133**	**296**	**3,428**

Three different networks were trained, using different data as input. The first implemented network is exclusively an image analysis algorithm, using the clinical images as input. The second model utilises the age and gender information found in the dataset as features, along with the clinical images used in the first model. The age is normalised for the values to be between 0 and 1. The gender is represented in encoded into a binary format.

The third implemented model aims to validate the use of the clinical information extracted from the unstructured clinical notes in the Derm.AI algorithm, and adds those features to the previously mentioned inputs. The clinical entity extraction from the clinical notes is performed using the fine-tuned BERTpt model, described in Section 5.5. Out of the eleven clinical entities

defined in the dissertation work, three entities are chosen as features for the model - Anatomical Site, Evolution, and Value. There is a selection and encoding of the extracted entities for these to be used as features. Regarding the Value entity, the numerical values relating to the diameter of the lesion are selected and converted to millimetre. The information extracted in the Anatomical Site entity is subject to an one-hot encoding method and divided into four main groups - face and neck, torso, upper limb and lower limb. One-hot encoding is also employed in the Evolution entity. The sectioning of the information considers cases without improvement, recent evolution (less than one year), prolonged evolution (more than one year), and occurrence of haemorrhage in the lesion. The process of encoding the information in the notes is done by identifying certain expressions in the clinical notes related to each of the previously mentioned sections. Therefore, each clinical case can have more than one evolution scenario.

Regarding the model training process, the data is split into training and test sets with a ratio of 80:20, considering a stratified distribution of the classes. To mitigate possible overfitting issues due to imbalanced data, stratified batches were considered, where the batch size was chosen to match the 13 classes from the Derm.AI algorithm. This results in oversampling of the classes with fewer examples, with the training data augmentation done using image data augmentation techniques such as rotation, flipping, width shift, zooming and brightness.

6.2.1 Results and Discussion

The results for the Derm.AI risk prioritisation are presented in this section, considering the precision, recall and F1-score evaluation metrics.

Table 6.6 displays the results for the model that only uses the images as input. Additional information regarding true labels and predicted labels can be found in the Appendix, in Figure B.0.1. It is observable that the results for diagnosis classes with less instances reveal lower results. The model is not able to predict classes 8 and 12, corresponding to solar lentigo and BCC cases, that have the value 0 as outcome for all analysed evaluation metrics. On the other hand, class 1, that corresponds to Seborrheic Keratosis diagnosis and has the larger instance number, achieves the better evaluation metrics for precision and F1-score.

The results for the model combining image analysis and normalised age and gender information are shown in Table 6.7. The corresponding confusion matrix can be found in Figure B.0.2 in the Appendix. Similarly to the previous approach, class 1 achieves the best evaluation metrics, with a F1-score of 0.63. The lowest scores belong to class 5 that, as observable in Table 6.5, corresponds to only 70 clinical cases from the Derm.AI selected dataset. The results for classes 8 and 12 slightly increased, obtaining F1-scores of 0.10 and 0.07, respectively. Overall, the results slightly increased, with the macro average precision value increasing to 0.29, and recall and F1-score maintaining the value of 0.29.

Table 6.6: Results for risk prioritisation using image analysis Derm.AI algorithm.

	Precision	**Recall**	**F1-Score**
1	**0.64**	0.50	**0.56**
2	0.49	**0.57**	0.52
3	0.51	0.43	0.47
4	0.38	0.36	0.37
5	0.18	0.14	0.16
6	0.11	0.16	0.13
7	0.39	0.43	0.41
8	0.00	0.00	0.00
9	0.22	0.22	0.22
10	0.46	0.56	0.51
11	0.18	0.26	0.21
12	0.00	0.00	0.00
13	0.13	0.20	0.16
macro avg	0.28	0.29	0.29

Table 6.7: Results for risk prioritisation using image analysis Derm.AI algorithm with age and gender data.

	Precision	**Recall**	**F1-Score**
1	**0.66**	**0.59**	**0.63**
2	0.57	0.56	0.56
3	0.51	0.52	0.52
4	0.33	0.21	0.26
5	0.00	0.00	0.00
6	0.15	0.22	0.18
7	0.44	0.43	0.44
8	0.09	0.10	0.10
9	0.31	0.17	0.22
10	0.38	**0.59**	0.46
11	0.16	0.13	0.14
12	0.06	0.09	0.07
13	0.13	0.10	0.11
macro avg	0.29	0.29	0.29

Lastly, the results for the risk prioritisation model that also includes the information extracted from the clinical notes are presented in Table 6.8. Additional information that support these results can be found in Figure B.0.3 in the Appendix. Class 5 achieves the lowest results, similarly to the previous model. Classes 1 and 2, with more representative instances in the dataset, have the highest F1-score of 0.59. Although there is not a visible increase in all classes when analysing the evaluation metrics, the macro average evaluation metrics slightly increased. The achieved values for precision, recall and F1-score are 0.32, 0.31, 0.31, respectively. These results show that, to some extent, the information extracted from the textual clinical data improves the Derm.AI risk prioritisation algorithm outcome. The usage of more information from the clinical notes, related to other clinical entities, can be a strategy to improve the model predictions.

Table 6.8: Results for risk prioritisation using image analysis Derm.AI algorithm with age, gender, and clinical entity data.

	Precision	**Recall**	**F1-Score**
1	**0.62**	0.56	**0.59**
2	0.57	**0.61**	**0.59**
3	0.47	0.56	0.51
4	0.42	0.36	0.38
5	0.07	0.07	0.07
6	0.14	0.06	0.08
7	0.43	0.43	0.43
8	0.08	0.10	0.08
9	0.30	0.13	0.18
10	0.58	0.54	0.56
11	0.15	0.35	0.21
12	0.07	0.09	0.08
13	0.20	0.20	0.20
macro avg	0.32	0.31	0.31

6.2.2 Comparison of Derm.AI Risk Prioritisation Approaches

A comparison of the macro and weighted average F1-scores for the analysed risk prioritisation models implemented in the scope of Derm.AI project can be found in Table 6.9.

Analysing the results, it can be stated that the addition of age and gender data does not have a significant effect on the model outcome, with equal recall and F1-scores of 0.29. In the last approach, using the extracted information from the clinical notes, there was a slight increase in the macro average evaluation metrics, achieving a recall and F1-score of 0.31. Although the increase in the evaluation metrics values is not accentuated, it can be stated that using information extracted from the unstructured clinical notes improve the model performance.

Table 6.9: Macro average scores for Derm.AI risk prioritisation models.

Model	Precision	Recall	F1-Score
Image Analysis	0.28	0.29	0.29
Image Analysis + Age and Gender	0.29	0.29	0.29
Image Analysis + Age and Gender + Clinical Entity Extraction	0.32	0.31	0.31

The implemented model using the clinical entities only use information extracted by three out of the eleven defined entities. A way to improve the results is to use encoded information from other clinical entities as features.

6.3 Summary

This chapter included two risk prioritisation models using the textual data extracted from the clinical notes as input, in order to prove that the extracted information is valuable for the risk prioritisation task.

The first implemented model was an LSTM neural network, with the tokenised clinical notes, the clinical entity tags and the ICD-9 code corresponding to each case as inputs. The chosen data for the risk prioritisation task is justified, based on the class imbalance found for the clinical case priority values, that was lower for the dataset composed by the non-annotated clinical notes. Since the priority distribution is still strongly imbalanced, two different approaches using oversampling and undersampling strategies were adopted to surpass the imbalanced distribution and evaluate how it affects the results. Recall was utilised for the outcome assessment, since it represents the fraction of clinical cases whose priority value is accurately predicted by the model. In the risk prioritisation context, it is specially important that a high recall value is obtained for the higher priority value, since it represents the maximisation of the correct predictions for that class, and consequently, a shorter waiting time for the Dermatology specialty appointment. Since the macro average recall was higher for the random and SMOTE oversampling models than for the data-imbalanced model, it can be affirmed that the data-balancing strategies improved the prediction results. Overall, random oversampling together with random undersampling resulted in the best macro average recall of 0.70, as well as the best recall value for the high priority class.

The Derm.AI risk prioritisation implementation is also described, and the results obtained by the model that uses information extracted from the clinical entities Anatomical Site, Evolution and Value as input are compared with the models that only perform image analysis and use age and gender data. There is a slight increase when using encoded data extracted from the selected clinical entities, achieving a macro recall and F1-score of 0.31. Analysing the results, it can be concluded that the textual data positively influence the risk prioritisation task. The usage of more clinical entities would add more information regarding the Dermatology clinical case and, therefore, could improve the model prediction.

Chapter 7

Conclusions and Further Work

Every year, millions of people worldwide are diagnosed with skin cancer, and that number tends to increase over the years for various reasons. There is an economic burden in public health services caused by increased skin cancer incidence. Given the existing limitations in the diagnosis process, namely the lack of formation and resources for the primary care physicians to perform a skin cancer diagnosis and the reduced number of speciality dermatology physicians in the Portuguese NHS, the development of AI applications to support medical decisions has increased. The Derm.AI project intends to support the patient referral from primary care to Dermatology appointments, by performing a prioritisation according to the clinical case associated risk, using both image analysis and NLP.

The main goal of the dissertation was to develop a pipeline for risk prioritisation using Dermatology clinical notes from the Portuguese NHS. For this goal to be accomplished, the work was divided into two main tasks: the extraction of clinical information of interest from the clinical notes, and the risk prioritisation algorithm itself.

Eleven clinical entities were selected to represent valuable information in the scope of the Dermatology domain. The manual annotation of 980 clinical cases was performed using the selected clinical entities, and part of the annotated dataset was further validated by two dermatologists, achieving a total agreement ratio of $87.21 \pm 0.10\%$. The vectorised representation of each token is fundamental in both main tasks. The FastText algorithm was implemented, using clinical text from Dermatology and Neurology as input for word embedding training.

Considering the clinical entity extraction task, four different ML and DL approaches were implemented, and evaluated using the manual annotated corpus. BERTpt obtained the best F1-score of 0.76. This outcome is in accordance with the literature, since Transformer architectures are the state-of-the-art for several NLP tasks, such as entity extraction. When compared with the other fine-tuned BERT models, BERTpt has a better outcome because it is pre-trained with general domain Portuguese text, achieving better results than the multilingual pre-trained models.

Taking the best results obtained in the clinical entity extraction task into account, the BERTpt model is applied to the remaining non-annotated clinical notes in the Dermatology dataset. Along with the tokenised clinical notes and the ICD codes, the resulting clinical entities are used as input for the risk prioritisation model, that uses the LSTM neural network architecture to

predict the prioritisation of a clinical case. The analysed evaluation metric was recall, because it represents the model ability to correctly predict the case priority value, which is important because errors in the prediction can lead to an appointment wait time that is not compatible with the condition severity. Data-balancing methods were utilised to counteract the case priority unbalance found in the dataset. The models in which data-balancing strategies were performed achieved better recall results than the model in which the imbalanced dataset was used. Random oversampling was the best approach, achieving a recall of 0.70. A selection of extracted information from the clinical notes was also used as input to the already implemented Derm.AI risk prioritisation model, that utilised an image analysis approach. The model achieved the best outcome when using the information extracted from the textual data as feature, proving that the extraction of information from the clinical notes positively impacts the risk prioritisation of dermatological clinical cases.

In conclusion, it can be asserted that the goals for this dissertation work were achieved. The developed approaches are a first step towards including information from Dermatology clinical notes in the Derm.AI project algorithm and, for that reason, can contribute to future scientific publications.

7.1 Future Work

Regarding further work to be developed in the Derm.AI project context, a better adjustment of the approaches to the Dermatology and clinical domain can positively impact the results. Firstly, there would be interest in training the FastText word embedding model using only Dermatology clinical cases in a substantial amount, so that only Dermatology-specific vocabulary used in the WE training.

Regarding the clinical entity extraction, although the BioBERTpt model, pre-trained with Portuguese clinical notes, is available in the HuggingFace repository, it was not possible to fine-tune it for the Named Entity Recognition task. Since BERTpt was the model that achieved better results, it is expected that BioBERTpt could achieve a better performance since it is pre-trained with text from clinical domain. In terms of the risk prioritisation task, it would be interesting to evaluate the predictions of other neural network architectures for this specific problem, as well as analysing other data-balancing methods.

There are other tasks in the NLP field that would be relevant to explore using the Dermatology dataset. For example, an automatic prediction of the ICD code using extracted information about the observed condition would be a valuable feature for the Derm.AI algorithm.

Bibliography

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[2] C. P. Wild, E. Weiderpass, and B. W. Stewart, *Cancer research for cancer prevention World Cancer Report*. 2020, ISBN: 9789283204473.

[3] Z. Khazaei, F. Ghorat, A. Jarrahi, H. Adineh, M. Sohrabivafa, and E. Goodarzi, "Global incidence and mortality of skin cancer by histological subtype and its relationship with the human development index (hdi); an ecology study in 2018," *World Cancer Res J*, vol. 6, no. 2, e13, 2019.

[4] J. E. Gershenwald and G. P. Guy, "Stemming the rising incidence of melanoma: Calling prevention to action," *JNCI: Journal of the National Cancer Institute*, vol. 108, no. 1, 2016.

[5] T. L. Diepgen and V. Mahler, "The epidemiology of skin cancer," *British Journal of Dermatology*, vol. 146, pp. 1–6, 2002.

[6] Z. Apalla, A. Lallas, E. Sotiriou, E. Lazaridou, and D. Ioannides, "Epidemiological trends in skin cancer," *Dermatology practical & conceptual*, vol. 7, no. 2, p. 1, 2017.

[7] A. F. Duarte, A. da Costa-Pereira, V. Del-Marmol, and O. Correia, "Are general physicians prepared for struggling skin cancer?—cross-sectional study," *Journal of Cancer Education*, vol. 33, no. 2, pp. 321–324, 2018.

[8] T. Mudigonda, D. J. Pearce, B. A. Yentzer, P. Williford, and S. R. Feldman, "The economic impact of non-melanoma skin cancer: A review," *Journal of the National Comprehensive Cancer Network*, vol. 8, no. 8, pp. 888–896, 2010.

[9] S. Carr, C. Smith, and J. Wernberg, "Epidemiology and risk factors of melanoma," *Surgical Clinics*, vol. 100, no. 1, pp. 1–12, 2020.

[10] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[11] Instituto Nacional de Estatística, "Sociedade de informação Inquérito à Utilização das Tecnologias de Informação e da Comunicação nos Hospitais Proporção de hospitais com processos clínicos eletrónicos quase duplicou numa década," pp. 1–7, 2014.

[12] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: Past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[13] M. E. Matheny, D. Whicher, and S. T. Israni, "Artificial intelligence in health care: A report from the national academy of medicine," *Jama*, vol. 323, no. 6, pp. 509–510, 2020.

[14] H. Tsao, J. M. Olazagasti, K. M. Cordoro, J. D. Brewer, S. C. Taylor, J. S. Bordeaux, M.-M. Chren, A. J. Sober, C. Tegeler, R. Bhushan, *et al.*, "Early detection of melanoma: Reviewing the abcdes," *Journal of the American Academy of Dermatology*, vol. 72, no. 4, pp. 717–723, 2015.

[15] W. R. Shaikh, S. W. Dusza, M. A. Weinstock, S. A. Oliveria, A. C. Geller, and A. C. Halpern, "Melanoma thickness and survival trends in the united states, 1989–2009," *JNCI: Journal of the National Cancer Institute*, vol. 108, no. 1, 2016.

[16] V. Madan, J. T. Lear, and R.-M. Szeimies, "Non-melanoma skin cancer," *The lancet*, vol. 375, no. 9715, pp. 673–685, 2010.

[17] Z. Apalla, D. Nashan, R. B. Weller, and X. Castellsagué, "Skin cancer: Epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches," *Dermatology and therapy*, vol. 7, no. 1, pp. 5–19, 2017.

[18] P. A. Bath, "Health informatics: Current issues and challenges," *Journal of information science*, vol. 34, no. 4, pp. 501–518, 2008.

[19] J. M. Gesulga, A. Berjame, K. S. Moquiala, and A. Galido, "Barriers to electronic health record system implementation and information systems resources: A structured review," *Procedia Computer Science*, vol. 124, pp. 544–551, 2017.

[20] K. Häyrinen, K. Saranto, and P. Nykänen, "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," *International journal of medical informatics*, vol. 77, no. 5, pp. 291–304, 2008.

[21] K. Cresswell, A. Worth, and A. Sheikh, "Implementing and adopting electronic health record systems," *Clinical Governance: An International Journal*, 2011.

[22] G. H. Shah, J. P. Leider, B. C. Castrucci, K. S. Williams, and H. Luo, "Characteristics of local health departments associated with implementation of electronic health records and other informatics systems," *Public Health Reports*, vol. 131, no. 2, pp. 272–282, 2016.

[23] E. W. Ford, N. Menachemi, and M. T. Phillips, "Predicting the adoption of electronic health records by physicians: When will health care be paperless?" *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 106–112, 2006.

[24] R. Hillestad, J. Bigelow, A. Bower, F. Girosi, R. Meili, R. Scoville, and R. Taylor, "Can electronic medical record systems transform health care? potential health benefits, savings, and costs," *Health affairs*, vol. 24, no. 5, pp. 1103–1117, 2005.

[25] S. V. Jardim, "The electronic health record and its contribution to healthcare information systems interoperability," *Procedia technology*, vol. 9, pp. 940–948, 2013.

[26] E. D. Liddy, "Natural language processing," 2001.

[27] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.

[28] F. Y. Choi, P. Wiemer-Hastings, and J. D. Moore, "Latent semantic analysis for text segmentation," in *Proceedings of the 2001 conference on empirical methods in natural language processing*, 2001.

[29] D. D. Palmer, "Tokenisation and sentence segmentation," *Handbook of natural language processing*, pp. 11–35, 2000.

[30] K. Cohen, "Biomedical natural language processing and text mining," *Methods in biomedical informatics: a pragmatic approach*, vol. 141, 2013.

[31] J. J. Webster and C. Kit, "Tokenization as the initial phase in nlp," in *COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics*, 1992.

[32] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information fusion*, vol. 36, pp. 10–25, 2017.

[33] D. Sarkar, "Text analytics with python," 2016.

[34] A. G. Jivani *et al.*, "A comparative study of stemming algorithms," *Int. J. Comp. Tech. Appl*, vol. 2, no. 6, pp. 1930–1938, 2011.

[35] M. Chary, S. Parikh, A. F. Manini, E. W. Boyer, and M. Radeos, "A review of natural language processing in medical education," *Western Journal of Emergency Medicine*, vol. 20, no. 1, p. 78, 2019.

[36] V. Balakrishnan and E. Lloyd-Yemoh, "Stemming and lemmatization: A comparison of retrieval performances," 2014.

[37] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2017.

[38] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 41, 2016.

[39] X. Lv, Y. Guan, J. Yang, and J. Wu, "Clinical relation extraction with deep learning," *International Journal of Hybrid Information Technology*, vol. 9, no. 7, pp. 237–248, 2016.

[40] T. Tran, T. D. Nguyen, D. Phung, and S. Venkatesh, "Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm)," *Journal of biomedical informatics*, vol. 54, pp. 96–105, 2015.

[41] O. Jacobson and H. Dalianis, "Applying deep learning on electronic health records in swedish to predict healthcare-associated infections," in *Proceedings of the 15th workshop on biomedical natural language processing*, 2016, pp. 191–195.

[42] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, 2016, pp. 301–318.

[43] A. N. Jagannatha and H. Yu, "Bidirectional rnn for medical event detection in electronic health records," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, NIH Public Access, vol. 2016, 2016, p. 473.

[44] A. N. Jagannatha and H. Yu, "Structured prediction models for rnn based sequence labeling in clinical text," in *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, NIH Public Access, vol. 2016, 2016, p. 856.

[45] J. A. Fries, "Brundlefly at semeval-2016 task 12: Recurrent neural networks vs. joint inference for clinical temporal information extraction," *arXiv preprint arXiv:1606.01433*, 2016.

[46] Y. Liu, T. Ge, K. S. Mathews, H. Ji, and D. L. McGuinness, "Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion," *arXiv preprint arXiv:1804.04225*, 2018.

[47] B. Mohit, "Named entity recognition," in *Natural language processing of semitic languages*, Springer, 2014, pp. 221–245.

[48] H. Shelar, G. Kaur, N. Heda, and P. Agrawal, "Named entity recognition approaches and their comparison for custom ner model," *Science & Technology Libraries*, vol. 39, no. 3, pp. 324–337, 2020.

[49] Y. Wu, J. Xu, M. Jiang, Y. Zhang, and H. Xu, "A study of neural word embeddings for named entity recognition in clinical text," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2015, 2015, p. 1326.

[50] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural language processing using very large corpora*, Springer, 1999, pp. 157–176.

[51] D. Demner-Fushman and N. Elhadad, "Aspiring to unintended consequences of natural language processing: A review of recent developments in clinical and consumer-generated text processing," *Yearbook of medical informatics*, no. 1, p. 224, 2016.

[52] R. H. Baud, A.-M. Rassinoux, C. Lovis, J. Wagner, V. Griesser, P.-A. Michel, and J.-R. Scherrer, "Knowledge sources for natural language processing.," in *Proceedings of the AMIA Annual Fall Symposium*, American Medical Informatics Association, 1996, p. 70.

[53] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, and G. O. Barnett, "The unified medical language system: An informatics research collaboration," *Journal of the American Medical Informatics Association*, vol. 5, no. 1, pp. 1–11, 1998.

[54] B. L. Humphreys, G. Del Fiol, and H. Xu, "The umls knowledge sources at 30: Indispensable to current research and applications in biomedical informatics," *Journal of the American Medical Informatics Association*, vol. 27, no. 10, pp. 1499–1501, 2020.

[55] M. Becker, S. Kasper, B. Böckmann, K.-H. Jöckel, and I. Virchow, "Natural language processing of german clinical colorectal cancer notes for guideline-based treatment evaluation," *International journal of medical informatics*, vol. 127, pp. 141–146, 2019.

[56] F. B. Putra, A. A. Yusuf, H. Yulianus, Y. P. Pratama, D. S. Humairra, U. Erifani, D. K. Basuki, S. Sukaridhoto, and R. P. N. Budiarti, "Identification of symptoms based on natural language processing (nlp) for disease diagnosis based on international classification of diseases and related health problems (icd-11)," in *2019 International Electronics Symposium (IES)*, IEEE, 2019, pp. 1–5.

[57] Y. Baştanlar and M. Özuysal, "Introduction to machine learning," in *miRNomics: MicroRNA Biology and Computational Analysis*, Springer, 2014, pp. 105–128.

[58] M. W. Berry, A. Mohamed, and B. W. Yap, *Supervised and unsupervised learning for data science*. Springer, 2019.

[59] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, Ieee, 2016, pp. 1310–1315.

[60] R. Gentleman and V. J. Carey, "Unsupervised machine learning," in *Bioconductor case studies*, Springer, 2008, pp. 137–157.

[61] A. Vlachos, "Evaluating unsupervised learning for natural language processing tasks," in *Proceedings of the First workshop on Unsupervised Learning in NLP*, 2011, pp. 35–42.

[62] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[63] A. Bibal, M. Lognoul, A. De Streel, and B. Frénay, "Legal requirements on explainability in machine learning," *Artificial Intelligence and Law*, vol. 29, no. 2, pp. 149–169, 2021.

[64] J. Haneczok and J. Piskorski, "Shallow and deep learning for event relatedness classification," *Information Processing & Management*, vol. 57, no. 6, p. 102 371, 2020.

[65] A. H. Marblestone, G. Wayne, and K. P. Kording, "Toward an integration of deep learning and neuroscience," *Frontiers in computational neuroscience*, vol. 10, p. 94, 2016.

[66] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and trends in signal processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[67] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, `http://www.deeplearningbook.org`.

[68] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[69] S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, *et al.*, "Deep learning in clinical natural language processing: A methodical review," *Journal of the American Medical Informatics Association*, vol. 27, no. 3, pp. 457–470, 2020.

[70] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[71] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[72] L. Medsker and L. C. Jain, *Recurrent neural networks: design and applications*. CRC press, 1999.

[73] S. Hochreiter, "Recurrent neural net learning and vanishing gradient," *International Journal Of Uncertainity, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 2, pp. 107–116, 1998.

[74] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[75] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[76] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[77] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[78] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[79] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *Journal of Biomedical Informatics: X*, vol. 4, p. 100 057, 2019.

[80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[81] C. Wang, M. Li, and A. J. Smola, "Language models with transformers," *arXiv preprint arXiv:1904.09408*, 2019.

[82] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[83] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[84] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[85] D. Berrar, "Cross-validation," *Encyclopedia of bioinformatics and computational biology*, vol. 1, pp. 542–545, 2019.

[86] L. Ferreira, A. Teixeira, and J. P. S. Cunha, "Medical information extraction in European Portuguese," *Handbook of Research on ICTs for Human-Centered Healthcare and Social Care Services*, pp. 607–626, 2013.

[87] F. Lopes, C. Teixeira, and H. G. Oliveira, "Contributions to clinical named entity recognition in portuguese," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 223–233.

[88] ——, "Comparing different methods for named entity recognition in portuguese neurology text," *Journal of Medical Systems*, vol. 44, no. 4, pp. 1–20, 2020.

[89] A. C. Peters, A. M. P. da Silva, C. P. Gebeluca, Y. B. Gumiel, L. M. M. Cintho, D. R. Carvalho, S. A. Hasan, C. M. C. Moro, *et al.*, "Semclinbr–a multi institutional and multi specialty semantically annotated corpus for portuguese clinical nlp tasks," *arXiv preprint arXiv:2001.10071*, 2020.

[90] L. Chen, Y. Gu, X. Ji, C. Lou, Z. Sun, H. Li, Y. Gao, and Y. Huang, "Clinical trial cohort selection based on multi-level rule-based natural language processing system," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1218–1226, 2019.

[91] D. J. Feller, J. Zucker, M. T. Yin, P. Gordon, and N. Elhadad, "Using clinical notes and natural language processing for automated hiv risk assessment," *Journal of acquired immune deficiency syndromes (1999)*, vol. 77, no. 2, p. 160, 2018.

[92] Z. Yun-tao, G. Ling, and W. Yong-cheng, "An improved tf-idf approach for text classification," *Journal of Zhejiang University-Science A*, vol. 6, no. 1, pp. 49–55, 2005.

[93] H. S. Chase, L. R. Mitrani, G. G. Lu, and D. J. Fulgieri, "Early recognition of multiple sclerosis using natural language processing of the electronic health record," *BMC medical informatics and decision making*, vol. 17, no. 1, pp. 1–8, 2017.

[94] A. Al-Aiad, R. Duwairi, and M. Fraihat, "Survey: Deep learning concepts and techniques for electronic health record," in *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 2018, pp. 1–5.

[95] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *Journal of biomedical informatics*, vol. 69, pp. 218–229, 2017.

[96] K. Xu, Z. Zhou, T. Hao, and W. Liu, "A bidirectional lstm and conditional random fields approach to medical named entity recognition," in *International Conference on Advanced Intelligent Systems and Informatics*, Springer, 2017, pp. 355–365.

[97] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[98] J. Costa, I. Lopes, A. Carreiro, D. Ribeiro, and C. Soares, "Fraunhofer aicos at clef ehealth 2020 task 1: Clinical code extraction from textual data using fine-tuned bert models," in *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2020.

[99] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than english: Opportunities and challenges," *Journal of biomedical semantics*, vol. 9, no. 1, p. 12, 2018.

[100] N. Nishimoto, S. Terae, M. Uesugi, K. Ogasawara, and T. Sakurai, "Development of a medical-text parsing algorithm based on character adjacent probability distribution for japanese radiology reports," *Methods of information in medicine*, vol. 47, no. 06, pp. 513–521, 2008.

[101] R. Costumero, Á. Garcı´a-Pedrero, C. Gonzalo-Martı´n, E. Menasalvas, and S. Millan, "Text analysis and information extraction from spanish written documents," in *International Conference on Brain Informatics and Health*, Springer, 2014, pp. 188–197.

[102] H.-U. Krieger, C. Spurk, H. Uszkoreit, F. Xu, Y. Zhang, F. Müller, and T. Tolxdorff, "Information extraction from german patient records via hybrid parsing and relation extraction strategies.," in *LREC*, 2014, pp. 2043–2048.

[103] H. Moen, F. Ginter, E. Marsi, L.-M. Peltonen, T. Salakoski, and S. Salanterä, "Care episode retrieval: Distributional semantic models for information retrieval in the clinical domain," in *BMC medical informatics and decision making*, BioMed Central, vol. 15, 2015, pp. 1–19.

[104] O. Perez-de-Viñaspre and M. Oronoz, "Snomed ct in a language isolate: An algorithm for a semiautomatic translation," in *BMC medical informatics and decision making*, BioMed Central, vol. 15, 2015, pp. 1–14.

[105] A. de Souza, P. Nohama, C. Moro, *et al.*, "A rule-based method for continuity of care identification in discharge summaries.," *Studies in health technology and informatics*, vol. 192, pp. 1221–1221, 2013.

[106] M. Fernandes, R. Mendes, S. M. Vieira, F. Leite, C. Palos, A. Johnson, S. Finkelstein, S. Horng, and L. A. Celi, "Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing," *PloS one*, vol. 15, no. 3, 2020.

[107] M. Fernandes, R. Mendes, S. M. Vieira, F. Leite, C. Palos, A. Johnson, S. Finkelstein, S. Horng, and L. A. Celi, "Risk of mortality and cardiopulmonary arrest in critical patients presenting to the emergency department using machine learning and natural language processing," *PloS one*, vol. 15, no. 4, 2020.

[108] E. T. R. Schneider, J. V. A. de Souza, J. Knafou, L. E. S. e Oliveira, J. Copara, Y. B. Gumiel, L. F. A. de Oliveira, E. C. Paraiso, D. Teodoro, and C. M. C. M. Barra, "Biobertpt-a portuguese neural language model for clinical named entity recognition," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 2020, pp. 65–72.

[109] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

[110] M. Skeppstedt, M. Kvist, G. H. Nilsson, and H. Dalianis, "Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study," *Journal of biomedical informatics*, vol. 49, pp. 148–158, 2014.

[111] P. V. Ogren, G. K. Savova, C. G. Chute, *et al.*, "Constructing evaluation corpora for automated clinical named entity recognition," in *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, IOS Press, 2007, p. 2325.

[112] Y. Wang and J. Patrick, "Cascading classifiers for named entity recognition in clinical notes," in *Proceedings of the workshop on biomedical information extraction*, 2009, pp. 42–49.

[113] L. Akhtyamova, P. Martı´nez, K. Verspoor, and J. Cardiff, "Testing contextualized word embeddings to improve ner in spanish clinical case narratives," *IEEE Access*, vol. 8, pp. 164 717–164 726, 2020.

[114] Y. Zhang, H.-J. Li, J. Wang, T. Cohen, K. Roberts, and H. Xu, "Adapting word embeddings from multiple domains to symptom recognition from psychiatric notes," *AMIA Summits on Translational Science Proceedings*, vol. 2018, p. 281, 2018.

[115] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[116] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989.

[117] S. Okada, M. Ohzeki, and S. Taguchi, "Efficient partition of integer optimization problems with one-hot encoding," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[118] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[119] I. Girardi, P. Ji, A.-p. Nguyen, N. Hollenstein, A. Ivankay, L. Kuhn, C. Marchiori, and C. Zhang, "Patient risk assessment and warning symptom detection using deep attention-based neural networks," *arXiv preprint arXiv:1809.10804*, 2018.

[120] C. Li, G. Zhan, and Z. Li, "News text classification based on improved bi-lstm-cnn," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, IEEE, 2018, pp. 890–893.

[121] G. Rao, W. Huang, Z. Feng, and Q. Cong, "Lstm with sentence representations for document-level sentiment classification," *Neurocomputing*, vol. 308, pp. 49–57, 2018.

[122] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[123] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[124] M. Schubach, M. Re, P. N. Robinson, and G. Valentini, "Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[125] I. Mani and I. Zhang, "Knn approach to unbalanced data distributions: A case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, ICML United States, vol. 126, 2003.

[126] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17–26, 2017.

Appendix A

Manual Annotation Tool

Figure A.0.1 shows the Prodigy[14] interface for the manual annotation tool. The top part of the figure contains the clinical entities defined in the scope of the dissertation work for the clinical entity extraction task, previously described in Section 4.2.1. The annotation tool displays one sentence at the time, which is manually annotated by choosing the corresponding entity and selecting the tokens that should be included in that entity. The green box must be selected when the user wants to accept the sentence annotation and move to the next sentence. The red box and grey box with a forbidden symbol represent the Reject and Ignore command, and should be used when the user does not want to include a certain sentence in the annotated dataset, or because the user does not know how to annotate that sentence. The last grey box with an arrow can be used to return to previous annotated sentences in case a revision should be made.

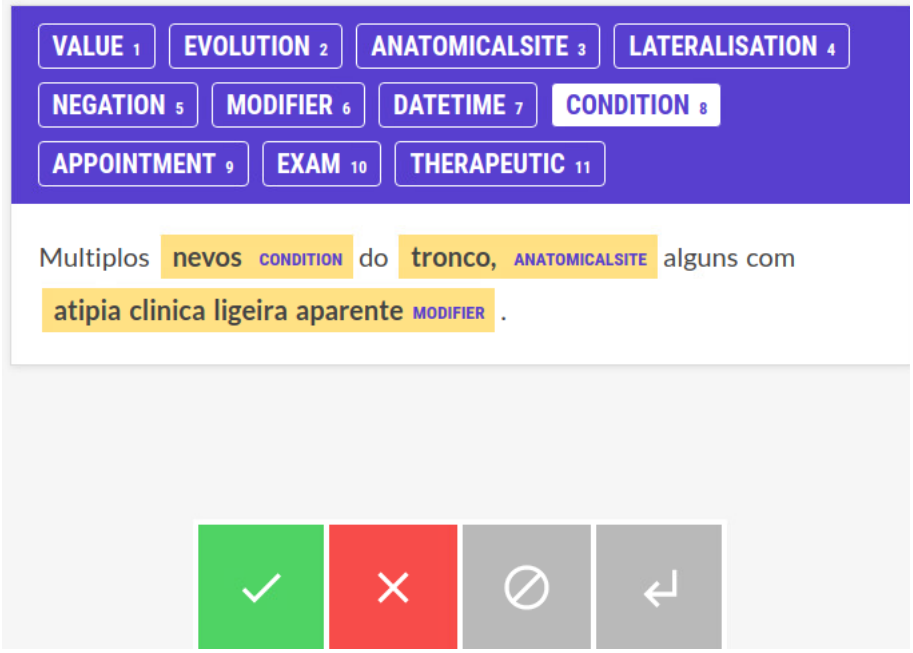


Figure A.0.1: Prodigy interface for manual annotation task.

Appendix B

Derm.AI Risk Prioritisation Algorithms - Confusion Matrices

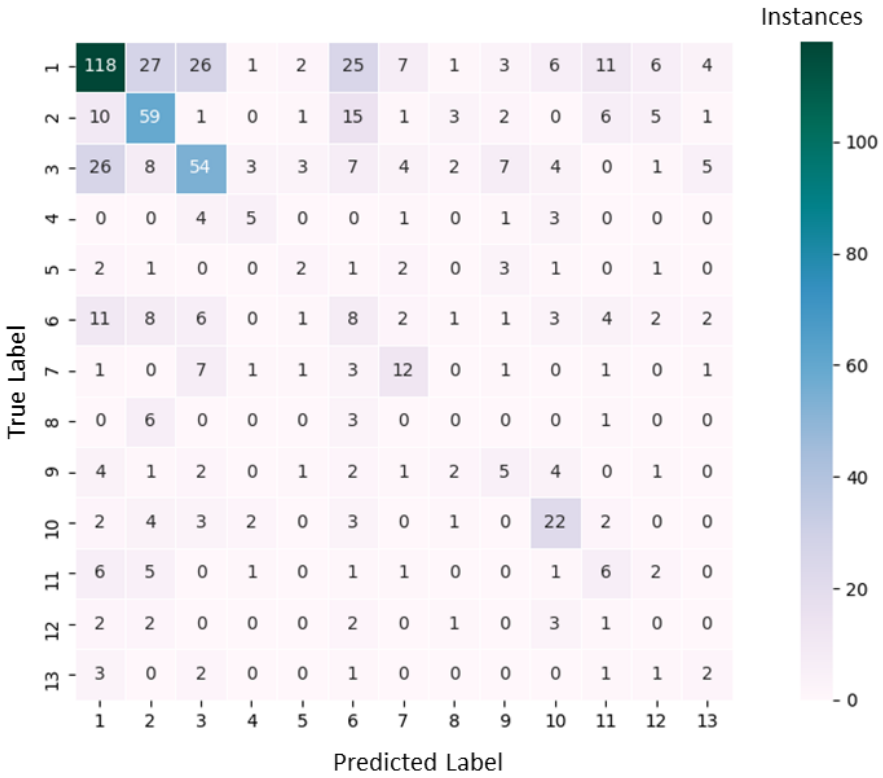


Figure B.0.1: Confusion matrix for risk prioritisation using image analysis Derm.AI algorithm.

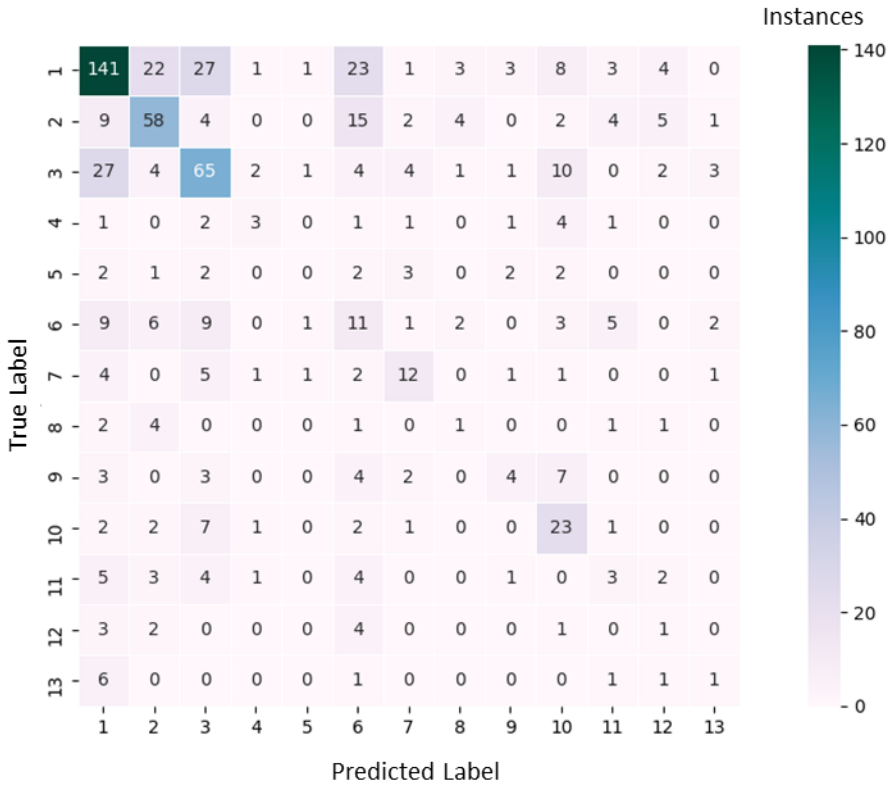


Figure B.0.2: Confusion matrix Results for risk prioritisation using image analysis Derm.AI algorithm with age and gender data.

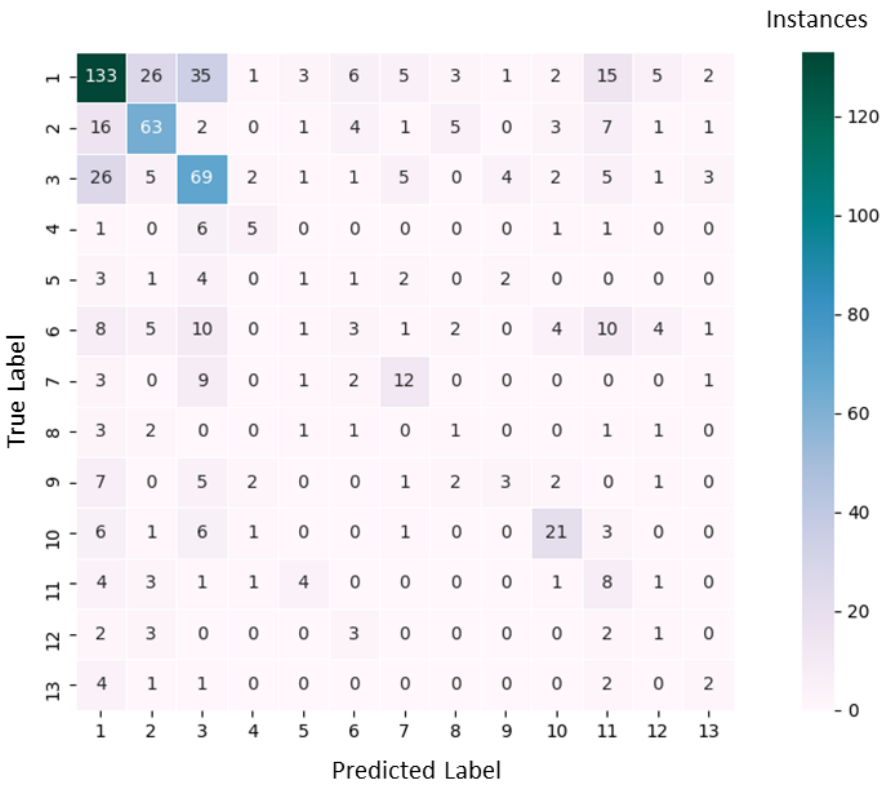


Figure B.0.3: Confusion matrix for risk prioritisation using image analysis Derm.AI algorithm with age, gender, and clinical entity data.