UNIVERSITY OF TWENTE.

Identifying Influential Variables for an Explainable AI based Clinical Decision Support System in the Healthcare Industry

by

Douwe Rotink

A thesis submitted to the
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
in partial fulfilment of the requirements for the degree of

MSc in Business Information Technology

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)

University of Twente

Enschede, Overijssel, The Netherlands

June 2024

ABSTRACT

The medical field has greatly benefited from the collection and analysis of large datasets, gaining critical insights into patient health, disease patterns, and treatment efficacy. Recognizing these benefits, healthcare providers have structured their data collection efforts to maximize value, primarily storing data in Electronic Health Records (EHR). This structured data, being organized and easily searchable, opens new avenues for analysis, providing a comprehensive understanding of healthcare outcomes.

Company X has developed standardized reports to capture this structured data, aiming to enhance business processes. This research focuses on the anamnesis process, which marks the initial interaction between the practitioner and the patient, centring on the patient's medical history and symptoms. The objective is to develop a machine learning (ML) based critical decision support system (CDSS) tailored for the anamnesis conversation. The ML model functions as a classification tool to assist in diagnosing patients, with practitioners retaining final decision authority. Hence, the model incorporates explainable AI (XAI) to ensure transparency, enabling practitioners to discern correct from incorrect model outputs. Achieving this goal promises standardized patient care for greater consistency and reliability, efficient patient evaluation to reduce consultation times, and reduced misdiagnoses, leading to better treatment plans and minimized adverse health outcomes. The central research question addressed is:

> *How to design and integrate a Explainable Artificial Intelligence (XAI) based Clinical Decision Support System (CDSS) to identify the most influential variables and support practitioners in diagnosing patients?*

The research begins with an analysis of the anamnesis conversation, detailing the completion of three structured reports: anamnesis, examination, and treatment reports. The anamnesis report contains variables used by the ML model to classify diagnoses, primarily nominal data on pain location and type. The treatment report provides the final diagnosis, serving as the research's label, encompassing 177 diagnoses. This study focuses on a subset of these diagnoses, narrowing it down to 20.

After understanding the business process and data, the data was prepared for ML use, starting with a star schema model to gain insights into input values. After the modelling, missing data was addressed using averages or medians. Feature reduction was also implemented to minimise variables from 299 to 51. This data was then used to train three models: two single-label classifiers (SLC) using automated machine learning, and a multi-label classifier (MLC) random forest model. The MLC model outperformed the others based on sensitivity (0.45), precision (0.64), and F1-score (0.48), with high precision being crucial for identifying correct

diagnoses. The research also emphasizes the explainable output of the ML model. Four XAI outputs were created and evaluated by practitioners, leading to a final output format that highlights: variables leading to the diagnosis, variables deemed irrelevant, and alternative options for non-correlated variables.

In conclusion, this study offers a thorough literature review of XAI in healthcare and presents an XAI-verified diagnosis classification model. Additionally, it underscores the data modeling and preparation of EHR data for ML use, contributing valuable insights into the integration of ML in clinical settings.

Keywords: CDSS; Clinical Decision Support System, XAI, Explainable AI, Key Variables, EHR, Electronic Health Records, Single-Label Classification, SLC, Multi-Label Classification, MLC

AUTHOR'S DECLARATION

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Twente to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Twente to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

Douwe Rotink

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to everyone who supported and guided me throughout this project. Your contributions have been invaluable, and this thesis would not have been possible without your assistance.

First and foremost, I would like to thank my thesis supervisors, Marcos Machado, João Rebelo Moreira, and Wallace Corbo Ugulino. Your expertise and guidance were critical in shaping this research. In particular, I am profoundly grateful to Marcos Machado for his support and consistent feedback at every stage of my research. Your insights were essential to the completion of this work.

I am also deeply appreciative of my colleagues at Company X and the Healthcare Group Y. Conducting my research in such an enjoyable working environment was a pleasure. Special thanks go to my supervisors at the company. Their guidance during our weekly meetings and their support were instrumental in providing me with the clarity and direction needed to advance my research.

To all of you, thank you once again for your invaluable contributions to this thesis. I hope you enjoy reading it.

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

AB AdaBoosting.

AI Artificial Intelligence.

AUC Area Under the Curve.

AUROC Area Under the Receiver Operating Characteristic.

AutoML Automated Machine Learning.

Bi-LSTM Bi-directional Long Short-Term Memory.

CDSS Clinical Decision Support System.

CICST Category Imbalance and Cost-Sensitive Thresholding.

CNN Convolutional Neural Networks.

Crisp-DM Cross-Industry Standard Process for Data Mining.

DRF Distributed Random Forest.

DT Decision Tree.

EHR Electronic Health Records.

FWCI Field-Weighted Citation Impact.

GB Gradient Boosting.

GBM Gradient Boosting Machines.

GLM Generalized Linear Models.

GRU Gated Recurrent Unit.

KNN K Nearest Neighbour.

LASSO LASSO Regression.

LIME Local Interpretable Model-Agnostic Explanations.

LR Linear Regression.

LRP Layer-wise Relevance Propagation.

LSTM Long Short-Term Memory.

LSV Linear Support Vector.

ML Machine Learning.

MLC Multi-Label Classification.

MLP Multilayer Perceptron.

MSI-PTDM Multi-Stream Integration Tuberculosis Diagnosis Model.

NB Naïve-Bayes.

NN Neural Networks.

NRS Numerical Rating Scale.

RF Random Forest.

ROC Receiver Operating Characteristic.

SHAP Shapley Additive Explanations.

SLC Single-Label Classification.

SMOTE Synthetic Minority Oversampling Technique.

SS-PTDM Single Stream Tuberculosis Diagnosis Model.

SVC Support Vector Classifier.

SVM Support Vector Machine.

XAI Explainable Artificial Intelligence.

XGBoost eXtreme Gradient Boosting.

1

INTRODUCTION

1.1. INTRODUCTION

The amount of medical data produced and collected in recent years has been growing at a staggering speed [3]. Each phase of a health-related procedure, ranging from scheduling appointments to conducting surgeries, is meticulously documented and preserved within data storage systems. The healthcare industry is experiencing a significant increase in the volume of data being collected, including a wide variety of data types [4]. These data types range from basic details, such as personal and demographic statistics [5], to more complex forms of information. For instance, healthcare data can include medical images and unstructured reports, such as clinical notes or medical reports [6].

In recent years, the medical field has substantially benefited from collecting large amounts of data [4]. This helped gain insight into patient health, disease patterns, and the effectiveness of various treatments [7]. Recognizing these benefits, the healthcare sector has begun to structure its data collection efforts to maximize the value of the data. Currently, healthcare professionals are often required to fill out reports using structured fields that have been pre-determined, rather than the previously used open-text fields [8]. Structured data, which is organized and easily searchable, has opened up new avenues for analysis and understanding, providing a richer, more comprehensive picture of healthcare outcomes [5].

As data collection in the healthcare industry has advanced, so too has the effectiveness of Machine Learning (ML) algorithms in terms of accuracy, efficiency, and practical utility, marking significant progress in both domains. The computational limits to create useful ML models have mostly been overcome [9]. In multiple industries, like finance or energy, ML has already shown how effective it is in predicting or classifying complex issues [10–12]. ML models are capable of assisting professionals with intricate tasks, such as supporting financial analysts and investors in predicting stock prices. Additionally, ML could execute entire processes au-

tonomously, as seen in energy management systems to save energy and reduce costs. Given the advancements in the field of ML, it is reasonable to expect that the medical field would also reap significant benefits. The vast array of medical data, available in diverse formats, aligns seamlessly with the capabilities of ML models. Despite the development of numerous effective models within the medical field, their translation into practical applications remains limited [13].

In healthcare, the impact of a decision directly affects the health of a patient. Therefore, a wrong decision could have detrimental effects. This means that there are multiple barriers to implementing ML models because of the in-depth expertise needed in computer science and data analysis. Additionally, there is a requirement for in-depth expertise in other domains like health science and decision science [14]. Moreover, the interpretability of ML models presents a significant challenge to their practical implementation. For ML models to be integrated into existing processes, they must first earn the trust of medical experts. Models that operate as 'black boxes', providing only a final output without any explanation, are insufficient for the rigorous demands of the medical field [15–17].

1.2. RESEARCH BACKGROUND

Company X is a leading healthcare provider in the Netherlands, with an extensive network of over 280 locations nationwide. Operating under Healthcare Group Y, a conglomerate of enterprises in the healthcare industry, Company X leverages this affiliation to provide comprehensive and specialized care services to its clientele.

The recent rapid growth of Company X and Healthcare Group Y is evident, characterized by the assimilation of numerous companies. This expansion, resulting in a doubling of the group's size in the last 2 years, highlights the increasing demand for specialized services. This surge in acquisitions signifies a significant influx of new patients seeking healthcare. Additionally, the challenges posed by demographic shifts are apparent. A report commissioned by the Dutch government emphasizes the escalating demand for healthcare services due to an ageing population and a rise in chronic illnesses. It also highlights the strain of a shrinking healthcare workforce and escalating costs. Without intervention, accessibility to quality healthcare for all is at risk [18].

In pursuit of providing the best care possible for its patients, Healthcare Group Y recognizes the challenge of maintaining quality standards amidst rapid growth. The influx of new patients and practitioners raises concerns about potentially lowering quality. Therefore, the company is actively seeking ways to standardize its healthcare processes. Standardized processes lead to greater control, enabling continuous improvements and ultimately resulting in better patient care.

To create these standardized processes, the data-driven care department has developed structured forms in collaboration with practitioners. These forms facilitate structured data collection during the different processes at Company X, ensuring that relevant information is cap-

tured effectively. These processes may include patient screening or diagnostic interviews, all of which are integral to comprehensive patient care. By providing a standardized framework for assessing patient problems, these questionnaires offer practitioners a systematic approach to understanding patient needs. Furthermore, the collected data can be analyzed to assist practitioners in decision-making processes, thereby enhancing the quality of care provided to patients. This research centres on the anamnesis process, which marks the initial encounter between the practitioner and the patient. During this appointment, the focus is on discussing the patient's medical history and current symptoms. The goal of this conversation is to gather insights into the patient's past medical experiences and their present condition, this process is further discussed in Section 4.2.

1.3. RESEARCH MOTIVATIONS AND OBJECTIVES

In recent months, a meticulously selected group of practitioners has filled in structured reports during their appointments with new patients. The selection of practitioners for the pilot program was intentionally diverse, encompassing a wide range of personalities and professional backgrounds. Creating a focus group of practitioners with different experience levels and motivations helps identify whether the reports capture data reliably and effectively. It showcases that less experienced or motivated practitioners also fill in these reports in a good manner, which could however also lead to a worse data quality. Now that more than a thousand structured reports have been filled in during the last year, a ML model could be trained. The primary task for this ML model would be classification, where the goal is to predict patient diagnoses based on reported complaints and other relevant data from the structured reports. Successfully implementing this ML model holds the potential to predict patient diagnoses accurately, offering invaluable assistance in establishing a standardized approach to patient care.

By harnessing the wealth of historical data encapsulated within these reports, the ML model can function as a complementary opinion for practitioners, providing consistent diagnostic insights across the various branches of Company X or Healthcare Group Y. This collaborative approach promotes providing patients with uniform and high-quality treatment, irrespective of the clinic they visit.

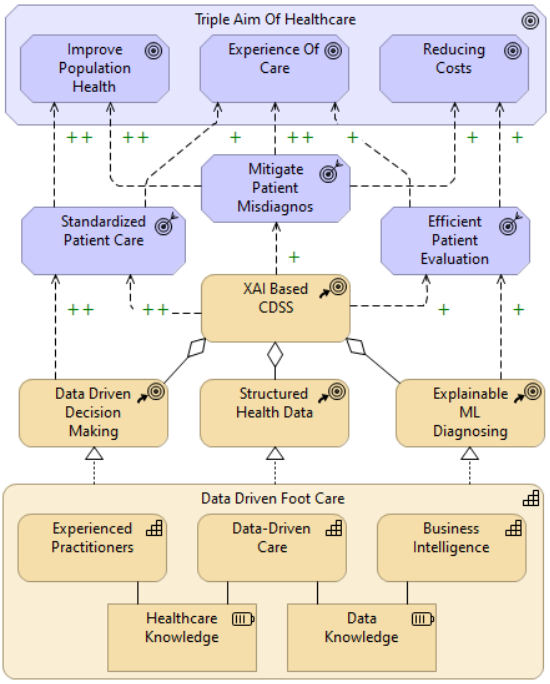


Figure 1.1: Motivation & Strategy View for Data Process

The primary objective of this research endeavour is twofold: firstly, to discern the most influential variables within the reports for disease diagnosis, and secondly, to develop an Explainable Artificial Intelligence (XAI)-based Clinical Decision Support System (CDSS) to aid practitioners in their diagnostic processes. It is important to assess the data quality for predictive purposes thoroughly and to determine the requisite level of explainability necessary for practitioners to comprehend and trust the model's outputs. Such insights are pivotal for the eventual deployment of the model in real-world clinical practice. Upon successful deployment, multiple benefits are expected:

- **Standardized Patient Care:** Implementing a CDSS to assist practitioners in decision-making promotes a uniform approach to patient management. By leveraging data from previous diagnoses made by various practitioners, the CDSS provides consistency across patient consultations, regardless of the attending practitioner. This standardized approach supports providing patients with consistent and reliable care irrespective of the practitioner conducting the examination.
- **Efficient Patient Evaluation:** During an initial patient consultation, practitioners often encounter a multitude of potential diagnoses, requiring extensive research to confirm. By leveraging a Clinical Decision Support System (CDSS), practitioners can filter and prioritize likely diagnoses, significantly reducing research time and shortening the duration of appointments.
- **Mitigate Patient Misdiagnoses:** Identifying the correct diagnosis for individual patients can pose challenges for practitioners, particularly for those early in their practice. Mis-

diagnoses can result in inappropriate treatment plans, potentially compromising patient health and straining healthcare resources. Implementation of a CDSS can aid practitioners in achieving accurate diagnoses initially, thus minimizing the risk of adverse health outcomes and optimizing healthcare efficiency.

The benefits mentioned above are also illustrated in Figure 1.1. This figure highlights the overall goals of Company X, which centre around the triple aim of healthcare as described by Berwick et al. 2023. The triple aim framework focuses on improving healthcare costs, enhancing the experience of care, and boosting overall population health. The figure outlines three potential outcomes or benefits that align with the triple aim. These outcomes are achieved through specific courses of action, which are the crucial steps implemented during this research. At the bottom of the figure, the relevant stakeholders involved in this research are shown, highlighting their roles and capabilities. The established objectives serve as the basis for the main research question which is formulated below:

How to design and integrate a XAI based CDSS to identify the most influential variables and support practitioners in diagnosing patients?

To be able to answer this main research question, this research delves into different aspects through the formulation of sub-questions. These sub-questions serve to explore various important factors related to the main research question and are formulated below:

1. How can Electronic Health Records (EHR) data be used to classify patient diagnoses?
   (a) How can the EHR data be extracted to be used for analysis?
   (b) What kind of variables should be used to classify diagnoses?
2. How can diseases be automatically diagnosed during an anamnesis to support the decision-making of a practitioner?
   (a) What ML techniques can be used to classify diagnoses based on patient symptoms?
   (b) What type of evaluation metrics can be deployed to assess the performance of the deployed ML models?
   (c) How should a ML model be implemented into the current anamnesis process?
3. How can XAI methods be used to support practitioners in their decision-making during the anamnesis?
   (a) What XAI techniques could be used to improve the interpretability of ML models?
   (b) What XAI techniques are considered to be trustworthy by practitioners?
   (c) How can XAI techniques contribute in making practitioners critical of CDSS output?

The research adopts the Cross-Industry Standard Process for Data Mining (Crisp-DM) framework for data mining, utilizing H2O's ensemble ML models like Random Forest (RF). Model

performance is validated using standard metrics such as F1-Score and Sensitivity. XAI methods, including Local Interpretable Model-Agnostic Explanations (LIME), are employed for enhanced interpretability. Further elaboration on these methods is provided in Chapter 3.

This research makes the following important contributions to the literature. First, it provides practical insights into the use of structured EHR data to diagnose patients. Given that EHR data is a relatively new concept in the healthcare industry, use cases for implementing this into a working ML model are limited. Second, this research provides a comprehensive literature review of XAI in healthcare. This offers insights into how ML should be implemented into healthcare processes and what this explainability could resemble. Finally, this is complemented with a practical implementation of this explainability, where the research utilizes the literature review as the basis for its practical implementation of XAI. This research contributes the literature with five possible XAI outputs, from which one is created in a qualitative focus group including medical experts.

The subsequent chapters are structured as follows. In Chapter 2, a systematic literature review was conducted. This chapter highlights the trends of ML models in decision support systems and showcases XAI techniques used in a healthcare setting. Following, Chapter 3 discusses the methodology which was followed to construct this research. The chapter explains the CrispDM framework which structured this research. It also defines the ML models and XAI techniques used. In Chapter 4, the experimental set-up is explained, in which the process from data extraction to model implementation is presented. Following, Chapter 5 analyzes the results of this implementation. The chapter covers the results of the different ML models and also showcases what XAI methods should be used when the CDSS is implemented in practice. Finally, Chapter 6 covers limitations and future recommendations related to this research. It also includes a conclusion answering all of the above research questions.

2

LITERATURE REVIEW

2.1. METHODOLOGY

This systematic review utilized two different databases to retrieve scientific papers that directly contribute to answering the research questions. The first database that was utilized is Scopus [1]. Scopus is a comprehensive bibliographic database that is widely recognized and used for its broad coverage of scientific literature. In the context of this review, Scopus served as a valuable resource for sourcing papers specifically in the field of computer science. The second database that was used for this review is PubMed [2]. Unlike Scopus, PubMed is a free database that provides access to the MEDLINE database of references and abstracts on life sciences and biomedical topics. It primarily focuses on healthcare-related papers, making it an ideal resource for sourcing literature in the field of healthcare for this review. The combination of these two databases ensures a balanced information retrieval. Information about the technical aspects of Computer Science and the foundational elements of Healthcare is combined into a multidisciplinary view.

Once these two databases were selected, keywords were chosen to ensure a comprehensive search of the literature relevant to the research questions. These keywords are selected based on the research questions formulated in the previous section. The first set of keywords focuses on the general application of Artificial Intelligence (AI) and ML in the design, implementation, and deployment of CDSS in healthcare. Furthermore, the second set of keywords narrows it down to XAI inside of a CDSS. Both queries used over both databases exclude mobile and remote in the title because these are types of papers that are assumed to be meant for CDSS supporting patients directly instead of aiding healthcare professionals.

1. 'AI' OR 'Artificial Intelligence' OR 'Machine Learning' OR 'ML'

[1] https://www.scopus.com/home.uri
[2] https://pubmed.ncbi.nlm.nih.gov/advanced/

AND 'Implementation' OR 'Design' OR 'Deployment'
AND 'CDSS' OR 'Clinical Decision Support System' OR 'Clinical Decision Aid' OR 'DSS'
AND 'Healthcare' OR 'Medical Care'
AND NOT 'Mobile' AND NOT 'Remote'

2. 'Explainable AI' OR 'XAI' OR 'Transparent AI' OR 'Interpretable AI' OR 'Explainable ML'
AND 'CDSS' OR 'Clinical Decision Support System' OR 'Clinical Decision Aid' OR 'DSS'
AND 'Healthcare' OR 'Medical Care'
AND NOT 'Mobile' AND NOT 'Remote'

Once these keywords were selected, a second filter needed to be made. The process of this second filter can be found in Figure 2.1. This filter was first based on the article type. For the Scopus database, only journal papers were selected, while for Pubmed, journal papers and literature review papers were selected. After this, the papers were filtered to only include free papers written in English, after 2018 in a certain field. For Scopus, this field was Computer Science, while for Pubmed, this field was the medical field. Table 5.2 summarizes the in- and exclusions during this phase. The final queries used on Pubmed and Scopus can also be found in A.

Table 2.1: Table containing a summary of the criteria used to select the articles

Criteria	**Decision**
Pre-defined keywords are included in the title, abstract, or in the keyword list of the paper	Inclusion
The paper was published in a scientific journal and written in English	Inclusion
The paper was published before 2018 or not available for free	Exclusion
Duplicates of an original paper	Exclusion
The paper's abstract, title, and content are not relevant to the research objective	Exclusion

After the final filtering, the results of the four different queries are combined, and duplicates are removed to create a list of 54 unique papers. From this list, a manual selection identifies the most relevant studies for the literature review. The manual selection process involves an initial screening, where relevant information and keywords are extracted from the titles, abstracts, and results of each paper. This information is then analyzed to determine which papers address the main research question or sub-questions. For instance, the research focuses on CDSS systems that support medical experts. Therefore, papers discussing CDSS systems designed to directly support patients are excluded. Similarly, papers on XAI methods for image data are excluded, as they are not relevant to the study's focus. This careful selection process ensures the inclusion of papers addressing problems similar to those in this research. Ultimately, 26 papers are identified for in-depth analysis in the literature review.

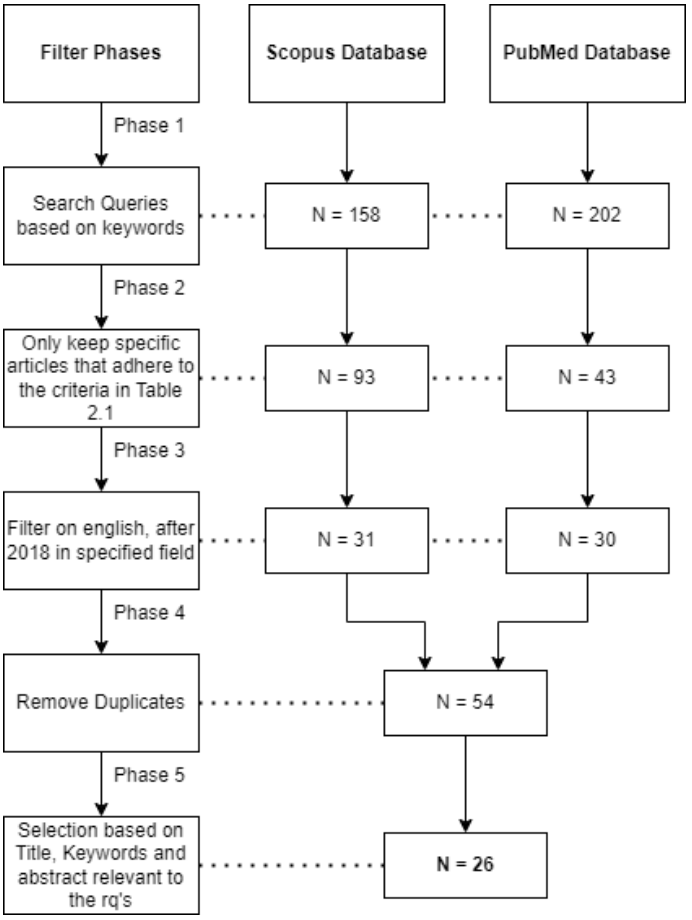


Figure 2.1: Articles selection process

2.1.1. TECHNICAL EXPLORATION

The final selection of papers mentioned above contains great and relevant information about the research question. Given that the ultimate research is predicated on a ML model used with structured EHR data, this subject matter was crucial for important information. Therefore, an additional search was conducted using the keywords "Machine Learning" and "Electronic Health Records", which further enriched the existing literature. This supplementary search resulted in the inclusion of five additional papers to increase the knowledge of EHR information in the literature that was already found [5, 20–23].

2.2. RELEVANT THEMES AND TRENDS IN LITERATURE

This section provides a comprehensive overview of the themes and trends that have emerged in the literature under review. It is divided into three subsections, each focusing on a specific aspect of the literature. The first subsection delves into the recent trends in the literature, examining the years in which the papers were published. This analysis allows us to identify patterns and shifts in the focus of research over time, providing a temporal perspective on the evolution of the field. The second subsection focuses on the quality and type of papers that have been found in the literature. The quality of the papers is assessed using a specific set of criteria, pro-

viding an objective measure of their academic rigour and contribution to the field. Additionally, this subsection categorizes the papers into types such as systematic reviews, case studies, review articles, and case studies. Finally, the third and final subsection explores the keywords used by the authors in their papers. This analysis reveals the main focus points of the research, highlighting the themes and topics that are currently at the forefront of academic discourse in the field.

The complete list of literature consulted during this research can be found in Tables B.1, B.2, and B.3. Table B.1 contains information about the literature papers focused on the implementation of ML models in a healthcare setting. This Table highlights the data types used in this research, the ML models, and the evaluation methods of these models. It also shows the results of the research and some important challenges faced. In Table B.1, two significant papers are highlighted. The first one is by Pavon et al. [21], which uses structured data to predict functional impairment. The paper demonstrates how clustering classifications together can handle missing data and make better predictions. The second paper is by Wang et al. [20], which focuses on predicting tuberculosis using EHR. This paper combines structured EHR data with unstructured report data and provides a comprehensive architecture of how the model is used and continuously trained.

Table B.2 contains information about literature focused on the topics XAI. The table covers the different XAI techniques, how they help in gaining the trust of the end-users, and useful information on what the explanation of the output should look like. A paper by Naiseh et al. [24] provides insights into how different XAI outputs can help in gaining trust. It presents useful visualizations of these outputs and their pros and cons, emphasizing that humans are more willing to engage with explanations when they are familiar, simple, and casually relevant. Another paper by Barda et al. [13] discusses the importance of the context in which the explanation is provided. It suggests that the XAI technique depends on who the explanation is provided to and why it requires an explanation.

Finally, Table B.3 is focussed on the implementation of CDSS and its impact on the triple aim of healthcare [19]. A paper by Jia et al. [25] proposes implementations of different XAI methods and their contribution to the safety inside CDSS models. It also shows the life-cycle of a ML system from development, to usage and the feedback loop for when it's in production. This paper highlights that even if a model's rules can perfectly predict outcomes on a test dataset (i.e., they have 100% accuracy), this does not automatically inspire confidence in those rules among clinicians. Therefore, the paper underscores the importance of explainability in ML models used in healthcare. It's not enough for a model to be accurate; it also needs to be understandable and trustworthy to the clinicians using it [16, 26].

2.2.1. EXPLORING LITERATURE TRENDS

The exploration of literature trends indicates that the papers found are recent, spanning the years 2019 to 2023. As visualized in Figure 2.2, a majority of the papers have been published in

the last two years. This trend can be attributed to search queries using relatively new keywords or emerging techniques. Keywords such as ML and CDSS are relatively new in the healthcare industry. The integration of these advanced technologies into healthcare practices has gained momentum only in the last few years, leading to a surge in related research and publications. Furthermore, The growing interest in XAI has led to a rise in research exploring its applications in healthcare [1]. XAI aims to make the decision-making process of ML models transparent and understandable to human users [27].

The spike in papers in 2023 [1, 5, 9, 16, 20, 21, 23, 24, 26–32], showcased in Figure 2.2, can also be explained by the extra literature included from Section 2.1.1. These papers mainly focussed on EHR data, which is an emerging term in the literature. Healthcare companies are investigating strategies to fully leverage their data. The approach of standardizing processes and capturing structured data, like EHR data, represents one notable discovery [29].

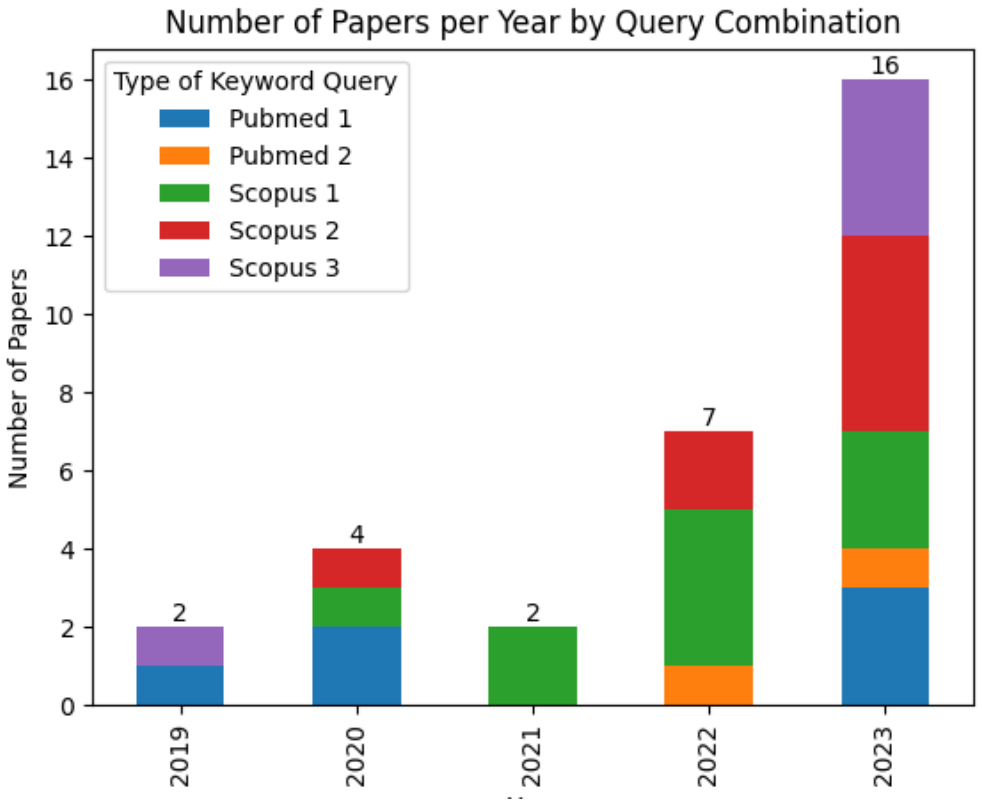


Figure 2.2: Year Distribution of Papers per Query Type

2.2.2. TYPE OF RESEARCH

This section highlights the types of research that were analyzed, as can be seen in Table 2.2. It is important to include different types of research to create an inclusive overview of the literature. Almost a third of the papers analyzed are systematic review papers. The critical analysis of these reviews was instrumental in providing a comprehensive understanding of the best practices and pitfalls in the implementation of CDSS and XAI.

In addition to the types of research papers, the impact of these papers on the academic community was also considered. This was measured through the number of citations and the Field-Weighted Citation Impact (Field-Weighted Citation Impact (FWCI)) of each paper. The impact factor, the last column in Table 2.2, is a metric provided by Scopus that reflects the yearly average number of citations to recent articles published in that journal. Assessing the impact of research articles requires the consideration of metrics such as FWCI, which provides a standardized measure accounting for disciplinary differences in research behaviour. FWCI adjusts for the expected number of citations a paper should receive, with different weight factors for different fields. For instance, the medical field tends to produce a significantly larger number of papers than the mathematical field. Therefore, citations in fields with fewer papers are weighted more heavily than those in fields with more articles. An article that is not included on Scopus or that does not have any citations, does not have this score. Papers with a high impact factor and/or a large number of citations are considered to have a significant impact on the literature. For example, the paper of Giuste et al. [27] has retrieved significant attention, which can be seen by its 14 citations and impact score of 15.17. Papers with a higher score often represent key findings or innovative ideas in the field, and their high citation count indicates that many other researchers have built upon or responded to this work in their research. These types of papers are therefore crucial in shaping this literature review.

Table 2.2: Review of Type and Quality of Literature

Author	Type of Research	#Citations	Impact[3]
(Barrera Ferro et al. 2020)	Empirical study	19	1.67
(Bartels et al. 2022)	Perspective Article	2	0.53
(Choudhury 2022)	Observational study	4	1.06
(Naiseh et al. 2023)	Comparative study	11	11.13
(Barda et al. 2020)	Qualitative Research	37	2.79
(Bienefeld et al. 2023)	Review Article	4	5.61
(Blanes-Selva et al. 2023)	Case study	2	2.91
(Murari et al. 2022)	Case study	0	-
(Hasan et al. 2021)	Observational study	7	0.9
(Jagadamba et al. 2023)	Case study	0	-
(Aiosa et al. 2023)	Case study	0	-
(Tapia-Galisteo et al. 2020)	Design Study	2	0.17
(Pumplun et al. 2023)	Design Study	0	-
(Calisto et al. 2022)	Case study	31	8.28
(Chen et al. 2021)	Case study	27	2.99
(Jia et al. 2022)	Systematic Review	13	3.13
(Giuste et al. 2023)	Systematic Review	14	15.17
(Du et al. 2022)	Comparative User Stu.	4	0.96

[3] FWCI score Scopus: https://service.elsevier.com/app/answers/detail/a_id/14894/supporthub/scopus/

(Lambert et al. 2023)	Systematic Review	4	2.96
(Mahadevaiah et al. 2020)	Systematic Review	49	6.94
(Nuutinen and Leskelä 2023)	Systematic Review	0	-
(Magrabi et al. 2019)	Systematic Review	105	3.9
(Iqbal et al. 2023)	Review Article	1	-
(Rundo et al. 2020)	Review Article	47	1.32
(Antoniadi et al. 2022)	Case Study	4	1.44
(Di Martino and Delmastro 2023)	Survey Article	1	-
(Wang et al. 2023)	Case Study	0	-
(Khodadadi et al. 2023)	Case Study	0	-
(Pavon et al. 2023)	Case Study	0	-
(Raita et al. 2019)	Case Study	191	11.67
(Chiu et al. 2023)	Case Study	1	-

Incorporating FWCI scores into research is essential for gauging the impact and relevance of academic papers. High FWCI scores, papers such as Naiseh et al. (11.13), Giuste et al. (15.17), and Raita et al. (11.67), showcase significant recognition within their respective fields, indicating influential contributions widely acknowledged by the academic community. Including such high-impact papers in a literature review improves the credibility and foundation of this research. However, a balanced approach incorporating papers with varying FWCI scores, including moderate and low ones, ensures a comprehensive review that includes established, emerging, and potentially innovative research areas. While high FWCI scores validate the inclusion of certain papers as foundational or highly influential, moderate and low scores illuminate the breadth of research and ongoing field development, aiding in identifying research gaps and future directions.

2.2.3. KEYWORDS

Understanding the keywords used in the literature allows us to gain a deeper understanding of the prevailing interests and priorities of researchers in the field. The authors choose the keywords to reflect the core ideas and themes of their research. The frequency of these keywords is particularly important as it indicates the primary focus areas of the literature. The frequency of these keywords has been visualized using word clouds, where the most predominant terms are shown as the largest words.

Figure 2.3: Word cloud of keywords in papers

The initial word cloud, as depicted in Figure 2.3, provides a snapshot of the primary themes prevalent in the reviewed literature. The terms “Artificial Intelligence”, “Machine Learning”, “Clinical Decision Support System”, and “Electronic Health” emerge as the most frequently occurring keywords in the research corpus. These topics dominate the discourse in the analyzed papers, reflecting their centrality in the field under study. This is not surprising, given that these were the key terms employed in the literature search across various databases. Therefore, another word cloud is created.

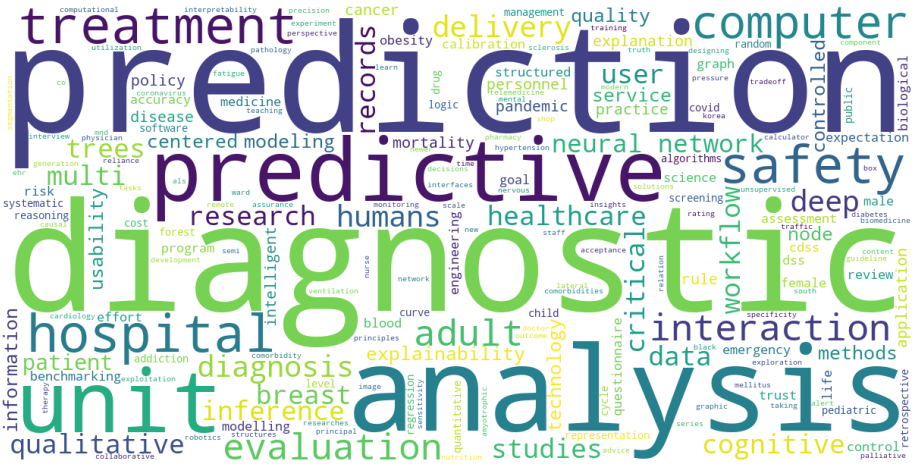


Figure 2.4: Word cloud of keywords in papers without search query keywords

To gain a more nuanced understanding of the literature, a revised word cloud was generated, excluding the initial search keywords. This refined word cloud, as illustrated in Figure 2.4, offers a more in-depth perspective on the thematic landscape of the papers. The first aspect that stands out is the "prediction", "diagnostic" and "analysis" keywords. These refer to the use case of the ML models. Second, in this revised visualization, keywords such as "safety", "interaction", "centred modelling", "evaluation", and "critical" gain prominence. This shift in keyword frequency suggests a paradigmatic shift in the literature towards explainability and

user-centricity. It underscores the increasing emphasis on the end user in the design and implementation of the final product.

2.3. MACHINE LEARNING IN HEALTHCARE DECISION SUPPORT

This section discusses the use of ML models in healthcare and their application in CDSS models. First, the types of ML research that was gathered are discussed. Following, the impact of ML based CDSS on the triple aim will be covered. Finally, the implementation of CDSS in healthcare will be highlighted.

2.3.1. MACHINE LEARNING RESEARCH

This section reviews ML research, analyzing the models and data used in various studies. It provides detailed insights into five papers, discussing their settings, use-cases, and results. Additionally, it explores how different challenges were addressed in the literature, which can also be found in Table B.1.

Setting

From the medical papers included in this literature review, two models have emerged as the most frequently utilized: Neural Networks (NN) and Support Vector Machine (SVM), both of which are often employed as ML approaches [17, 31]. A multitude of studies have incorporated eXtreme Gradient Boosting (XGBoost) in their final models or as part of the models tested [7, 20–23, 35]. In a similar vein, NN have been extensively used in the final model or the models tested across a wide array of researches [4–6, 20, 22, 23, 26, 29, 30, 33, 35]. The data utilized in these studies varies in form, ranging from structured to unstructured data types. However, some research also combined these two types of data [4, 6, 20, 23, 29].

Aiosa et al. [26] predicted obesity comorbidities diagnoses of patients using unstructured text datasets. They gave feature importance values, in which it was observed that BMI is the most important feature. The models developed using Multilayer Perceptron (MLP) and XGBoost were classified as the best prediction models. Following this, the research by Tapia-Galisteo et al. [6] aimed to predict cocaine inpatient treatment success using high-dimensional unstructured data of healthcare reports. Their best classifier, the Random Forest (RF), which is an ensemble method composed of multiple decision trees, achieved an accuracy of 82.12%, outperforming models like SVM, Linear Regression (LR) and MLP. This can be used in identifying patients who may need extra attention to prevent them from dropping out of care.

Wang et al. [20] diagnosed patient with tuberculosis using EHR data. They used an oversampling technique to counter the imbalanced data. The Multi-Stream Integration Tuberculosis Diagnosis Model (MSI-PTDM) was the highest with 90.91% accuracy, which is a model that can process sparse data, dense data, and unstructured text data concurrently [20]. However SVM and XGBoost also had great results. This research could be used to support clinicians in diagnosing patients with tuberculosis. Khodadadi et al. [5] developed an end-to-end approach for learning patient representations from tree-structured information for readmission and mortal-

ity prediction tasks. They used structured EHR data including demographic data, and personal data, and had a lot of missing data and imbalanced classes. They used Random Forest (RF) and extracted an Area Under the Receiver Operating Characteristic (AUROC) score of 0.87. This research can be used to support medical experts and management in predicting mortality. Lastly, the research by Pavon et al. [21] created a scalable process for the identification of functional impairment using structured EHR data. This research used k-means clustering and XGBoost to enable population-based strategies for identifying functional impairment, and precise targeting of prevention or treatment resources within health system populations to patients likely to benefit most.

In conclusion, the papers we analyzed demonstrated diverse applications of ML within healthcare. Some papers utilized ML models to diagnose various diseases, including diabetes and tuberculosis [3, 5, 20, 26]. On the other hand, several other papers aimed to predict future cases [6, 22, 23, 33, 35], This prediction could be related to the no-show behaviour of patients or predicting clinical outcomes. Table B.1 includes more findings of different ML based research. The diversity in the objectives of these studies underscores the versatility and potential of ML in healthcare, opening up new avenues for patient care and management.

Challenges Faced

This section delves into the various challenges encountered in the reviewed literature, Table B.1. A recurring theme was class imbalance, where certain diseases or diagnoses were more prevalent than others. Wang et al. [20] addressed this by employing down-sampling and up-weighting strategies. Here, down-sampling refers to reducing the frequency of the dominant classes, while up-weighting involves assigning greater weights to the less frequent classes. On the other hand, Khodadadi et al. [5] chose to maintain the data imbalance and trained their model accordingly.

Another issue pertains to the presence of missing data in the studies. Pavon et al. [21] tackled this by using a clustering approach, where data clusters with missing values were assigned the same value. Subsequently, Chiu et al. [23] utilized XGBoost, which is known for its superior handling of missing data. Lastly, the task of identifying the most significant features posed a challenge. Aiosa et al. [26] employed feature tables to discern the most crucial features, while Tapia-Galisteo et al. [6] used a correlation-based feature search to determine the features to be included in the model.

2.3.2. IMPACT ON TRIPLE AIM

The implementation of ML based CDSS has had a profound impact on the triple aim of healthcare. The triple aim of healthcare is a framework developed to improve and optimize overall care and health [19]. It consists of the following three categories:

1. **Experience of Care:** Improving the experience individual patients have when accessing medical care

2. **Improving Population Health:** Improving the quality of the medical processes and acting preemptively
3. **Reducing Costs:** Control costs more effectively and create more value for patients for less money

The framework can be used to determine the benefits of implementing new technology. Below, we cover the impact on the three categories of the triple aim framework.

Experience of Care

ML-based CDSS have been instrumental in enhancing the quality of care. By providing more accurate and efficient protocols [28, 32], these systems have improved service quality. They have also increased patient safety by reducing advice against protocol [28, 32]. Furthermore, they have contributed to improved diagnostic accuracy [15], thereby enhancing the overall patient experience.

Improving Population Health

ML-based CDSS have also shown promise in improving population health. They have improved clinician satisfaction [9] and have proven beneficial for training novice clinicians [9]. Moreover, practitioners have observed that AI could enhance consistency in healthcare delivery [31, 34], thereby contributing to better population health outcomes.

Cost Reductions

Lastly, ML-based CDSS have contributed to cost reductions in healthcare. By speeding up processes, these systems have enabled healthcare providers to assist more patients [15]. They have also reduced repetitive tasks, freeing up clinicians to engage in other, more critical activities [38]. Additionally, by reducing instances of incorrect diagnoses, they have led to fewer costs associated with misdiagnosis [32].

2.3.3. IMPLEMENTATION CDSS IN HEALTHCARE

Implementing CDSS systems into healthcare has been challenging. Despite the potential benefits of CDSS, only a small fraction of these systems have found their way into existing healthcare processes [14].

The adoption rate of ML based CDSS is surprisingly low. van de Sande et al. [39] revealed that 89.3% of their papers did not make it past the prototyping or development phase. There are different reasons for this low adoption rate. First, practitioners are likely to prioritize risk factors [34], medical professionals do not have the feeling that the systems are accurate enough to provide support in their decision-making. Adding to this, many practitioners are unfamiliar with Artificial Intelligence (AI) and technology, this further increases the fear of wrong outputs of the CDSS [28, 34, 38]. When these employees do not trust the output, the model will not be used. There is a fear among healthcare professionals of being replaced or losing their autonomy [9, 28, 34]. It was stated in multiple papers that clinicians fear that the system would start taking over and their knowledge would not be necessary anymore. Finally, there is often a

misalignment between expectations and the actual output of the system [28, 31].

Mitigating the discrepancy between the output of a ML model and the expectations of the users is crucial. Stakeholders should therefore be included in the development stages of the project. Their knowledge and goals should be included when designing the CDSS [25, 36, 37]. The ability for clinicians to accept or decline the CDSS's recommendations can enhance their sense of control and positively influence their experience with the model [15]. Finally, quality measures and assurance of the model are needed to guarantee safety [7, 14, 34, 36].

When it comes to integrating ML by implementing it into a medical process, several strategies have proven effective. Establishing a helpdesk or central point for ML malfunctions has been beneficial [14]. Forming a multidisciplinary team for model output has also shown promise [28, 38]. The importance of training and communication cannot be overstated [9, 28]. Rigorous initial and ongoing evaluation is critical to ensuring safe and effective integration [36]. Lastly, the system should be designed in a way that it does not take too much time to document or use [38].

2.4. XAI IMPLEMENTATION IN HEALTHCARE

This section covers the implementation of XAI in healthcare and contains three subsections. The first subsection focuses on the different XAI methods. Following, the second and third subsection highlights different parts of the explainability process. The second subsection covers the ante-hoc explanation and the third subsection describes the post-hoc explanations.

2.4.1. XAI METHODS

Transparency in AI decision-making is crucial in many fields, particularly in healthcare, where XAI has shown significant benefits in CDSS [17]. Medical experts advocate for XAI due to its potential to enhance the interpretability and explainability of AI models, which is essential for ensuring their reliability and safety in healthcare settings. They need to understand why a particular decision or recommendation has been made to trust the system, especially when it contradicts their clinical judgment. There are different methods to make a model explainable, and it is not always clear what works best for the end-users of the models [13].

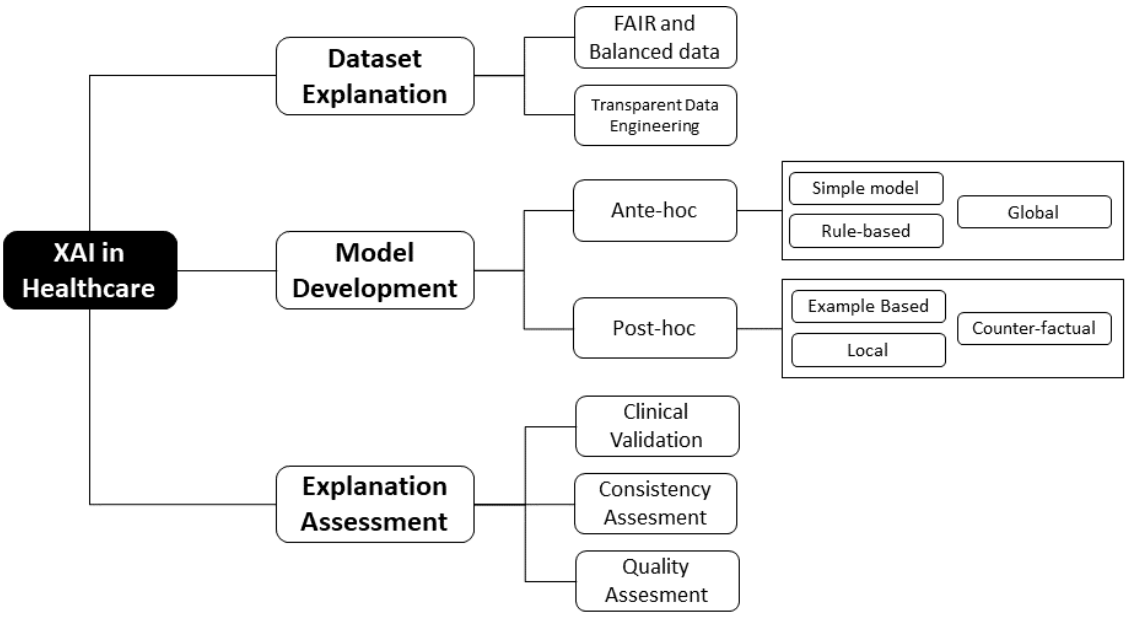


Figure 2.5: XAI Methods in Healthcare Overview, Inspired By Research Di Martino and Delmastro [1]

In the exploration of XAI methods, different researchers created a unique syntax and perspective of XAI methods. Jia et al. [25] distinguish between explanations that are inherent to the model (model explanations) and those generated after the model has made a prediction (post-hoc explanations). Pumplun et al. [30] further categorizes explanations into global and local. Global explanations aim to provide an overall understanding of the model's behaviour, while local explanations focus on individual predictions. Di Martino and Delmastro [1] propose a temporal perspective, dividing the explanation process into pre-modeling, model-development, and post-modeling stages. Interestingly, they further divide the model-development stage into ante-hoc (before the fact) and post-hoc (after the fact) explanations. Naiseh et al. [24] delves deeper into post-hoc explanations, dividing them into four different categories:

1. **Local Explanation:** It quantifies the importance of each input data feature to the recommendation.
2. **Example Based:** The AI justifies its decision by providing examples from the dataset with similar characteristics.
3. **Counterfactual:** The AI answers "what-if" questions to observe the effect of a modified data feature on the recommendation.
4. **Global Explanation:** It explains the overall logic of the model, including presenting the weights of different data features as decision trees, rules, or ranking styles.

The findings about XAI techniques during this literature research and the different perspectives from these papers have led to Figure 2.5. This overview provides a clearer picture of the current state of XAI and the different aspects that can be focussed on during the design and implementation of XAI systems.

2.4.2. ANTE-HOC EXPLANATIONS

Ante-hoc explanations in XAI refer to the explainability of the model itself [1]. These explanations are used to distinguish important features and interpret the inner structure of the model [1]. Ante-hoc explanations can also be a simple model or a rule-based approach, as demonstrated in the research by Chen et al [3].

Ante-hoc XAI plays a crucial role in evaluating reliable output, improving trust, revealing new insights, identifying potential weaknesses, and tuning [26]. Moreover, XAI can help in the training and creation of machine learning models. It provides insights into the model's decision-making process, which can lead to a better understanding of the model's behavior and, consequently, improved performance [27].

During the training phase, XAI can highlight important variables and features [17]. This allows for a more focused and efficient training process as the model can concentrate on the features that significantly influence the predictions. By understanding the importance of different features, engineers can fine-tune the model to capture the underlying patterns in the data better.

Moreover, XAI methods can help improve models based on better understanding and faster debugging. For instance, if a model is not performing as expected, XAI can provide insights into why the model is making certain decisions. This can help identify any issues or biases in the model, leading to targeted measures to improve the model [16]. It is also important that this information is supported by domain knowledge, to create better output [13]. Overall, XAI can be critical in making an understandable and fair model. However, XAI can also lead to worse performance, so the trade-off between model explainability and performance also has to be taken into account [25]

2.4.3. POST-HOC EXPLANATIONS

Post-hoc explanations in XAI refer to the interpretability methods applied after a model has made a prediction. These methods aim to shed light on the reasoning behind the model's decision, thereby enhancing transparency and trustworthiness [24].

There are several types of post-hoc explanations, with the most commonly used models being LIME and Shapley Additive Explanations (SHAP) [25, 40, 41]. These models provide visualizations that help users understand the contribution of each feature to the prediction.

To gain trust and reduce errors, it is essential to provide training or tutorials on how to interpret these visualizations[24]. However, the implementation of these explanations should take into account the specific needs of the target group, as different stakeholders may have different goals [16].

When displaying the final explanation, it is crucial not to overwhelm the user with too much information, as this could lead to the user skipping important details[13, 24]. The techniques used to present the information should depend on the audience[13]. Consistency and good interpretation are vital for effective communication[27], and evidence is needed at the cohort

level[1]. While there is a need for more assessment of explanation methods, it is crucial to involve end-users in the design process. They need to understand how the explanation works and what information is important[1, 13, 24].

2.5. CONCLUSION

This systematic literature review investigated the implementation, design and deployment of XAI in CDSS within the healthcare domain. The research retrieved valuable information from both the Scopus and PubMed databases using queries based on specific keywords. The systematic nature of the literature search ensured the collection of relevant journal papers and research, enhancing the multidisciplinary perspective by synthesizing computer science and medical literature. The most crucial information of the literature is collected in Tables B.1, B.2, and B.3.

After the collection of the data, the study delves into the dominant themes and trends in the literature. This is done in Section 2.2, which also covers the research types in the literature. Following, Section 2.3, shifts the focus to an in-depth exploration of ML applications within CDSS and healthcare contexts, encompassing both research insights and practical implementations. Finally, Section 2.4 reviews the landscape of XAI methods, explaining how the literature has approached and implemented these methods in the context of CDSS. The findings of these sections created insights used to answer the research questions:

1. **Main trends and themes in literature:** The literature review reveals a significant trend of CDSS and XAI models in 2022 and 2023. The analysis of publication years indicates a pronounced focus on these topics. Furthermore, the types of research encompass a variety of methodologies, with systematic reviews and case studies being prominent. Finally, the themes of the literature show an emphasis on user-centric approaches and explainability, highlighting a shift towards more transparent and user-friendly applications of ML in healthcare decision-making.

2. **ML in healthcare and impact on triple aim:** The analysis of ML research in healthcare reveals a predominant use of models such as NN, SVM, and XGBoost across diverse medical settings. The studies varied in the usage of unstructured or structured EHR data. Importantly, the impact of ML-based CDSS on the triple aim of healthcare is significant, enhancing the patient experience, improving population health, and contributing to cost reductions through increased efficiency and accuracy in healthcare delivery.

3. **CDSS implementation healthcare:** The implementation of CDSS in healthcare faces significant challenges, leading to a low adoption rate. The reluctance among practitioners stems from practitioner concerns about accuracy, unfamiliarity with AI, fear of job displacement, and misalignment of expectations. Overcoming these challenges requires collaborative development with stakeholders, empowering clinicians to control system recommendations, implementing quality assurance measures, and transparent output of the CDSS.

4. **XAI methods:** XAI is integral to healthcare, ensuring trust and transparency in decision-making. The diverse landscape of XAI methods, as illustrated in Figure 2.5, encompasses ante-hoc and post-hoc explanations. Ante-hoc explanations, such as rule-based models, play a pivotal role during model training, aiding in the identification of crucial features and refining the model's understanding of data patterns. Post-hoc explanations, facilitated by models like LIME and SHAP, offer transparency into the decision-making process after predictions, providing visualizations that elucidate the contribution of each feature. To ensure effective implementation, it is crucial to tailor these XAI methods to the specific needs of stakeholders, incorporating user-friendly interfaces, and targeted tutorials, and involving end-users in the design process for optimal understanding and trust.

In conclusion, the literature review highlights a growing interest in CDSS and XAI over the past two years. The diverse research methodologies employed reflect a nuanced exploration, with a notable trend towards user-centric and transparent approaches in healthcare decision-making. ML research in healthcare, particularly using models like NN, SVM, and XGBoost, demonstrates a significant impact on improving patient experience, population health, and reducing costs. However, the implementation of CDSS faces challenges, necessitating collaborative strategies and transparent outputs. XAI remains integral for ensuring trust and transparency, with diverse methods tailored to stakeholders' needs. Involving end-users in the design process is crucial for effective XAI implementation and fostering understanding and trust.

3

METHODOLOGY

3.1. CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING

The Crisp-DM model is a framework used in the field of data mining and analytics. Its purpose is to provide a structured approach for conducting data mining projects, guiding practitioners through the various stages from understanding the business problem to deploying a solution [2]. The model is iterative, allowing for flexibility and adaptation as the project progresses, as shown in Figure 3.1.

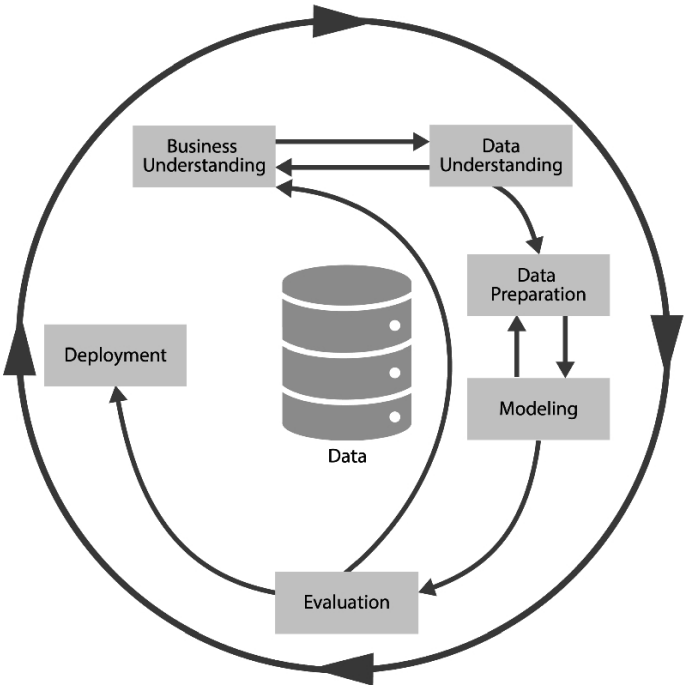


Figure 3.1: CRISP-DM Process Model Kristoffersen et al. [2]

1. **Business understanding:** This phase involves understanding the business objectives and

requirements for the project. It requires identifying the problem to be solved and determining how data science can contribute to addressing it. Finally, it includes gaining information about the processes and businesses involved.

2. **Data understanding:** The second phases focuses on collecting and exploring the available data relevant to the problem at hand. This includes identifying data sources, assessing data quality, and gaining insights into the characteristics of the data.

3. **Data preparation:** Once the data of the previous phase is collected, it needs to be analysed. This phase involves cleaning the data, integrating data from different sources, and transforming it into a suitable format for analysis.

4. **Modeling:** This phase is where the actual data analysis takes place. It involves selecting appropriate models and algorithms, building predictive or descriptive models, and evaluating their performance.

5. **Evaluation:** After building and testing the models, they need to be evaluated to ensure they meet the business objectives and requirements. This involves assessing the performance of the models, validating their accuracy, and determining their effectiveness in addressing the problem at hand.

6. **Deployment:**The final phase involves deploying the solution into the operational environment. This includes implementing the models into production systems, integrating them with existing processes, and providing support for ongoing monitoring and maintenance.

The Crisp-DM model offers a structured framework for initiating data science projects. By adhering to its defined steps, practitioners can systematically leverage data for various initiatives.

3.2. AUTOMATED MACHINE LEARNING

Automated Machine Learning (AutoML) is a process where ML models are created automatically with minimal human intervention. The goal is to make the process of building ML models more accessible to non-experts and to streamline the workflow for experts, allowing them to focus more on the problem at hand rather than the nitty-gritty details of model building.

AutoML systems typically handle tasks such as data preprocessing, feature engineering, model selection, hyperparameter tuning, and model evaluation. They leverage techniques like evolutionary algorithms, Bayesian optimization, and meta-learning to search through the space of possible models and configurations efficiently.

This research utilized the auto ML models from H2O.ai [1]. H2O.ai is a company that offers a popular open-source platform called H2O, which includes tools for AutoML. H2O.ai's AutoML functionality automates the process of training and tuning ML models, allowing users to quickly experiment with different algorithms and configurations to find the best model for their

[1] https://h2o.ai/

data. H2O.ai's platform is known for its scalability and performance, making it suitable for both small-scale and large-scale ML tasks. H2O provides multiple different kinds of ML models to be trained. The models used during this research are shortly explained below:

- **Gradient Boosting Machines (GBM):** GBM are ensemble learning methods that build a series of decision trees sequentially. Each tree corrects the errors of the previous one, leading to a strong predictive model [42].
- **Distributed Random Forest (DRF):** A DRF is an ensemble learning method that constructs multiple decision trees across a distributed computing environment to enhance accuracy and robustness. Each tree is trained on a subset of the data, often using different subsets on different nodes in the computing cluster. The final prediction is made by aggregating the predictions from all individual trees, typically through averaging for regression tasks or voting for classification tasks. This distributed approach allows for handling larger datasets and reduces computation time by leveraging parallel processing [43].
- **Generalized Linear Models (GLM):** GLM are a class of linear models that generalize linear regression to accommodate different types of response variables and error distributions. GLMs are widely used for regression and classification tasks and are particularly useful when the relationship between the predictors and the response variable is not linear [44].

3.3. EXPLAINABLE ARTIFICIAL INTELLIGENCE

3.3.1. PARTIAL DEPENDENCY

Partial dependency plots are a popular method for visualizing the effect of a set of features on the predictions made by a model. They can help to understand the interaction between these features and the target variable, and can also provide insights into the behavior of the model in different regions of the feature space.

3.3.2. LIME

LIME is a novel explanation technique that explains the predictions of any classifier in an interpretable manner, by learning an interpretable model locally around the prediction[40]. LIME provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one[40]. The utility of explanations via LIME has been demonstrated in various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted[40].

3.4. VALIDATION METHODS & METRICS

Validation methods and metrics are crucial in assessing the performance of a ML model. They provide quantitative measures that reflect how well the model will generalize to unseen data.

3.4.1. METHODS

Cross-Validation This research uses a validation method called x-fold cross-validation, one of the most common validation methods used in ML. This technique involves partitioning the data into x subsets or "folds". The model is trained x times, each time using $x-1$ folds for training and the remaining fold for validation. This process ensures that every data point is used for both training and validation, which provides a comprehensive evaluation of the model's performance [45].

3.4.2. METRICS

Precision Precision, also known as the positive predictive value, is the proportion of true positive predictions among all positive predictions [46]. It is given by the formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

A high precision indicates a low false positive rate, but does not take into account any false negatives.

Sensitivity Sensitivity, also known as recall or true positive rate, measures the proportion of actual positives that are correctly identified as such [46]. It is given by the formula:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

A model with high sensitivity will detect most of the positive examples, but may also produce many false positives.

F1-Score The F1-score is the harmonic mean of precision and sensitivity, and it tries to find the balance between these two values [46]. The F1-score is given by the formula:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

The F1-score is a good metric to consider if you need both precision and sensitivity to be high.

Accuracy Accuracy is the proportion of correctly predicted instances (both true positives and true negatives) among the total instances evaluated [46]. It is given by the formula:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

While accuracy is a simple and intuitive metric, it can be misleading in cases of imbalanced datasets where one class significantly outnumbers the other.

4

Experimental Set-Up

4.1. Experimental Set-Up

This study explores the potential of a ML-based CDSS within the domain of healthcare. The foundation of this CDSS will be EHR data, collected using structured reports created over a year ago. After a year of data retrieval, the dataset now contains sufficient information for deploying AI. This research will examine the performance and utility of the accumulated data, providing insights and guiding future advancements. In light of the advancements and motivations outlined in Chapter 1, the CDSS is expected to fulfill the following requirements:

- **Selection of Relevant Variables:** It is important to ensure that the CDSS selectively integrates variables directly linked to specific diseases. Unrelated variables should be excluded to maintain the interpretability of the CDSS. This precaution is crucial for fostering trust in its reliability among users.
- **Accurate Disease Diagnoses:** The foremost objective of the CDSS is to establish accurate diagnoses for patients to support practitioners in decision-making. While minor inaccuracies may occur, correct diagnostic options within its output are the most important factor. Misclassification of the CDSS can be rectified by practitioners during subsequent clinical evaluation, ensuring the accuracy of the final diagnosis. This should still be prevented as much as possible but has a lower priority than the correct classification being included.
- **Explanatory Transparency:** The CDSS should explain its decision-making process clearly and understandably. This transparency is essential for facilitating comprehension and critical evaluation by practitioners. The CDSS aims to improve clinical decision-making by offering insights into the reasoning behind its outputs. It is essential that this reasoning be validated to ensure it is reliable and useful for practitioners who will use the CDSS in practice. Therefore, this research will validate the transparency of the CDSS using a

questionnaire among practitioner and a focus group including domain experts.

Meeting these critical requirements presents significant challenges, necessitating thorough research into the problem statement. In Chapter 4, a structured approach using the Crisp-DM framework is adopted, with modifications tailored to the specific needs of the study. The experimental setup, depicted in Figure 4.1 , begins with a focus on understanding the business processes. This section of the experimental setup focuses on the anamnesis process. It is the only section on the business level, which can be seen by the yellow colour. Moving forward, the experimental setup extends to the application level, which starts with the data understanding. This step delves into the data generated by the anamnesis process, including the information contained within the three reports completed during anamnesis. With a grasp of both the data and business process, the crucial step of data preparation needs to be done. Here, the focus is to first model the data into a star schema, which offers a robust format for analysis and querying. Additionally, this phase involves data cleaning and feature enhancement to ensure the data is useable for ML tasks. This means preparing the data so that it is accurate, consistent, and in a suitable format for ML algorithms to process effectively. Following data preparation, the transition to the modelling phase is made. Here, three distinct types of machine learning models are developed: two Single-Label Classification (SLC) models tailored to accommodate varying class distributions, and one Multi-Label Classification (MLC) model integrating weight factors. Finally, the performance of the best models is evaluated using the metrics covered in Section 3.4, and options for explainability are explored. Practical implementation considerations are also factored into the evaluation process.

All of the processes and sub-processes mentioned in the experimental setup were developed through the use of several Python libraries and tools supported by Jupyter Notebook[1], a web-based development environment for creating data science projects. The Jupyter notebooks were mainly used for data explorations. Once a structured and trustworthy data pipeline was created, it was turned into a python script that could run periodically.

[1] https://jupyter.org

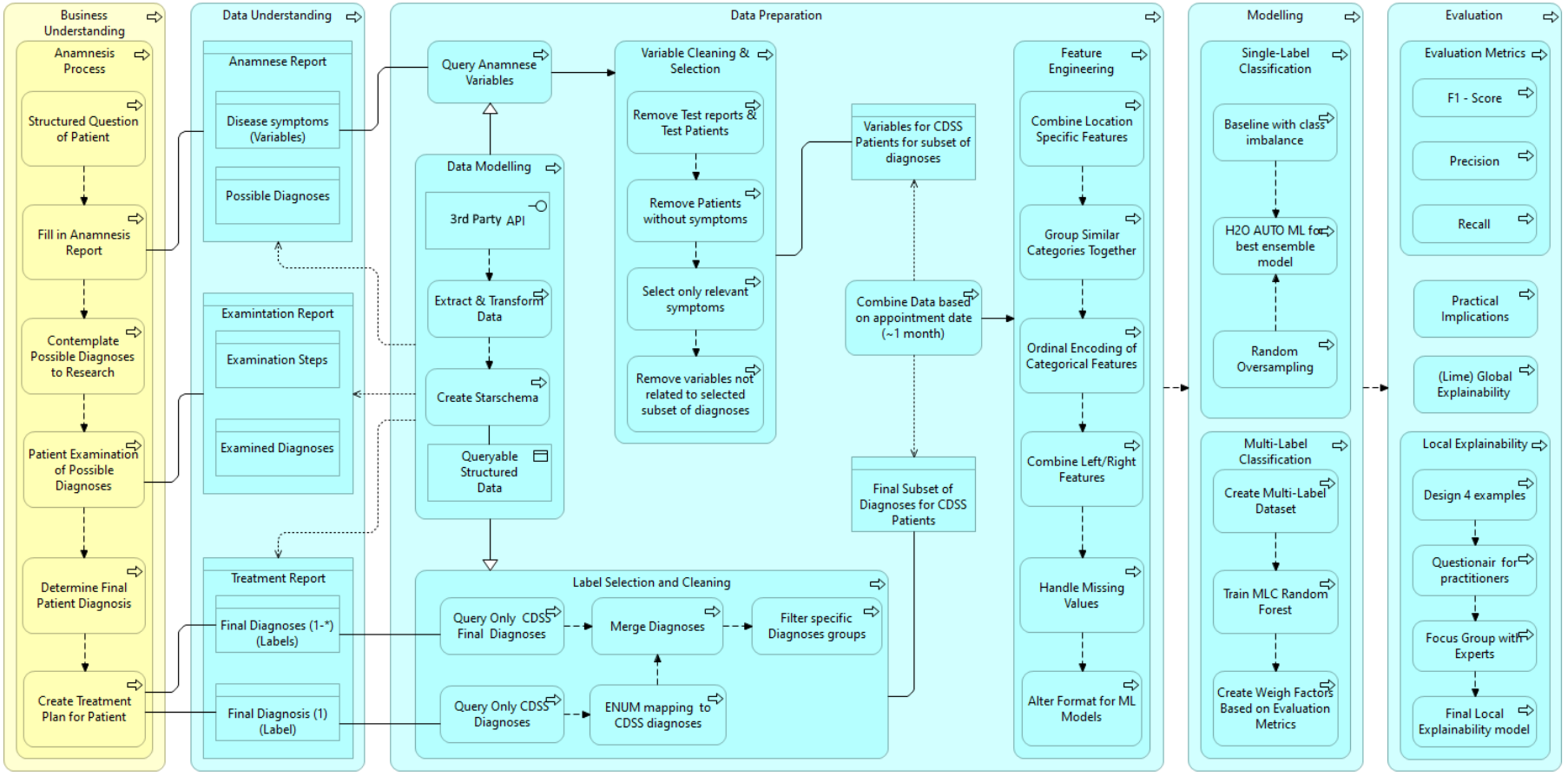


Figure 4.1: Overview of the Experimental Set-Up

4.2. BUSINESS UNDERSTANDING

The anamnesis process, a fundamental component of healthcare, is the primary focus of this study. An anamnesis involves a thorough questioning of past medical events, experiences, and pertinent information from patients. It serves as the initial appointment between a practitioner and a patient, crucial for understanding the patient's medical history and current symptoms. This comprehensive gathering of data is essential for accurate diagnosis and effective treatment [47]. For this research, the anamnesis process consists of more than just this anamnesis conversation. The process also includes the examination of the patient and the set-up of a treatment plan, as illustrated in Figure 4.2. The process typically initiates with the identification of a patient in need of treatment, either through self-observation or referral from other medical professionals such as physicians or physiotherapists. Subsequently, an appointment with a practitioner at the company is scheduled, marking the commencement of the anamnesis.

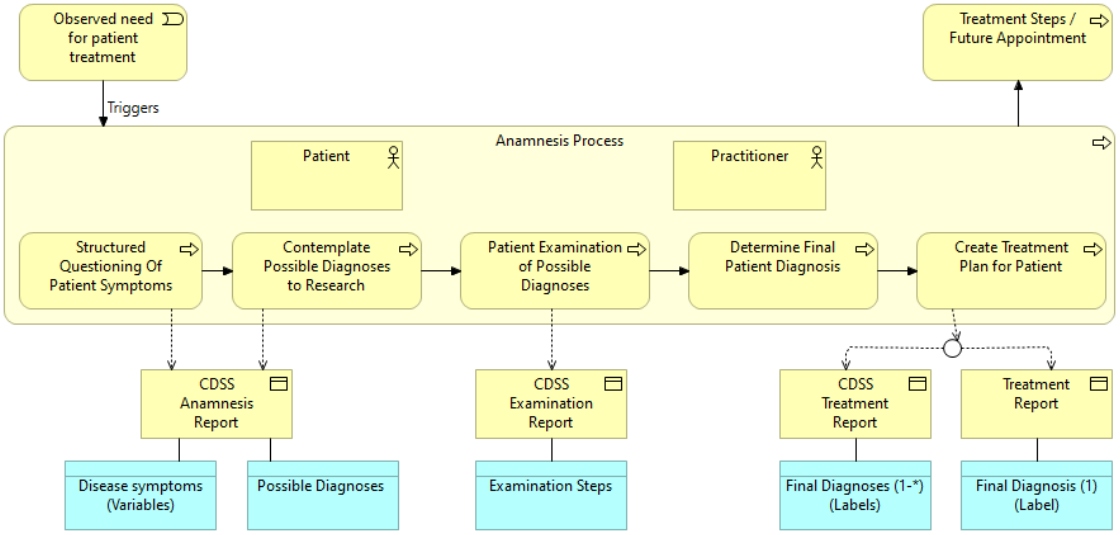


Figure 4.2: Business View of the Anamnesis Process

During the anamnesis appointment, the practitioner initiates the process by gathering the patient's medical history and current symptoms. This initial discussion serves as a crucial step in understanding potential underlying conditions. During this discussion, there are lots of different types of questions a practitioner could ask. For example, the practitioner asks the patient if they have consulted other healthcare professionals, inquires about the location of their issues or pain, and prompts the patient to describe their symptoms in detail, including the type of pain experienced or any inflammatory symptoms. This questioning gives the practitioner insights into the patient's condition, helping them formulate potential diagnoses. As mentioned in the introduction, a select group of practitioners currently fill in structured reports during this anamnesis process, which is the main focus point of this study. These structured reports will be adopted company wide when the current test phase is deemed a success. The reports offer a standardized format for gathering patient information, ensuring that important details are not overlooked. Furthermore, it also standardizes the data collection during the anamnesis

process, ultimately leading to more informed decision-making. Details regarding the standardization efforts and the impact on data understanding will be discussed in Section 4.3.

After the anamnesis conversation, the practitioner gains insight into potential conditions the patient may be experiencing. Typically, there are several possible diseases that the practitioner considers at this stage. Therefore, a comprehensive examination is essential to pinpoint specific diseases or issues. This phase entails an evaluation of the patient's pain location through observation or other examination techniques. Different types of examinations may be required depending on the suspected diseases, each tailored to identify the correct condition accurately. If the examination proceeds as planned, the practitioner can ascertain the patient's diagnosis, which may consist of one or multiple diseases.

Finally, after the examination is completed, a treatment plan is established. In this plan, the practitioner collaborates with the patient to discuss the next steps aimed at reducing symptoms or improving participation in day-to-day activities. This may involve prescribing specialized drugs, crafting custom-made gear, or referring the patient to another medical specialist. Currently, within the existing process, only one structured report is completed at this stage. In the current process, only one structured report is completed at this stage. However, in the CDSS pilot group, three reports are completed following this appointment: one for anamnesis, one for examination, and one detailing the treatment plan.

4.3. DATA UNDERSTANDING

As stated in the previous section and illustrated in Figure 4.2, three distinct reports are completed within the business process. For this section, our attention will be directed solely towards the first and final reports. This decision was made because the initial report contains the symptoms, akin to variables in our context, which will serve as inputs for the ML model during its training phase. Conversely, the concluding report contains the definitive diagnosis provided to a patient, the labels in our model's training dataset. Reports are filled in using a structure from an external company. This partner provides the interface for users to fill in these reports. After the report is filled in, the input data is sent to the database. Below, both reports will be explained profoundly and the data types in these reports will also be covered.

4.3.1. ANAMNESIS REPORT DATA

The anamnesis report contains information regarding the patient's medical background and prevailing symptoms. Within this dataset, the medical history component comprises a wide array of information, including past medications, treatments, surgeries, and consultations with medical professionals. However, for the purpose of this research, this dataset is excluded from the analysis. The rationale behind this exclusion stems from the limited correlation between this historical data and the specific disease under examination. Including such data could potentially introduce noise and bias into the predictive models due to their low relevance to the target outcome. Following, the symptom data is included in this research. This data encap-

sulates the patient's current symptoms, what kind of pain he has and where it is located. This data consists of three types: nominal, ordinal and interval. The rest of the section explains these three data types and their corresponding input values.

The first, and primary, data type utilized in the anamnesis report is nominal data. Represented as checkboxes within the report interface, this data type simplifies questions into binary options of 'yes' or 'no'. Table 4.1 showcases the different types of options a practitioner can fill in. Each checkbox corresponds to a specific question category, providing insights into various aspects of the patient's condition. For instance, one such category might pertain to the kinds of symptoms the patient has, in this case offering a selection of ten distinct options. Notably, patients may present with multiple symptoms, allowing for the selection of multiple checkboxes or none at all, thus ensuring comprehensive data capture. The amount of filled-in features can be found in Appendix F, which highlights the descriptive statistics of all included features.

Table 4.1: Question Categories and Input Options for Nominal Data

Question Category	**Input Options**
Diagnosis Location*	Diagnosis-Location-1, Diagnosis-Location-2, Diagnosis-Location-3, Diagnosis-Location-4, Diagnosis-Location-5, Diagnosis-Location-6,
Exact Location*	Exact-Location-1, Exact-Location-2, Exact-Location-3, Exact-Location-4, Exact-Location-5, Exact-Location-6, Exact-Location-7, Exact-Location-8, Exact-Location-9, Exact-Location-10, Exact-Location-11,
Symptoms*	Symptom-1, Symptom-2, Symptom-3, Symptom-4, Symptom-5, Symptom-6, Symptom-7, Symptom-8, Symptom-9, Symptom-10,
Inflammatory Symptoms*	Inflammatory-1, Inflammatory-2, Inflammatory-3, Inflammatory-4, Inflammatory-5,
Caused by trauma?	Yes / No
When does pain occur?*	Start complaints after rest, Rest, During ADL (Activities of Daily Living) activities, After ADL activities, During warm-up, During sports, After sports, During work/school, After work/school, End of the day, At night
Are the complaints recurring?	Yes / No
Side?	Left, Left>Right, Both, Right>Left, Right
Are the complaints equipment specific?	Yes, in for every type of equipment; Yes, in some types of equipment; No, only without equipment; No, with or without equipment

*Multiple options can be selected for this datatype

The second type of data is the ordinal input type. This data type is represented as radio buttons

or tiles on the interface. For this data type, only one option can be selected by the user. This constraint is needed because the values within each question are typically mutually exclusive or hold a hierarchical relationship. This makes selecting one option conflict with wanting to select a second option. Table 4.2 presents a structured overview of questions and corresponding input options.

Table 4.2: Questions and Input Data for Ordinal Data

Question	**Input Options**
Existence period symptoms?	<1 week, <1 month, <2 months, 2-6 months, >6 months, >1 year
The course of the complaints?	Complaints have decreased, Complaints are indifferent, Complaints are mixed, Complaints have increased
Pain during walking?	None, Light, Medium, Severe, Not able to do this activity, Unknown
Pain during daily activity?	None, Light, Medium, Severe, Not able to do this activity, Unknown
Pain during intensive activity?	None, Light, Medium, Severe, Not able to do this activity, Unknown

*Only one option can be selected for this datatype

The final data type is the interval data. This data type only occurs once for the Numerical Rating Scale (NRS). The NRS is a widely utilized tool for assessing pain intensity, providing a quantitative measure of the severity experienced by patients. Respondents are presented with a range of integers from 0 to 10, where 0 signifies no pain, and 10 represents the worst imaginable pain. This comprehensive scale enables patients to articulate their pain intensity with granularity, facilitating accurate pain assessment and monitoring over time. However, it could be argued that this data is not truly interval data because the differences between scale points may not represent equal intervals of pain intensity as perceived by patients.

It's essential to recognize a significant detail regarding all input types: they allow for multiple entries per pain location. As depicted in Table 4.1, the report delineates six distinct symptom locations. For instance, if a practitioner notes that a patient experiences symptoms in two different pain locations, each question will be posed twice, once for each symptom location. This would lead to twice as many variables for each inputted report. Further elaboration on this scenario will be provided in Section 4.4.3.

4.3.2. TREATMENT REPORT DATA

As mentioned, the Treatment Report contains the final diagnoses given to the patients. This report also covers the future steps that need to be taken to treat the patient, however, this data is less relevant for this research. Currently, there are two types of variants of the treatment

report included in this study. The first type is an older version which is built on the in-house IT system. Following, there is a new type of report that was created using the EHR structure. This report was created months after the anamnesis report and is currently used interchangeably with the other treatment report. Because of the large amount of data already captured using the older version of the treatment report, and the fact that it is still in use, both versions were included in this research.

There are two key differences between the treatment reports that should be noted. Firstly, the new report version includes more questions that could be asked by practitioners. These additional questions, such as those delving deeper into a patient's medical history, were excluded from the research. This exclusion was due to the data being relatively new and not widely filled out, as many questions were not relevant to a large group of patients. Following, there is a second crucial difference between the reports that should be stated. In the older treatment report, a practitioner could only input one diagnosis for a patient. For the new CDSS treatment report, this changed into the option for multiple diagnoses. Currently, 10% of the reports in this research contain more than 1 diagnosis, whereas <1% even contain 3 or more diagnoses. It should be noted, however, that the percentage of multiple diagnoses per report is likely to increase as the new CDSS report is used more extensively. Therefore, this has been taken into consideration when creating the models in Section 4.5. One step of the research involves identifying the best performing machine learning model. As the dataset gradually shifts from single-label to multi-label, the best-performing model may also change. Thus, it is important to highlight this potential bias and critically assess the evolving data when selecting the most appropriate model.

Finally, the diagnoses in the database are categorized into 15 distinct groups each representing a specific aspect of the human body. These categories, which can be seen in Table 4.3, have been created by medical experts. It must be stated that subjectivity exists in the classification of these diseases into categories. Conditions classified under one category may also be deemed relevant to the other categories due to their anatomical proximity. Despite potential differences in medical expert opinions, a final decision was made for each diagnosis. Following, the database encompasses 177 unique diagnoses, which can be filled into these treatment reports. To streamline the analysis and focus on a more cohesive subset of conditions, this research only focuses on a subset of the locations and includes corresponding diagnoses, which can be seen in Table 4.3. This table summarizes the diseases by category, indicating the total number of diseases within each category, the number of diseases included in our research dataset per category, and whether the disease category was considered in this study. For this research, only a subset of the classes were selected relevant to a specific part of the body, which led to a total of 20 classes. The decision for only a specific part of the body was made because of the amount of data available for these categories. Furthermore, by concentrating solely on one specific subset of diagnoses, the prediction process becomes less complex. This is because these disease categories frequently exhibit similarities, allowing for more effective comparison and contrast.

Table 4.3: Summary of Diseases By Category

Disease Category	Amount of Diseases	Amount Currently Used	Included?*
Category 1	23	6	Yes
Category 2	22	6	Yes
Category 3	14	3	Yes
Category 4	8	1	Yes
Category 5	6	1	Yes
Category 6	12	3	Yes
Category 7	10	0	No
Category 8	27	0	No
Category 9	5	0	No
Category 10	13	0	No
Category 11	8	0	No
Category 12	7	0	No
Category 13	1	0	No
Category 14	18	0	No
Category 15	3	0	No
Total	177	20	-

*Only a subset of 20 classes has been selected for research purposes.

4.4. DATA PREPARATION

After gathering insights from the data, the next step is to get it ready for use in ML models. Initially, the data was not structured in a format that was suited for analysis. Therefore, this chapter focuses on three main things. First, Section 4.4.1 talks about changing the data format to simplify the ability to handle and modify the data. Secondly, Section 4.4.2 discusses data selection, detailing the types of data incorporated into the models and the exclusion criteria applied. Finally, Section 4.4.3 covers the feature engineering, encompassing the refinement of the chosen data to optimize its suitability for ML models, thereby transforming its format for seamless integration into the modelling process. Figure 4.3 gives a visualized view of the data preparation process.

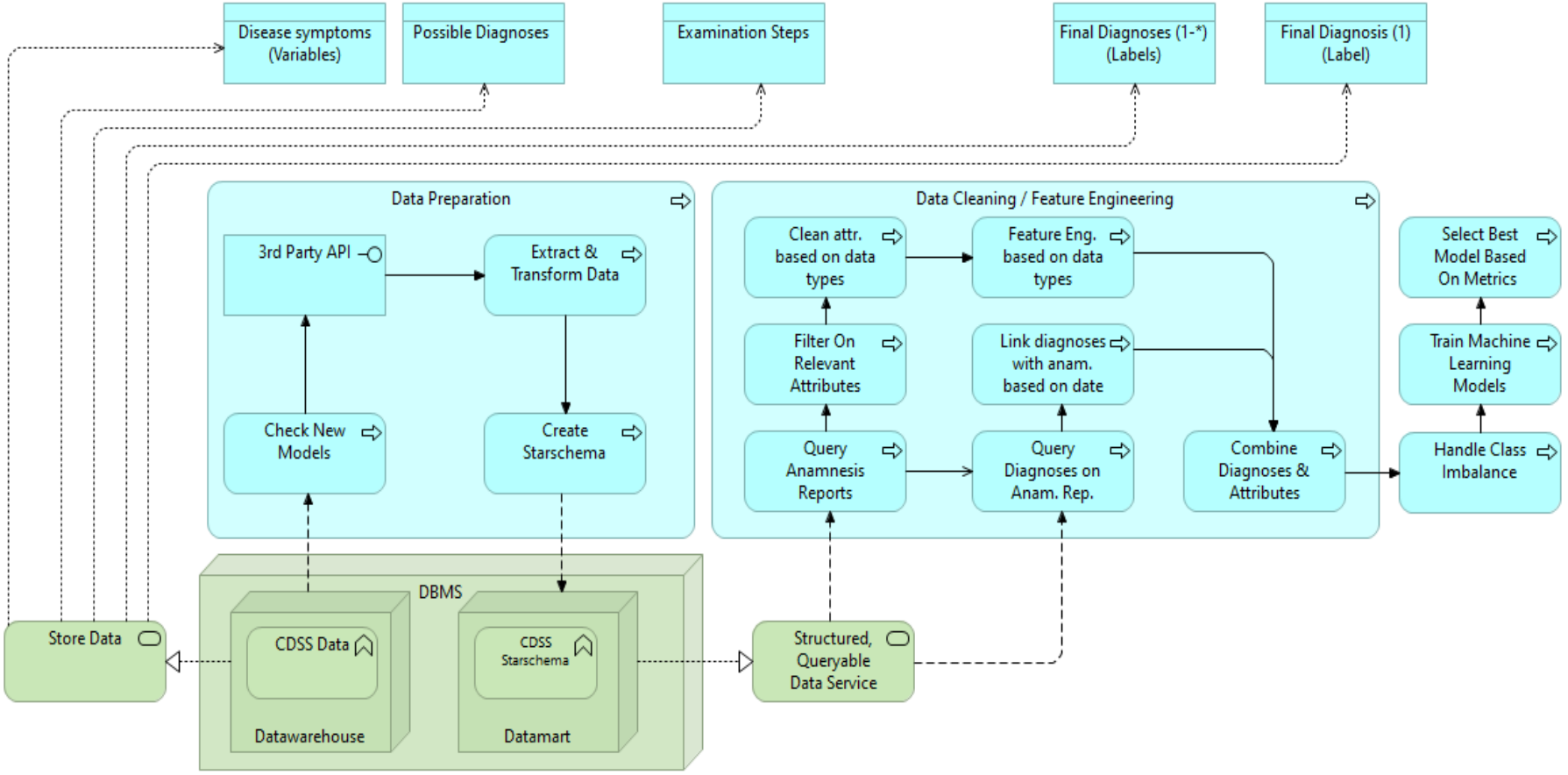


Figure 4.3: Technology and Application View for Data Process

4.4.1. DATA MODELLING

The initial step in data preparation involved the modelling of the dataset to render it conducive for analysis. Initially, the data provided consisted solely of questions accompanied by answers. Notably, the variables in the input values column of Table 4.1 lacked any linkage to their respective question categories. Consequently, disparate types of data, ranging from diagnoses and symptoms to medical history information, were intermingled. The current form of data storage lacks additional content to the data. The dataset can only be queried using specific questions outlined in the report, but the categories for these questions are not included. For instance, there are five different questions related to the inflammatory symptoms of a patient, but the category "inflammatory symptoms" is missing. Consequently, these five questions must be extracted individually, which hinders future-proofing. If the wording of a question changes, it would require updates in the data queries as well. Figure 4.4 shows an example of the old data format, in which headers, diseases, and different question types are all through each other in the database. To make the final output of this research scalable and future-proof, this data had first to be modelled in a way that was suitable for analysis.

Name	Value_Float	Value_String	Value_Enum
Diagnoses Category 1	[NULL]	0.0	[NULL]
What is your length?	182.3	[NULL]	[NULL]
Disease 2-1?	[NULL]	[NULL]	Yes
Symptom 1?	[NULL]	[NULL]	No
Symptom 8?	[NULL]	[NULL]	Yes
Amount of...?	8	[NULL]	[NULL]
When do you experi...?	[NULL]	Yes, During..	[NULL]
Disease 3-1?	[NULL]	[NULL]	No

Figure 4.4: Example of Data Storage Before Modelling

To enhance the utility of the data, a schema was devised to establish relationships between its various components. This star schema configuration facilitates richer and more efficient querying of the data. The provider of the report structures offers an API that furnishes information about the underlying structure in JSON format. Subsequently, a script was developed to interface with this API, triggered upon detection of new reports or report versions. Upon invocation, this script retrieves the pertinent data via the API, enabling the establishment of interconnections between different data elements. Consequently, the association between questions, answers, and their respective categories becomes significantly more discernible. The full process of capturing this underlying data format into a star schema can be seen in Appendix D.

The star schema, depicted in Figure D.1, comprises a fact table and four dimensions. The fact table encompasses the input values that users can populate, segregated by their respective input types such as enum, float, or string. Additionally, the fact table includes a unique identifier,

the report ID, which corresponds to the report being filled in. The primary dimension in our fact table is the CDSS structures dimension, which holds crucial details about the input values recorded. This dimension encompasses information regarding the source of the report and its corresponding version, alongside the linked question and its category. Categories could include diagnoses, symptoms, inflammatory symptoms, or medical history. Additionally, the fact table is linked to three other existing dimensions. The employee dimension specifies the employee who filled in the report, with their function and name being the relevant features. The date dimension solely captures the exact date variable. Lastly, the patient dimension is connected to the fact table but isn't currently utilized for data analysis.

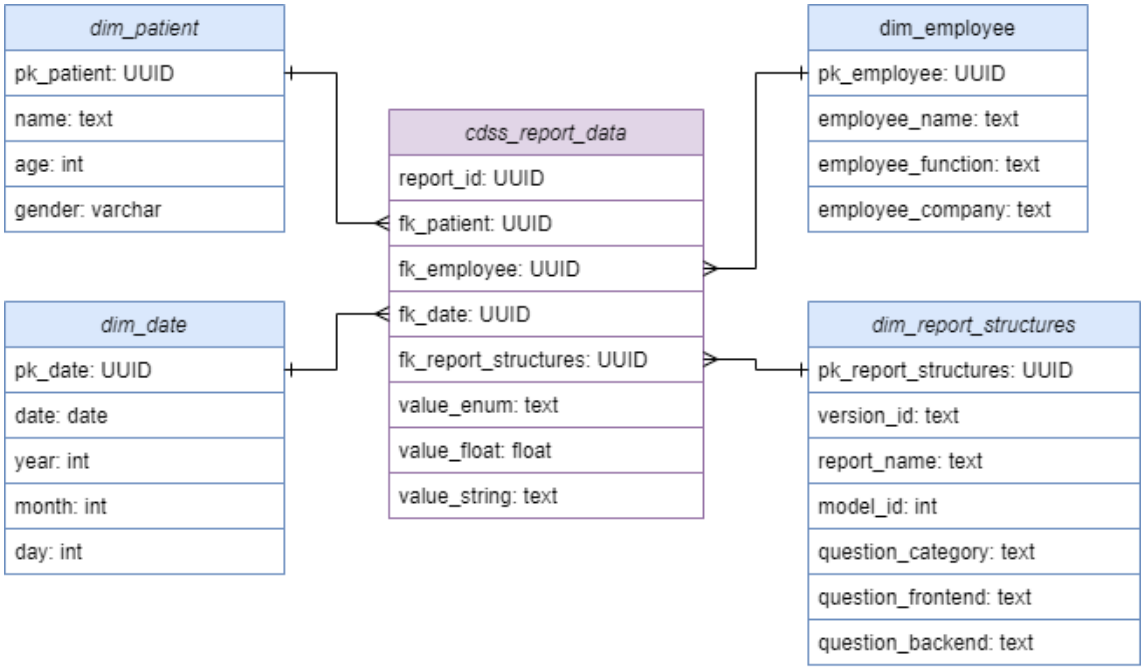


Figure 4.5: Starschema of Data After Modelling

The completion of data modelling has led to a more flexible approach to filtering question categories and made the system less vulnerable to changes in question formulations. By structuring the data in a star schema format with corresponding categories, the questions themselves can change, while still keeping most of the data queries working. An example of this data, in the form of the given star schema, can be found in Appendix C.

4.4.2. DATA SELECTION AND INTEGRATION

Upon data modelling, information can be selected, and features from reports can be excluded that are least relevant to the diseases or rarely filled in. The following section will cover the features and labels that were selected, and sequentially the data that was filtered out. After, this section will cover how the features of the anamnesis report were connected to the diagnoses of the treatment reports.

- **Relevant Reports:** During the data selection process, careful consideration was given

to select reports pertinent to the scope of the study. Test patients and their associated reports were excluded from the dataset. Furthermore, patients without any symptoms filled in were also excluded from the dataset. Finally, sometimes practitioners won't fill in any symptoms because of recurring patients or patients without any symptoms. These reports were also filtered out. This approach aimed to ensure the inclusion of cases about a new patient coming in without a diagnosis.

- **Final Diagnoses:** In some cases, a practitioner may only complete an anamnesis report before proceeding with treatment for a patient. Consequently, if a treatment report is not filled out for a patient, it indicates that a final diagnosis has not been recorded in the system. Since a label is essential for a supervised ML model to make predictions, patients without a final diagnosis are excluded from the dataset.
- **Specific Diagnosis:** The dataset was filtered on only diseases and corresponding symptoms related to a subset of the dataset. As mentioned in Section 4.3.2, everything of only a certain part of the human body was included. This decision aimed to enhance prediction accuracy and promote coherent data analysis by focusing on diagnosis groups with similar characteristics.
- **Symptom Features:** The features filled in during the anamnesis are filtered to only include direct symptoms of the patient to possible diseases. As mentioned in Section 4.3.1, the report also covers factors concerning medical history, such as visits to medical professionals or medical operations. However, these elements are deliberately excluded to focus solely on data possibly correlated to potential diagnoses for the patient.
- **Robust Representation:** As seen in Section 4.3, only a subset of the diseases were selected for this research. The selection of the total amount of diseases was essential to guarantee a foundational level of data for each diagnosis. This threshold not only provides a basic representation of each condition but also helps mitigate potential biases or anomalies that could arise from insufficient data.

After the selection of features and labels, the next step is to combine them into a cohesive dataset. This is achieved by linking entries based on patient identifiers and report creation dates. Patient numbers are assigned by the system to ensure that the right reports are linked to each other. Furthermore, since the reports were not linked to a certain appointment, they were matched based on creation dates within a 30-day window. This accommodated potential delays in report submission, which is especially common among novice practitioners, ensuring effective linkage despite temporal variations.

4.4.3. FEATURE ENGINEERING

Once the data was prepared and the relevant variables were selected, the features underwent refinement and modification for the ML models. The first type of feature engineering was a crucial step in the feature engineering process. Data points specific to a location were combined

to reduce the number of variables. As discussed in Section 4.3.1, when a practitioner records pain locations, they may input separate data for each instance, often duplicating values. Hence, these variables were grouped to significantly decrease the variable count. Following, Features that were similar to each other were also combined into one regularized feature. For example, left and right-specific features were combined into one overlapping feature.

The steps mentioned above reduced the feature space. As illustrated in Appendix E, the original 1153 features were already reduced by nearly 75% by selecting only features that have the highest correlation to diseases and were filled in regularly. However, with 299 features remaining, further reduction was imperative for the classification task. Feature engineering ultimately reduced the feature count to 54, a reduction of 95%. This consolidation ensured that the retained features provided richer information for model interpretability and improved model performance by prioritizing the most informative variables.

With the feature space reduced, the data types were altered to fit into the ML models. The nominal values which had True and False values were set to be one and zero respectively. Additionally, nominal data types with more than two options were converted to a numeric format, assigning a unique number to each option. Furthermore, ordinal values were standardized to a numeric format, ranging from 1 (lowest) to the highest value corresponding to the top category. For example, the degree of pain experienced while performing intensive activities, previously described from mild discomfort to complete inability to walk, was encoded from 1 to 4, enabling better predictive performance with fewer variables. These steps meant that all of the variables were now easily implementable into ML models for training.

Finally, once all of the features were improved, the missing values were handled. For nominal data types, missing values were set to 0, as their absence could imply either oversight by the practitioner or the absence of the symptom in the patient. For example, when one of the inflammatory symptoms is not filled in, it could imply that the practitioner forgot or that this inflammatory symptom did not occur. Because this distinction could not be made, these features are not handled. However, the missing numerical and categorical values could be identified, because these can only be zero when they are not filled in. For the numerical values, it was decided to change the missing values to the overall average, which was around 120 records in the dataset. Following, the categorical data, the median values were selected. This choice was made because these values were set to an incremental value, and the average value for these data types was sometimes the least filled in, which could lead to creating unnecessary noise for crucial information. Therefore, the missing values were set to the Median values for this datatype. The average amount of missing data for these features combined was around 16%, the exact amount of missing values can be seen in Table 4.4. Concluding this section, with the feature engineering complete, the models were equipped with well-suited data, and the reduction in feature noise supports model performance and interoperability.

Table 4.4: Table Which Shows How Missing Data is Handled

Feature	Amount Missing	Percentage Missing	Missing value replaced in
NRS	120	11.5%	Average, which is 5
Course of complaints	161	15.5%	Median, which is 3
Daily Activities	155	14.9%	Median, which is 2
Intensive Activities	308	29.6%	Median, which is 2
Walking	175	16.8%	Median, which is 2
Side of symptoms	24	2.3%	Median, which is 1

4.5. MODELING STRATEGIES FOR CLASS IMBALANCE

During the ML models implementation phase, we elucidate the implementation process and rationale behind various decisions made during this phase. Figure 4.6 illustrates a highly skewed distribution of diagnoses within the dataset. Notably, the most recorded diagnosis is 17 times more common than the least recorded diagnosis. This skew presents a potential challenge, as it may lead models to exhibit bias towards the predominant diagnoses. Consequently, a model could achieve seemingly high accuracy by simply predicting the most frequent diseases, which could yield misleading recommendations, particularly for cases outside this subset.

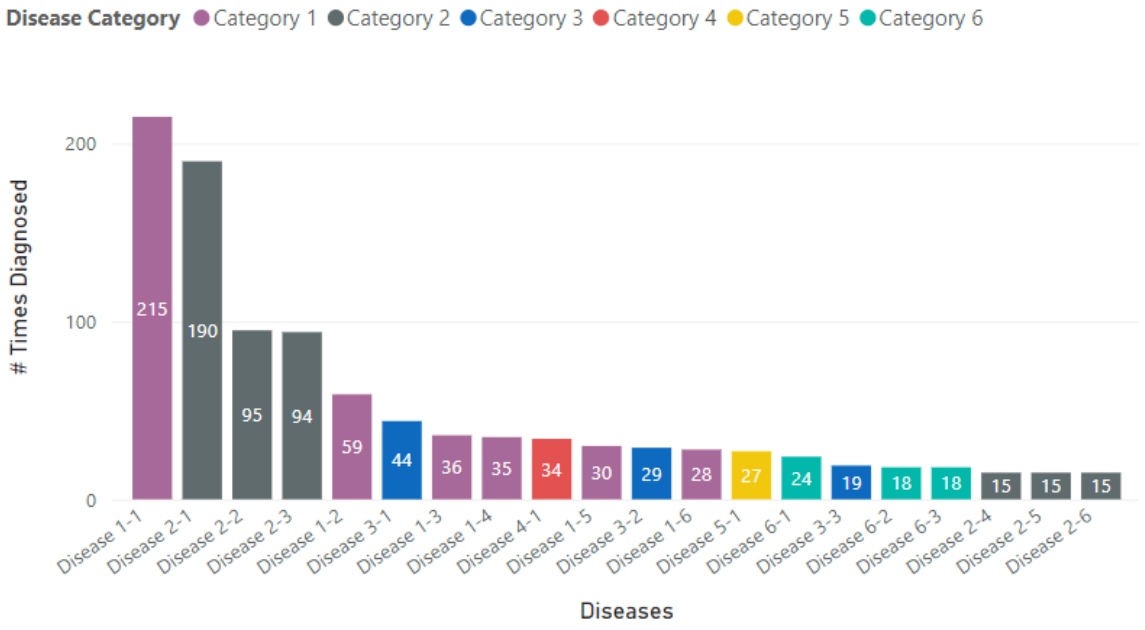


Figure 4.6: Distribution of Diagnoses Included in This Research

To address this imbalance, various class-balancing techniques can be employed. Among these techniques, undersampling involves removing data from the overrepresented classes, but given the limited dataset in this research, an oversampling approach was preferred. Specifically, random oversampling was chosen, wherein instances from the minority classes are duplicated.

While Synthetic Minority Oversampling Technique (SMOTE) is another option, it was deemed less suitable due to its potential impact on explainability; SMOTE generates artificial data, which could lead the model to make predictions based on synthetic instances not present in the original dataset, thus compromising interpretability.

Moreover, as discussed in Section 4.3.2, the nature of the problem is evolving from a SLC to a MLC scenario, owing to the possibility of multiple diagnoses being selected in the new treatment report. Hence, it becomes imperative to consider the appropriate classification approach. Below, we analyze three different options for addressing the classification problem, each tailored to suit the evolving problem type:

- **SLC with Class Imbalance:** Serves as a baseline, demonstrating model performance without any manipulation of class distribution.
- **SLC with Random Oversampling:** Randomly selects data from the minority class and duplicates them, offering fairer representation to underrepresented classes. However, it may inflate the importance of certain variables due to duplication.
- **MLC with Weight Factors:** Adapts the Category Imbalance and Cost-Sensitive Thresholding (CICST) model proposed by Liu et al. [48], assigning weight factors to each disease and triggering predictions when they exceed a certain threshold.

The training will involve exploring various strategies to address both SLC and MLC problems. In the subsequent sections, we will delve into a more detailed explanation of how we plan to tackle class imbalance and multi-label classification challenges using different techniques.

4.5.1. SINGLE-LABEL CLASSIFICATION

For single-label classification, the H2O platform was utilized due to its robustness and scalability [2]. The implementation involved training multiple models with varying strategies for handling class imbalance. Notably, two approaches were adopted:

- **Equal Representation:** Each label was duplicated to match the frequency of the most common class, ensuring balanced representation across all diagnoses. This most common class had a sample size of 198, which meant that the other diagnoses were duplicated until they reached the same size.
- **Shared Representation:** Alternatively, labels were associated with the same row, allowing models to learn from shared features across multiple diagnoses. These strategies aimed to assess the impact of class balancing techniques on model performance and interoperability.

[2] https://h2o.ai/

4.5.2. MULTI-LABEL CLASSIFICATION

Drawing from the CICST model proposed by Liu et al. (2021) [48], the multi-label classification framework was designed. This approach emphasizes the automatic classification of multiple helathcare conditions while addressing class imbalance and cost-sensitive thresholding.

The model structure and training process were adapted to suit the healthcare diagnostic context, with a focus on:

- **Automatic Multi-Label Classification:** Predicting multiple diagnoses simultaneously to cater to the complex nature of healthcare conditions.
- **Category Imbalance Handling:** Employing weight factors to account for class imbalances and ensure fair representation of all diagnoses.
- **Cost-Sensitive Thresholding:** Setting threshold values based on the severity and prevalence of each condition, thereby enabling prioritized predictions. The implementation closely followed the structure outlined in the literature [48], with necessary adjustments to align with the specifics of the healthcare domain.

4.5.3. EVALUATION METRICS

The three models in question will be evaluated using a variety of metrics. Given that two of the models are of the SLC type and one is a MLC type, we will assess them based on their precision and sensitivity scores for each disease. These scores will be contextualized within potential use cases to identify the most suitable model for each scenario. The evaluation will include:

- Sensitivity for Each Diagnosis
- Precision for Each Diagnosis
- F1-Score for Each Diagnosis

Additionally, the overall sensitivity, specificity, and F1-score will be examined. By comparing these metrics, insights into which models perform best for different types of use cases can be obtained. The values for these

4.6. EXPLAINABILITY

The final part of the research covers the explainability output of the ML model. This step is crucial in ensuring the model's outputs are interpretable and actionable. The explainability process is executed in three stages, illustrated in Figure 4.7. First, four different possible local explainability models were created, based on the research of Naiseh et al. [24]. These models are detailed in Section 5.3. These four models gave more tangible options for practitioners to decide what they would prefer.

Next, a questionnaire was conducted involving thirteen practitioners to account for variations in explainability needs across different sectors, companies, and individuals. Inspired by Naiseh

et al.'s work, the questionnaire aimed to capture practitioners' perceptions and preferences on three topics for each model: Understandability, Information Sufficiency, and Reliability Evaluation. The statements evaluated were:

- I understand the information that the visual is showing me.
- I receive enough information to determine which disease to examine.
- I can determine (in)correct diagnoses from the output.

Additionally, the questionnaire included a ranking of the models before and after an explanation of the information displayed. This ranking assessed the intuitiveness of the explainability models, as training can be costly for organizations. The goal was to ensure that the explainability felt intuitive to practitioners from the first use. The questionnaire concluded with a query on whether parts of the post-hoc explainability models should be combined. This question aimed to gather detailed insights, allowing the integration of highly rated elements from different models into a final, optimized model.

The results from the questionnaire informed the final step: a focus group discussion with key stakeholders and experienced practitioners. The focus group included the director of healthcare at Company X, the coordinator of integral care, and a member of the data-driven care team. This group reviewed the questionnaire results, selected the best-scored elements from each model, and refined features that received mixed feedback. This three-step process is done to make the final model well-suited to its specific implementation context because it is important to create the final model with the end-user in mind.

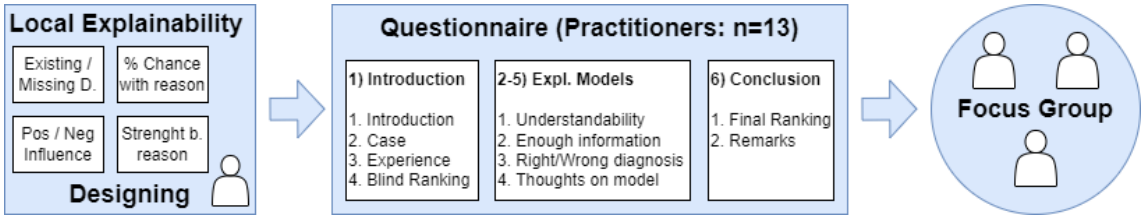


Figure 4.7: Explainability Research Process

5

RESULTS

5.1. EXPLORATORY DATA ANALYSIS

This section presents the global feature importance for the three models discussed in Section 4.5. We will first highlight the global feature importance, identifying the most critical variables for each model. Following this, we will examine partial dependency, providing insights into the most influential features of each specific diagnosis.

5.1.1. GLOBAL FEATURE IMPORTANCE

The global feature importances were calculated using the built-in global feature importance functions of H2O. Unlike local feature importance, which evaluates feature importance for individual predictions, global feature importance aggregates the importance of each feature over the entire dataset. Using H2O the global feature importance for the GBM and RF models is calculated based on the splits at each node within the decision trees. For every split, the reduction in variance is determined. The reduction in variance is the change in variance from the parent node to the corresponding child nodes, where a larger reduction indicates greater importance. The variance is calculated using the following formula [1]:

$$\text{VAR} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{5.1}$$

Where VAR represents the variance on a single node. In this multiclass problem, the y values are treated as binary using a One-vs-Rest approach. Specifically, the target value y_i for observation i is 1 if the observation belongs to the class of interest, and 0 otherwise. The mean ($\bar{y}$) is the average of these binary values across all observations in a node.

The variance reduction values are summed for every decision tree in the RF or for the single

[1] https://docs.h2o.ai/h2o/latest-stable/h2o-docs/variable-importance.html

decision tree in the GBM to obtain a final importance score for each feature. To normalize these scores and facilitate comparison, the final scores for each feature are summed, and the importance score of each feature is divided by the total sum of all scores. This provides a relative importance measure for each feature.

This method allows for the comparison of feature importance across the entire model, providing insights into which features have the greatest impact on the predictions. By revealing which features significantly influence the final predictions of the ML models, this approach offers an understanding of the model's behaviour. Features with a score of 0 are not utilized by the model, indicating that they have never determined a split in any of the trees. Conversely, features with high importance scores either determine multiple splits or significantly reduce variance when they determine a split. Additionally, the variables are categorized into seven groups based on their similarities. This categorization helps identify the most crucial groups of variables for the model and facilitates an understanding of the relative importance of variables within each group.

In the first model, shown in Figure 5.1, two key feature groups emerge: diagnosis location and the specific location of symptoms. These two feature groups respectively have the highest two feature importance scores, which means the baseline ML model focuses most on these feature sets. This distinction highlights the model's ability to differentiate diagnoses based on where the patient experiences symptoms most prominently. Furthermore, the symptoms and inflammatory symptoms surprisingly had a low feature importance score. These features normally are very important for practitioners to identify the right diagnosis of a patient but are not used quite as much by the model based on the baseline dataset. In machine learning models, features that strongly correlate with the most frequently occurring classes in the dataset tend to gain higher importance in the model's predictions. This is because the model learns that predicting these prevalent classes increases the likelihood of making correct predictions.

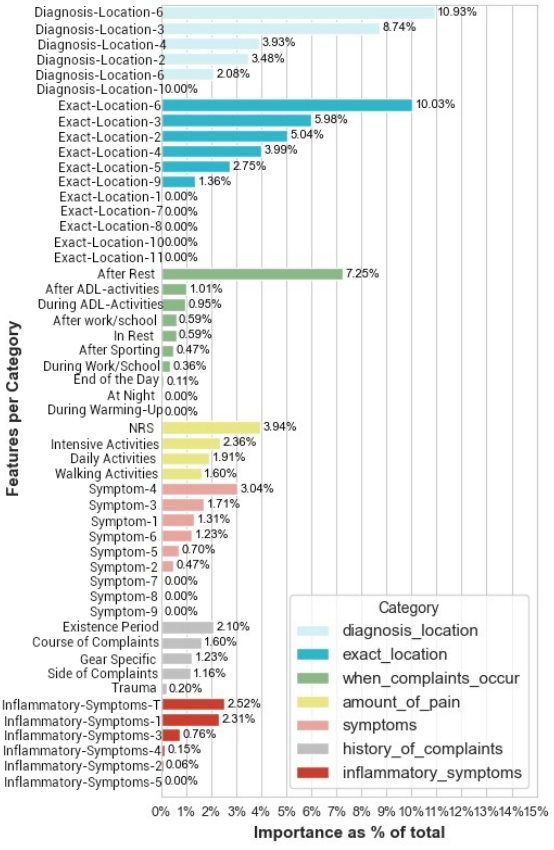


Figure 5.1: Feature Importance for Baseline Data

Consequently, features associated with these common classes become more influential, reflecting the model's bias towards predicting the most prominent categories within the dataset.

This bias towards frequent classes results in higher feature importance scores for characteristics closely linked to these categories. This is mainly seen by the fact that the two diagnosis-locations and exact-locations with the highest feature importance are also correlated to the two most prominent labels.

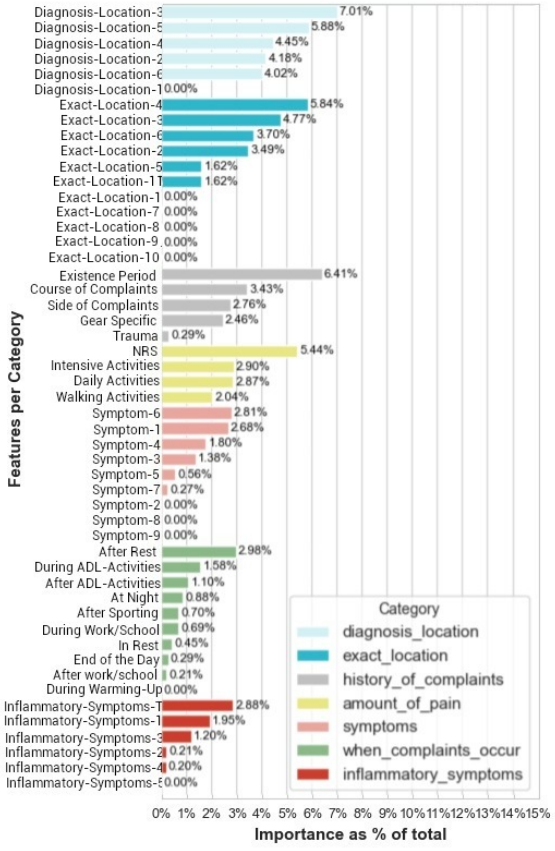


Figure 5.2: Feature Importance for Oversampled Data

In contrast to the previous model, the second model depicted in Figure 5.2, which utilizes oversampled data, exhibits a more balanced distribution of feature importance. Unlike before, where certain features stood out significantly, here we observe a more evenly spread of importance among features. While the general ranking of feature categories remains consistent, no conspicuous spikes indicate disproportionately influential features.

This shift can be attributed to the model's altered perspective on the data. By treating all classes as equal through oversampling, the model no longer favours features associated with frequently occurring diagnoses. Instead, it assigns greater importance to features distinguishing less common diagnoses. For instance, the feature 'After Rest' experiences a notable decrease in importance compared to the previous model, suggesting a recalibration of relevance.

Furthermore, there's a noticeable emphasis on the history of complaints in this model. With each diagnosis receiving equal consideration, features capturing the progression or duration of symptoms become more significant. It's plausible that these features are better suited to discerning accurate diagnoses when all conditions are weighted equally.

In the previous models, which were based on SLC, there was a noticeable emphasis on certain features. This was likely due to their prevalence or higher variance in the data. Although it was less prevalent for the oversampled data, it still existed. However, the latest model, which employs a MLC, presents a more balanced distribution of feature importance. Interestingly, the 'amount of pain' category has emerged as the highest scoring category in this model, a shift from the earlier models where the locations of the symptoms were deemed significant. Additionally, the patients' abilities to perform specific tasks have also gained prominence in influencing the model. This finding is somewhat unexpected. While the 'amount of pain' is indeed an important feature, it is surprising to see it considered as the most critical piece of information for decision-making by practitioners. This shift in feature importance underscores the nuanced capabilities of the MLC in balancing the influence of different features.

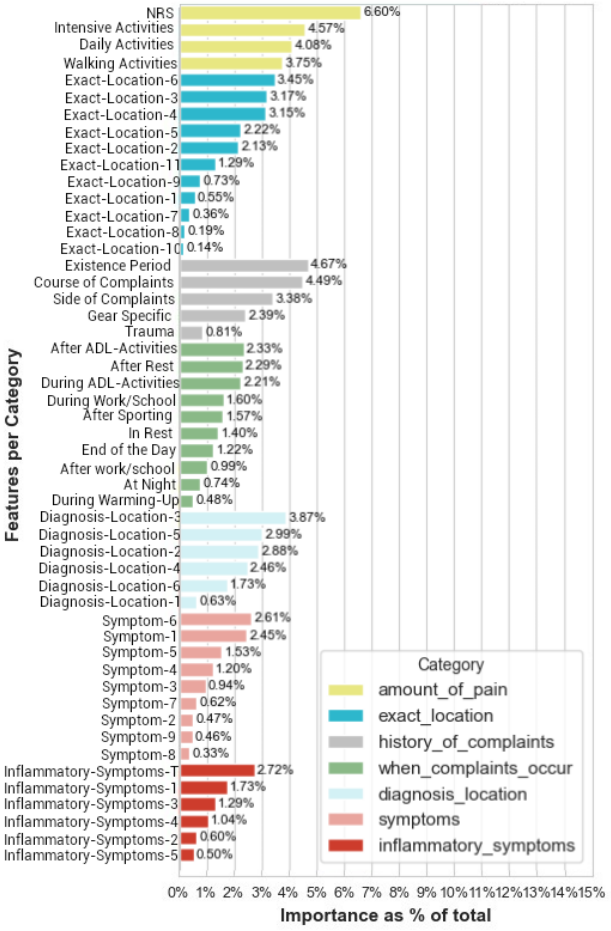


Figure 5.3: Feature Importance for CICST Data

In summary, the shift towards a more balanced distribution of feature importance in the oversampled model indicates a recalibration of the model's focus. By considering all diagnoses equally, the model highlights features that effectively differentiate between less common conditions, potentially enhancing its diagnostic accuracy across the board. Following, the introduction of an MLC seems to balance the feature importance further. All of the features in this final model seem to be around the same level of importance compared to each other. Another interesting condition to note is that the symptoms and inflammatory symptoms categories scored very low for every model. It would be expected that these would be higher scoring because practitioners consider this a crucial factor in diagnosing a patient. A closer look into the data showcases that most of the patients did not have inflammatory symptoms. Furthermore, two of the symptoms were filled in for most diagnoses, while others were far less common. It could be that these two feature sets were similar for every disease, and therefore not a good distinguishing factor for the models to base a decision.

Finally, identifying the most critical features mentioned above can assist with feature selec-

tion. For instance, the feature 'During Warming-Up' shows minimal importance in the current models. However, determining whether certain features should be combined, altered, or removed remains challenging because the models are trained on only a subset of the features. Features that appear insignificant in the current models might still hold importance for different diagnoses that have not yet been included. The combined global feature importance from all models is illustrated in Figure 5.4. This figure highlights that some features have little to no influence on the final predictions.

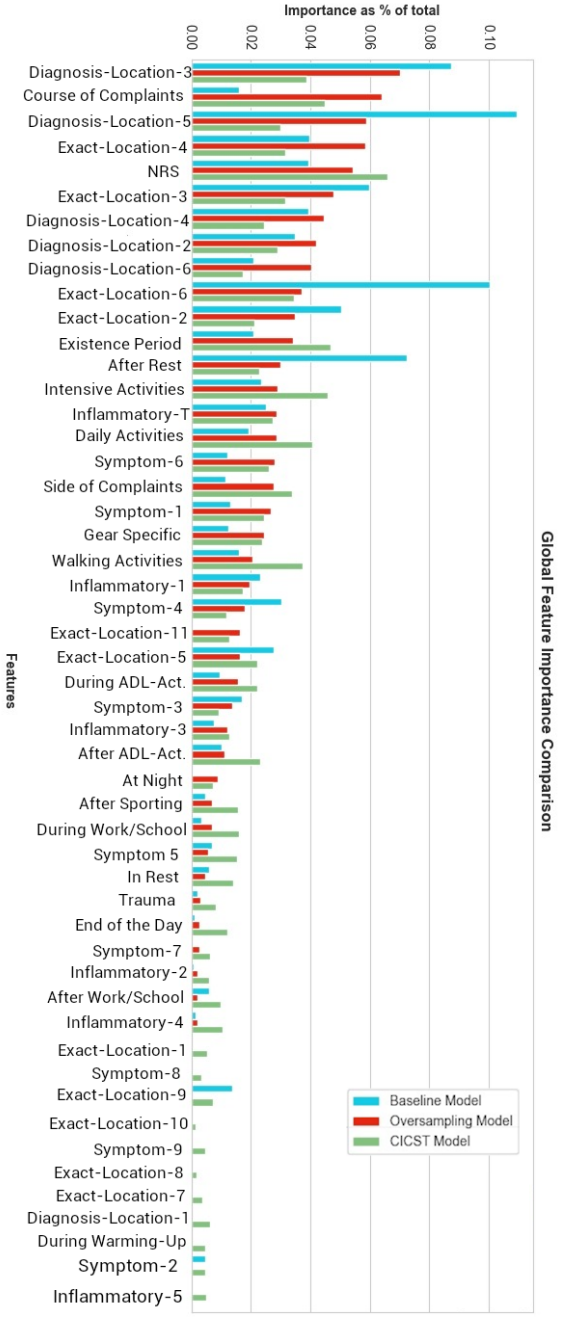


Figure 5.4: Comparison Feature Importance Between Models

5.1.2. PARTIAL DEPENDENCY

Partial dependency plots are traditionally effective tools for visualizing the influence of specific features on the final prediction. They are particularly valuable in multi-class problems for discerning feature-class correlations. However, given the extensive number of features and classes in this study, generating all necessary dependency plots is impractical. With three models, fifty-one features, and twenty labels, creating 3060 plots would be unfeasible. As a pragmatic alternative, the analysis focuses on examining the disparity in mean response values when certain features are either True or False. Given that most features are nominal values, this approach offers valuable insights into the impact of feature states on predictions. The resulting differences in response values indicate the strength and direction of correlation between a feature and a diagnosis, or lack thereof. These insights are encapsulated in heatmaps presented in Figures 5.5, 5.6, and 5.7, utilizing testing data to underscore the models' performance with new data. For instance, these figures reveal a strong positive correlation between the diseases and the True values for some of the features. This would mean that these features strongly correlate to these diseases.

The heatmaps underscore the varying degrees of influence that features exert on the models. Notably, the CICST model appears significantly more responsive to feature states compared to the other two models. Examination of the legend in Figure 5.7 reveals a wider axis range of 0.6 to -0.25, whereas the models in Figures 5.5 and 5.6 span from 0.2 to 0.2. This discrepancy suggests the potential superiority of the CICST model in disease prediction and feature-disease correspondence discernment. However, elevated influential values may also signal overfitting. For instance, the feature Inflammatory-Symptom-4 elicits a notably high response for this diagnosis despite lacking a strong correlation. Hence, meticulous identification of such instances aids in guiding subsequent feature engineering or modeling iterations. Moreover, the disparities in these values offer insights into operational discrepancies among different ML models. Take the variable 'nagelklachten' as an example: the two GBM models exhibit negligible importance assigned to this specific variable, with mean response differences close to 0. Conversely, the RF model presents a distinct pattern, assigning notably high or low scores to 'nagelklachten' for certain classes. This discrepancy underscores how the SLC GBM models interpret distinct patterns compared to the MLC RF model, leading to disparate conclusions.

The ML models play a crucial role in identifying features or labels that exhibit negligible impact across all scenarios. For instance, the feature 'eind van de dag' consistently elicits no response in any model. This suggests a potential lack of importance for this variable, warranting consideration for removal from the dataset. However, it's essential to acknowledge that these features might still hold value in diagnosing conditions not currently captured by the ML model. Additionally, it underscores the reality that certain diagnoses lack distinct feature sets. Upon closer examination of diagnosis values, it becomes evident that some lack high or low mean response values for any feature. This implies the model's inability to effectively differentiate this diagnosis based on existing features. This can be seen when comparing the values from Tables 5.5, 5.6 and 5.7 to the performance per diagnosis in Table 5.1.

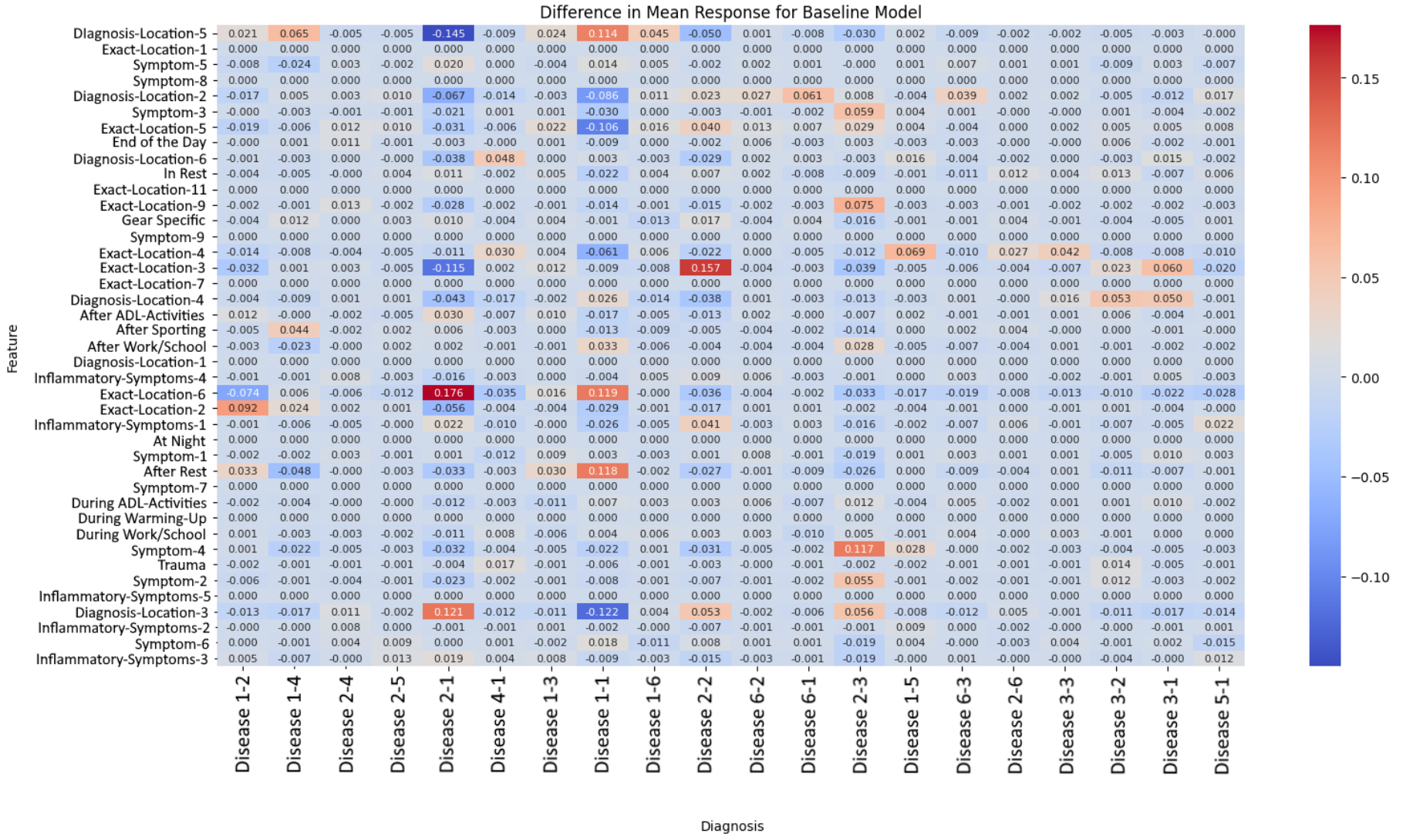


Figure 5.5: The Partial Dependency Values for Every Nominal Input Type for the Baseline Model Based on Test Data

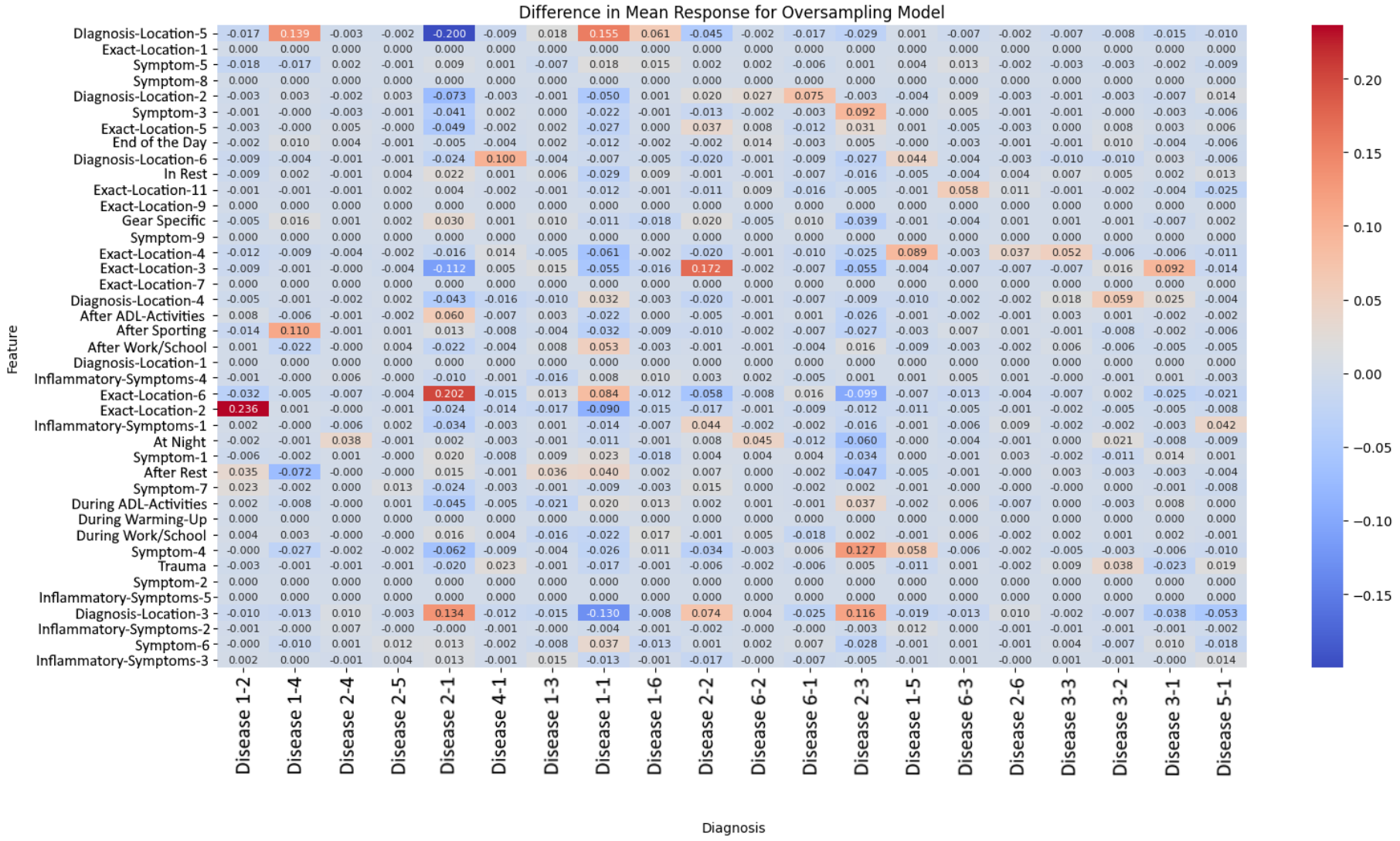


Figure 5.6: The Partial Dependency Values for Every Nominal Input Type for the Oversampling Model Based on Test Data

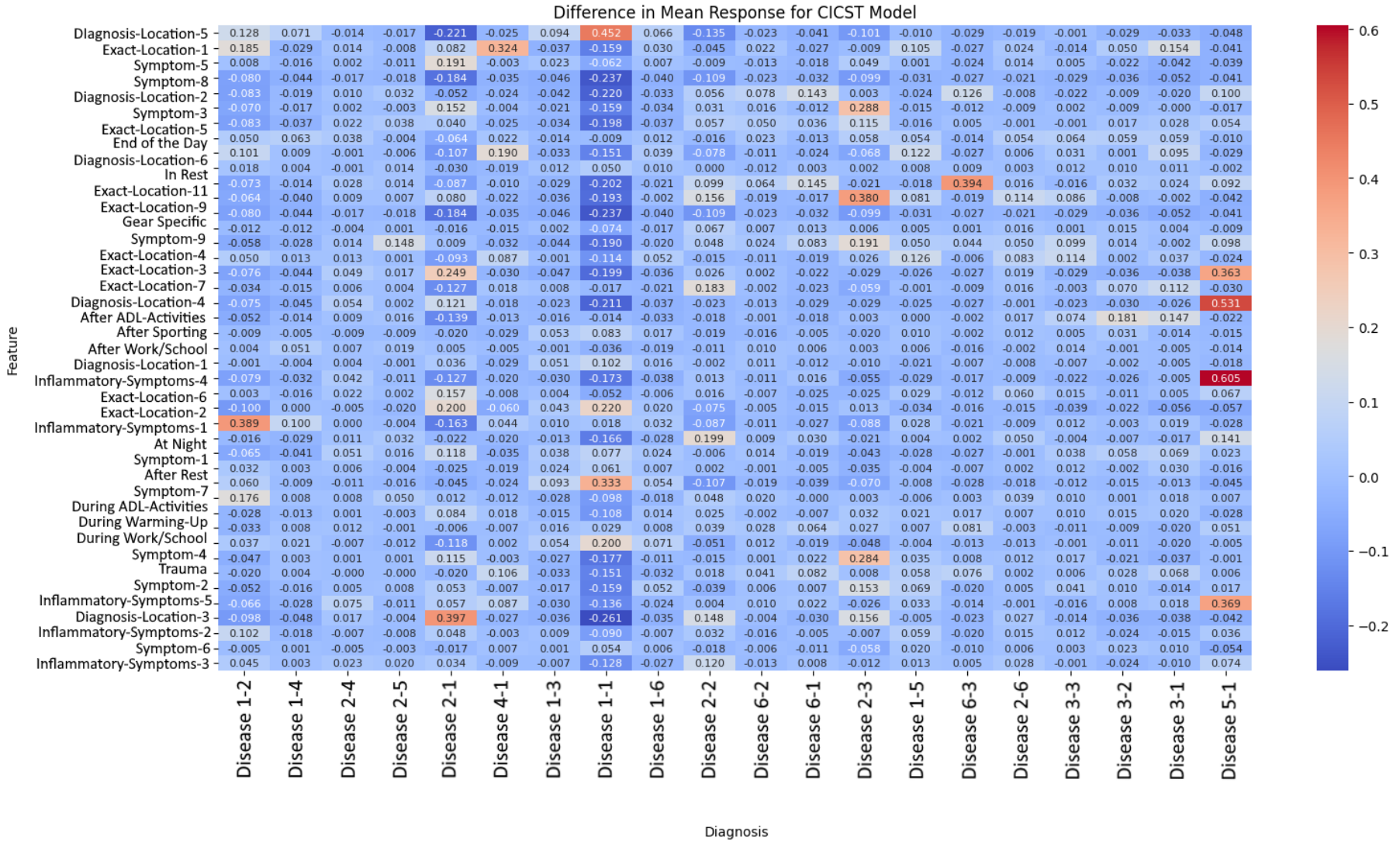


Figure 5.7: The Partial Dependency Values for Every Nominal Input Type for the CICST Model Based on Test Data

The diagnoses with the lowest performances like diagnoses 2-4 or 2-5 are also the diagnoses with the lowest total mean response. The limitation of effectively differentiating diagnoses could stem from biases inherent in the small sample size or the absence of defining features for this diagnosis. In conclusion, these findings indeed reflect the models' identification of important features for each diagnosis, emphasizing the need for more extensive sample sizes to discern the distinguishing characteristics of certain diagnoses.

5.2. MODEL COMPARISON BASED ON DIAGNOSIS-SPECIFIC PERFORMANCE

This section provides a comparative analysis of the performance of three trained ML models. Prior to discussing the actual performance of each model, the training conditions for developing the ML models will be explained.

The first two ML models were trained using H2O. H2O was configured to only train ensemble methods, in which H2O supports GBM, DRF, and GLM. Ensemble methods are chosen due to their robustness and ability to improve predictive performance by combining the strengths of multiple models, reducing the risk of overfitting and increasing generalization. In both instances, H2O was configured to train ten different models, selecting the one with the lowest mean per class error, a useful metric in multi-class classification problems as it provides an average measure of the classification error across all classes. Interestingly, in both the baseline and oversampling datasets, the model with the best performance was consistently a GBM. This preference for GBM may be attributed to its effectiveness in handling complex, non-linear relationships and its capacity to manage various data types and distributions [42]. GBM's iterative boosting approach also helps in sequentially correcting the errors of weaker models, leading to a more accurate and robust final model.

For the MLC, a RF model was selected. The choice of RF for the MLC model was motivated by several reasons. Firstly, RF models are known for their ability to handle high-dimensional data and provide robust performance even in the presence of noise. Secondly, RF models can capture complex interactions between features, making them particularly suitable for multi-label problems with a large number of features.

5.2.1. MODEL PERFORMANCE

After looking at the features that are important for each model, it is also crucial to know which models perform the best in terms of precision, sensitivity and F1 score. This chapter will first discuss each model separately, and then compare them afterwards.

Baseline: The initial model, trained on the original dataset with a skewed class distribution, demonstrates satisfactory performance predominantly for the most frequent classes. As indicated in Table 5.1, the model accurately predicts the top classes most of the time. For instance, Disease 1-1 achieves a sensitivity of 0.83, indicating high sensitivity for this class. However, the precision for this class is relatively low, at 0.56, suggesting a significant number of false positives. This outcome highlights the model's bias towards predicting the more common classes

due to their prevalence in the training data. The overall accuracy of the model stands at 54%, largely because the model tends to favour the prediction of frequent classes, thereby correctly predicting them more often. Nevertheless, the model's performance deteriorates substantially for less common classes. In fact, there are seven classes that the model fails to predict entirely. This failure occurs because these rare classes share feature similarities with the more common classes, but due to their lower representation in the dataset, the model is less likely to predict them.

Oversampling: The second model, which was trained on oversampled data, exhibits an overall accuracy of 49%. This slight decrease in accuracy compared to the baseline model can be attributed to the model's increased focus on less frequent classes, thereby lowering the average accuracy. While the performance of the most common classes has declined, there is a notable improvement in the prediction of less frequent classes. Specifically, the classes that performed poorly in the baseline model show improved scores in this model. Despite these gains, the overall precision and sensitivity of this model have decreased relative to the baseline. Interestingly, this model also fails to predict six classes entirely, which is unexpected since oversampling should theoretically balance the importance of all classes. This anomaly may be due to the nature of random oversampling, which can provide the model with information limited to the specific instances included in the training set. During dataset splitting, where 80% of the data is used for training and 20% for testing, minority classes are divided into small subsets. If the three samples in the testing set are not similar to the twelve samples in the training set, which are heavily duplicated, the model fails to predict these classes. Even though these 12 samples are cross-validated using H2O, they can still be different from the final 3 testing cases. While SMOTE could have mitigated this issue, as discussed in Section 4.5, it was not employed to preserve model explainability. Another significant improvement is observed in the top-k hit ratio. Although the accuracy remains at 49%, similar to the overall accuracy, the top-3 hit ratio has improved markedly. This model includes the correct diagnosis in its top 3 predictions 79% of the time, compared to 70% in the baseline model. This indicates that the model performs better when allowed to output multiple potential diagnoses.

CICST: The final model, an implementation of the RF model utilizing a one-vs-rest classifier for MLC, is employed. This model generates a one-vs-rest classifier for each diagnosis. If the expectancy of a diagnosis surpasses a predetermined threshold, it is predicted as True. The thresholds are determined through iterative exploration of various threshold values, selecting the one that maximizes the combined F1 and Sensitivity Scores. The final threshold values can be found in Appendix H. In terms of accuracy, this model outperforms previous iterations, achieving a significantly higher combined accuracy of 92%. However, this figure can be somewhat misleading, as an accuracy of 80% would still be attained even if the model predicted every diagnosis as false. Therefore, other evaluation metrics provide a more accurate assessment of the model's performance. On these alternative metrics, the RF model shows substantial improvements, achieving 1.5 times the sensitivity and F1-score, and double the precision compared to the other models. This model was optimized for higher precision, emphasizing

the importance of including the correct diagnosis as an option rather than ensuring it is the top prediction. Despite these improvements, two classes—Disease 2-4 and Disease 2-5—still receive no predictions, consistent with the results from the previous models. As discussed in Section 5.1.2, these conditions exhibit very low mean responses across their values. The model struggles to distinguish these diagnoses from others given the current dataset.

Overall, considering all metrics, the CICST model performs the best. This superior performance can be attributed to several factors. Firstly, to enable the models to function as SLC, some records had to be duplicated. For example, if a patient has two diseases, there will be one set of features for the first disease and a duplicate set for the second. This duplication can prevent the model from accurately predicting the correct diagnoses because the same features are repeated. Additionally, some diagnoses may have overlapping features for different diseases, making it challenging for SLC models to distinguish which features correspond to each disease. While this is still a challenge for MLC, it is easier for MLC models to handle. Furthermore, the CICST model was specifically trained to optimize the precision score, which is reflected in its superior performance. The first two models were trained using mean squared error as the objective, which limited their effectiveness. If these models had been trained with a focus on precision, their scores would likely be higher, as they would have been optimized with precision as the primary criterion.

5.2.2. PRACTICAL IMPLICATIONS

When considering the practical implications of implementing these models, the CICST model stands out as the best performer. Firstly, it's important to note that the CICST model excels in both precision and F1-score. This means it not only provides accurate diagnoses more frequently but also includes the correct diagnosis in its output more often. This is crucial for practitioners, who can then verify the diagnosis. The higher precision score is particularly valuable because it ensures that the correct diagnosis is likely to be among those suggested.

Additionally, as discussed in Section 4.3.2, the new treatment reports allow for multiple diagnoses to be recorded. With approximately 10% of records currently involving multiple diagnoses, this number is expected to grow. Therefore, the CICST model's ability to handle multiple diagnoses will become increasingly beneficial over time. This model is not only the best choice now but is also expected to outperform the other models in the future.

Lastly, the CICST model offers greater flexibility in practical application. If certain diagnoses become too frequent or too rare, the weight factors can be adjusted to improve performance. This adaptability empowers practitioners and data-driven care departments to fine-tune the model to better suit their needs, enhancing its practical utility. The final weight factors for this model are displayed in Appendix H.

Table 5.1: Comparison of Model Performances per Diagnoses

Class (n)	Baseline			Oversampling			Cicst		
	Sensitivity	**Precision**	**F1-Score**	**Sensitivity**	**Precision**	**F1-Score**	**Sensitivity**	**Precision**	**F1-Score**
Disease 2-1 (190)	0.78	0.59	0.67	0.70	0.59	**0.64**	0.54	0.82	0.65
Disease 1-1 (215)	**0.83**	0.56	0.67	0.67	**0.61**	**0.64**	0.66	0.84	0.74
Disease 1-2 (59)	0.78	**0.78**	**0.78**	0.67	0.55	0.60	0.56	0.77	0.65
Disease 1-6 (28)	0.00	0.00	0.00	0.14	0.14	0.14	0.31	0.63	0.42
Disease 1-4 (35)	0.75	0.75	0.75	0.50	0.44	0.47	0.67	0.29	0.40
Disease 2-2 (95)	0.62	0.65	0.63	0.43	0.60	0.50	0.85	0.68	0.76
Disease 5-1 (27)	0.75	0.33	0.46	**0.75**	0.43	0.55	0.75	**1.00**	**0.86**
Disease 3-3 (19)	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.50
Disease 4-1 (34)	0.57	0.40	0.47	0.71	0.50	0.59	0.70	0.78	0.74
Disease 3-2 (29)	0.29	0.33	0.31	0.14	0.17	0.15	0.25	0.75	0.38
Disease 1-5 (30)	0.17	0.25	0.20	0.33	0.50	0.40	0.39	0.70	0.50
Disease 2-3 (94)	0.39	0.54	0.45	0.39	0.33	0.36	0.32	**1.00**	0.49
Disease 3-1 (44)	0.25	0.38	0.30	0.17	0.33	0.22	0.63	0.56	0.59
Disease 2-6 (15)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.50	0.67
Disease 2-4 (15)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Disease 1-3 (36)	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.71	0.31
Disease 6-3 (18)	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.60	0.33
Disease 6-1 (24)	0.20	0.50	0.29	0.40	0.33	0.36	0.23	0.88	0.36
Disease 2-5 (15)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Disease 6-2 (18)	0.20	0.50	0.29	0.20	0.33	0.25	0.21	0.80	0.33
Average	**0.33**	**0.33**	**0.31**	**0.31**	**0.29**	**0.29**	**0.45**	**0.64**	**0.48**

5.3. EXPLAINABLE OUTPUT

Understanding the model is essential for effective implementation in practice. As practitioners start using the model, they need to be capable of critically assessing its classification outputs. Hence, tailoring the explainability to the end-user is crucial, as these techniques may vary depending on the user group [13, 16]. This chapter will initially explore four explainability options, followed by presenting the results of a questionnaire administered to the end-users regarding these options. Finally, it will discuss the development of a final design, achieved through a qualitative focus group session, incorporating insights from both the questionnaire and the four initial options. The full answers to the questionnaire with its corresponding questions can be found in Appendix I.

5.3.1. LOCAL EXPLAINABILITY OPTIONS

The four explainability options in this section are all based on the same model outcome. In these scenarios, the practitioner has conducted an initial assessment of the patient, collecting relevant information. Once the practitioner enters this data into the system, the model generates three possible diagnoses. The number of diagnoses presented may vary depending on the model's confidence level and the input provided by the practitioner. Following the model's output, the practitioner can access an "explanation" tab, indicated by the letter 'i', accompanying each diagnosis. In each example, this tab is available for both the first and second diagnoses. Below, each explainability option will be examined in detail.

Option 1 . Probability with Reasons: The first option for explainability presents two types of data. Firstly, it displays a percentage at the top, indicating the likelihood the model assigns to a particular diagnosis. This percentage differs between the SLC and the MLC. In the case of the SLC, the percentages sum up to 100%, while for the MLC, each percentage represents an independent probability. Secondly, the model provides information on the features contributing to its diagnosis prediction. It lists the four features with the strongest positive correlation to the predicted diagnosis. Each feature either positively or negatively influences the model's prediction, and this display highlights the four most positively influential features that are filled in by the practitioner.

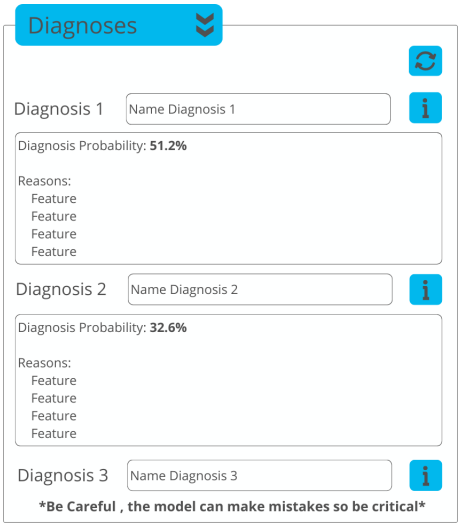


Figure 5.8: Example of Option 1, Percentage Chance with Reasoning

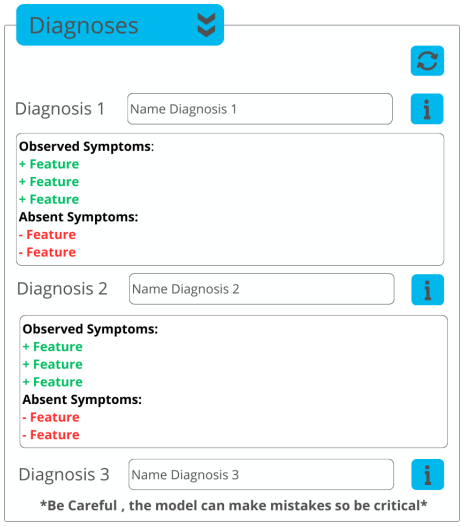


Figure 5.9: Example of Option 2, Existing and Missing Symptoms

Option 2 . Existing and Missing Diagnoses: The second option shows the practitioner two opposite types of explanations. The first values are values that the practitioner filled in as true and are correlated to this disease. For instance, if the patient experiences a sharp sensation, the practitioner documents this symptom, which aligns with the identified disease. Conversely, the second option also highlights symptoms that are absent or not disclosed by the patient. These are manifestations typically associated with the disease but are not documented in the report. This enables the practitioner to independently assess whether the condition remains a plausible diagnosis despite the absence of these symptoms.

Option 3 . Percentage of Positive and Negative Influences: The third option offers insight into the most impactful factors associated with the disease, irrespective of the practitioner's input. This model presents five values demonstrating the highest positive or negative influence on the disease. In this explanatory approach, the emphasis is on the significance of features influencing the disease's likelihood. Positive influences are visually highlighted in green, while negative influences are depicted in red. Additionally, the model provides the percentage of influence each value holds within the model. Such explanatory mechanisms offer the practitioner valuable insights into the behaviour of the model, aiding in informed decision-making processes.

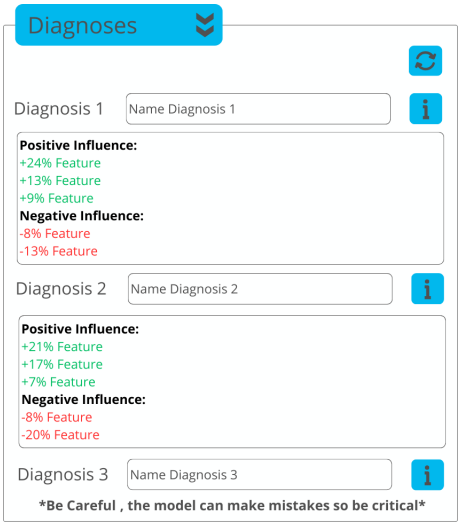


Figure 5.10: Example of option 3, Percentual Positive and Negative Influence

Option 4 . Strength-Based Reasoning: The last explainability option shows the strength a feature has on the prediction of the final diagnosis. It exclusively highlights values entered by the practitioner, ensuring that only documented symptoms are considered. For instance, if a patient lacks one symptom and it is omitted from the report, it does not factor into this explanation. Moreover, this model employs a visual representation of feature importance using plus and minus symbols. The prominence of a filled-in value is indicated by the number of plus or minus symbols assigned to it. This approach replaces percentages, offering an alternative ranking mode for individuals who may find numerical representations challenging.

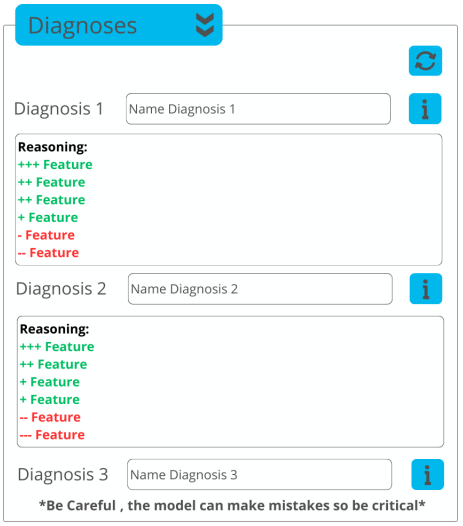


Figure 5.11: Example of option 4, Strength-Based Reasoning

5.3.2. QUESTIONNAIRE RESULTS

To create the best possible post-hoc explainability, a questionnaire was conducted involving thirteen practitioners. The evaluation of model explainability through a structured questionnaire, inspired by the work of Naiseh et al. [24], provided valuable insights into the perceptions and preferences of practitioners regarding different model presentations. This section covers the results of the questionnaire for each three of the topics. Furthermore, it highlights the rankings and extra information the practitioners filled in. This information was also used for the focus group which will be covered after.

1. Understandability: The initial segment of the questionnaire focused on assessing the intuitiveness of the models. Option 1 and Option 2 emerged as the top performers, a trend supported by participant feedback. Respondents expressed concerns regarding Option 3, citing an overload of percentage figures, while the interpretation of the +/- indicators in Option 4 remained ambiguous. Conversely, Option 1 and Option 2 were praised for their intuitive design and ease of comprehension. This preference for simplicity and clarity underscores the importance of user-friendly interfaces in enhancing model understanding.

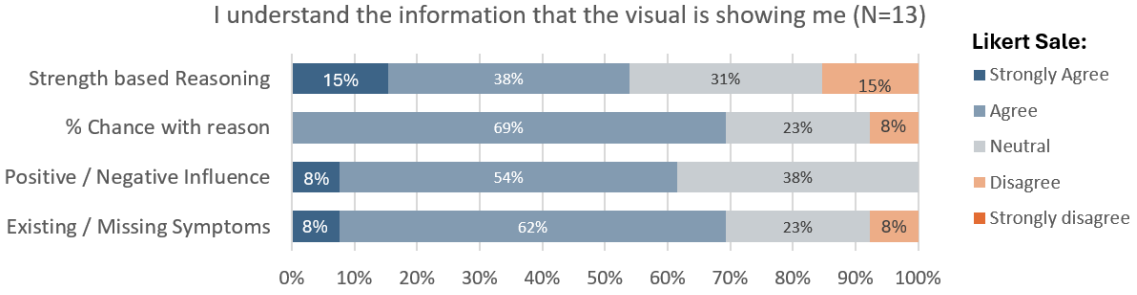


Figure 5.12: Filled in Answer per Model for the Question if the Model is Understandable and Intuitive

2. Information Sufficiency: The second category of questions revolved around the adequacy and clarity of information provided by the models. Participants noted that while Option 4 offered sufficient information, its presentation style was not well-received. Specifically, there was a desire to visualize which values exerted positive or negative influences on the model's decisions. Conversely, feedback indicated that both Option 2 and Option 1 left participants feeling as though they lacked some essential information necessary to fully understand the rationale behind the model's decisions. This sentiment underscores the importance of providing comprehensive insights to users, ensuring transparency and confidence in the decision-making process.

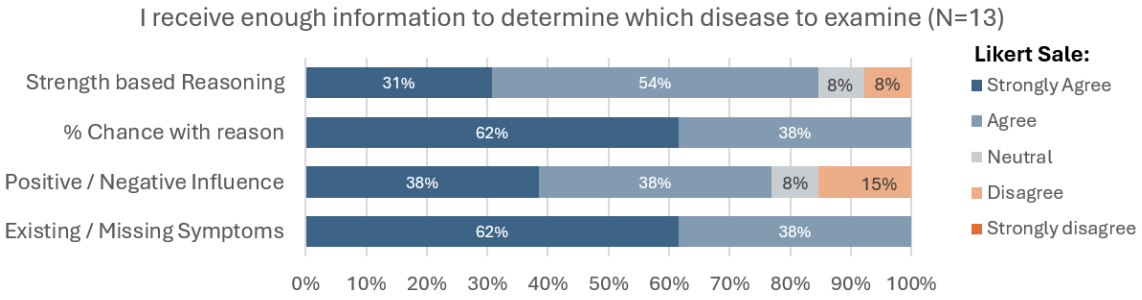


Figure 5.13: Filled in Answer per Model for the Question if the Model Provides Enough and Relevant Information

3. Reliability Evaluation: The final criterion assessed whether practitioners could ascertain the accuracy of the diagnosis output from the models. Reliability, particularly concerning the presentation of percentages, emerged as a contentious issue among respondents. While some participants expressed that the inclusion of percentages instilled confidence in the model and advocated for their presence in the final model, others found the percentage representation overwhelming and preferred a more simplified approach. Moreover, Option 4's structure elicited scepticism among participants, as its composition left them uncertain about its meaning, consequently diminishing trust in the model's reliability. In summary, the evaluation of reliability highlighted the importance of clear and transparent representations in fostering trust and confidence in the diagnostic models among practitioners.

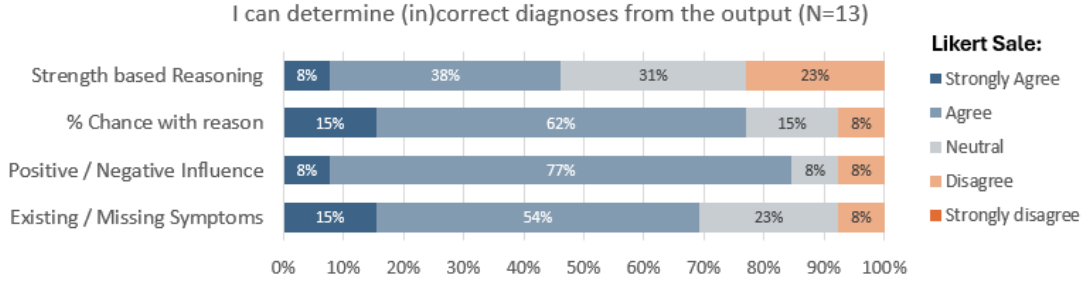


Figure 5.14: Filled in Answer per Model for the Question if it is Possible to Determine a (In)Correct Diagnosis Output

Table 5.2 outlines the preferences for various models both before and after explanations were provided. It was evident that Option 2 was the favoured choice by most respondents, both

initially and post-explanation, indicating its strong appeal. However, there was a consensus among participants that while Option 2 was preferred, it lacked some crucial information for it to be deemed perfect. Interestingly, after the explanation, there was a notable shift in preferences. Option 1 surpassed Option 3 in perceived efficacy, attributed to clearer reasoning provided at the bottom. This indicates the significance of transparent explanations in influencing model preference. It was also important to keep this fact in mind when starting the focus group, trying to solve the reasons why some models were not perceived as intuitive.

The final question on the questionnaire asked for the option of combining different elements of the explainable models. Responses were divided into two categories: some advocated for a synthesis of Option 2's chance information with Option 2's details, while others suggested incorporating either Option 3 or Option 4 to provide a comprehensive overview, enhancing the understanding of the subject matter. It's worth noting that while respondents favoured the use of colours, excessive colouration was deemed detrimental to comprehension. Thus, a balance between visual appeal and readability was recommended.

Table 5.2: Rankings of the Different Models Before and After Explanation

Model	**Before Explanation**		**After Explanation**	
	Avg. Rank	**Std. Dev.**	**Avg. Rank**	**Std. Dev.**
Existing and Missing Symptoms	**1.69**	**±0.82**	**1.77**	**±0.89**
% Chance with reasons	2.77	±1.05	2.38	±1.00
Positive and Negative Influence	2.54	±1.15	2.69	±1.14
Strength based Reasoning	3.00	±0.96	3.15	±0.95

5.3.3. FINAL LOCAL EXPLANATION FORMAT

The final model was developed through collaboration with a focus group comprising four individuals, including the director of healthcare at Company X, the coordinator of integral care, and a member of the data-drive care team. Integrating insights gathered from a questionnaire, a final explainability model aimed at clarity and trustworthiness was crafted. This model amalgamates key elements from four distinct explainability options. Its primary objective is for practitioners to critically evaluate model outputs, enabling them to validate potential diagnoses. Additionally, it serves to guide practitioners away from tunnel vision, encouraging consideration of alternative diseases they did not consider.

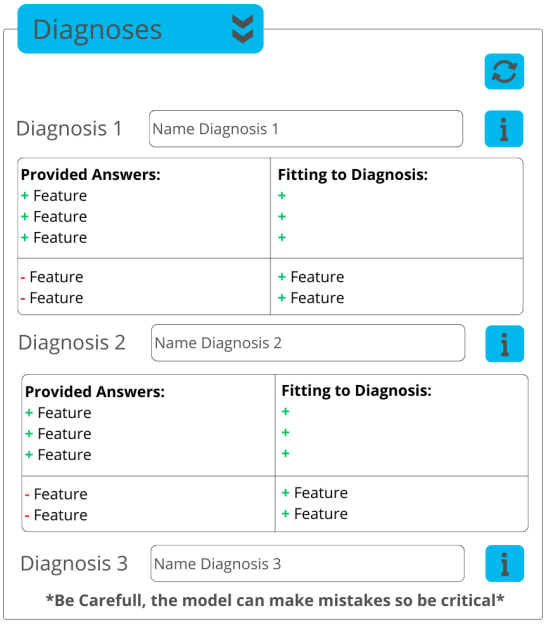


Figure 5.15: Example of the Final Version of Explainability

The final version delineates four quadrants, as illustrated in Figure 5.15. The left-side quadrants present information derived from values inputted into the patient's anamnesis report. The upper left quadrant highlights values with the highest positive correlation to the given diagnosis, affirming symptom correspondence. Conversely, the lower left quadrant identifies filled-in variables that lack correlation with the diagnosis, prompting reflection on their inclusion in the examination process. To further orient practitioners, the lower right quadrant showcases variables that align with the diagnosis, akin to those in the upper left. For instance, inputting an inflammatory symptom like redness prompts the model to suggest corresponding inflammatory symptoms specific to the diagnosis, rather than random variables. These groupings echo those discussed in Section 5.1.1. Finally, the upper right quadrant validates the correctness of variables in the upper left. The number of variables in the upper and lower halves may vary. For example, if no variables contradict the diagnosis, the count of positive variables may expand beyond the illustrated three in Figure 5.15.

It is important to note that this model does not include any percentages for the features or the final prediction score. This decision stems from its polarizing reception, with some practitioners favouring it while others disapprove. Moreover, omitting percentages aims to mitigate tunnel vision among practitioners. By encouraging consideration of less prevalent diseases alongside common ones, this approach fosters improved patient care through more nuanced diagnostic deliberation.

5.3.4. LIME EXPLAINABILITY

The final explainability of the model incorporates both positive and negative influences on the predictions. A prime example of such output is provided by LIME. For the CICST model, LIME was used during the modelling phase to illustrate its outputs and assist in refining the model. As demonstrated in Figure 5.16, LIME can help identify why a particular prediction was made, highlighting relevant factors. These insights can be integrated into the final explainability version discussed in Section 5.15.

However, while LIME is beneficial during model development, it is not ideal for the final output provided to practitioners. LIME tends to include many unfilled values, which might be useful for understanding what the model considers important but can be confusing for practitioners. These values are often not filled in by practitioners and typically default to "no", leading to unnecessary complexity in the explanations.

Figure 5.16: LIME Output of a True Positive value for Disease 1-2

Not only do LIME outputs aid in recognizing correct diagnoses, but they are also instrumental in identifying incorrect ones. Figure 5.17 illustrates a diagnosis that met the threshold for model output. The expected probability of 13% exceeds the threshold of 10%, as detailed in Appendix H. However, this diagnosis is incorrect. Notably, the patient does not experience any pain during sporting or warming-up activities. In a real-world scenario, this discrepancy could prompt a practitioner to ask follow-up questions, steering the diagnosis in the right direction.

Figure 5.17: LIME Output of a False Positive Value for Disease 1-4

Finally, LIME proves valuable in understanding why certain diagnoses are not selected. These insights during modelling are crucial for comprehending how the model reacts to data vastly different from the actual disease, shedding light on its decision-making process.

5.3.5. INTEGRATION OF ML BASED CDSS IN CURRENT ANAMNESIS PROCESS

The integration of the ML model into the current process primarily involves the improvement of filling in the anamnesis report. With the implementation of the ML model, this process will undergo slight modifications. Figure 5.18 illustrates how the final model could be integrated into the workflow.

In practice, once the practitioner completes the anamnesis report, the data will be submitted to the ML model. The model will then return a set of potential diagnoses, which is formatted using the explainability discussed in Section 5.3.3. Upon reviewing the model's output, the practitioner can determine which diagnoses warrant further examination. This may include the diagnoses suggested by the model, a subset of these diagnoses, or additional diagnoses that

the model may have overlooked. Following this decision, the subsequent steps in the diagnostic process will proceed as they currently do.

This sequence allows for effective monitoring of the model's performance in a real-world setting. By saving the output from step 3.1 alongside the outputs from steps 4 and 10, it will be possible to evaluate how often the practitioner adopts the model's suggested diagnoses. Furthermore, this comparison can highlight the model's accuracy in predicting correct diagnoses.

A prototype has been developed based on the CICST model, which mimics steps 2 and 3. In this prototype, features can be entered into a digital document, which then provides a list of potential diagnoses along with explanations. An visual example of this prototype can be found in Appendix J. This prototype serves as an initial draft, offering practitioners a preview of the model's functionality and the type of support they can expect. By integrating the ML model in this manner, the diagnostic process can become more efficient and potentially more accurate, benefiting both practitioners and patients.

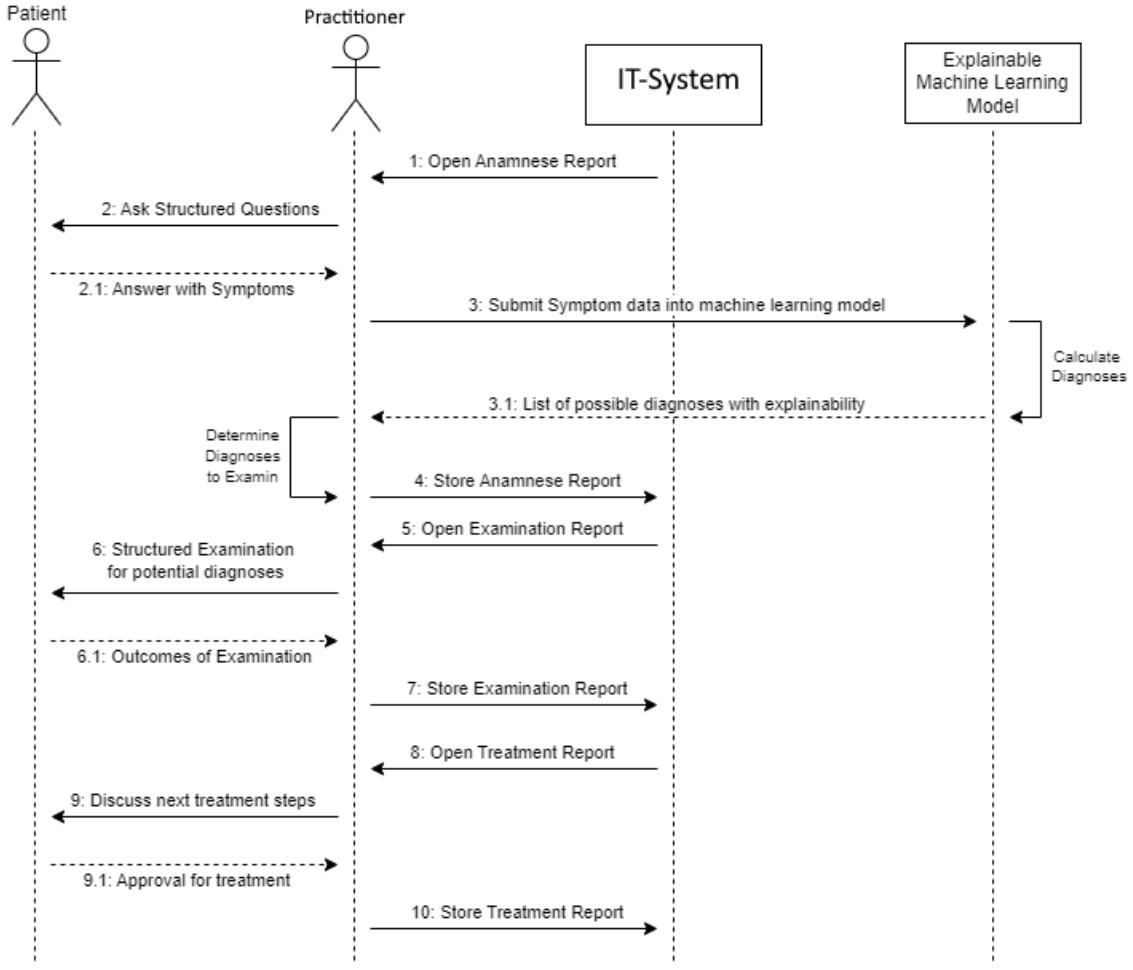


Figure 5.18: Sequence Diagram of the New Process When Implemented in Practice

6

Conclusion

This research focused on developing a XAI based CDSS utilizing EHR data. The primary objective of this CDSS is to assist practitioners in diagnosing patients during anamnesis appointments. The system aims to help practitioners by classifying potential diagnoses and providing explanations for these classifications based on the information recorded in the anamnesis report. In this chapter, we will address the main research questions using the insights gained throughout this study.

How can EHR data be used to classify patient diagnoses?

The data collected during CDSS reports during the anamnesis were initially presented in an unorganized format. It was therefore necessary to first model the data into a star schema format, with dimensions for the relevant information. After successfully modelling the data, it was possible to filter the features that were not correlated to the disease or were rarely filled in. This filtering resulted in a reduction of 1153 features to 299. The structured EHR data enables the application of predetermined answers, which is beneficial for feature engineering. By combining the data, the feature space was further reduced from 299 to 51 features. Additionally, the model identified and addressed missing data for both nominal and ordinal values. For the NRS feature, the missing values were replaced with the average value of the corresponding feature, resulting in the cleaning of approximately 13% of the data. For ordinal data, the missing values were set to the median, which was applied to five features on average, accounting for 16% of the missing data per feature. The organized structure of these reports will facilitate the application of ML and data science techniques.

How can diseases be automatically diagnosed during an anamnesis to support the decision-making of a practitioner?

Three distinct models were trained on the anamnesis data. The MLC models were trained using a GBM provided by the H2O platform, while the MLC model was trained using a RF. The

models were evaluated using Sensitivity, Precision, and F1-score metrics. Regrettably, the SLC models did not achieve high scores due to some diagnoses receiving zeros in all three metrics for multiple diagnoses. The two SLC models both had scores close to 0.3 for all three metrics respectively. However, the MLC model was deemed the best choice due to its flexible thresholds, current good performance, and anticipated superior performance in the future. The current performance of the MLC model stands at 0.64 for precision, 0.45 for sensitivity, and 0.48 for F1-score. In this particular use case, precision is of particular importance. Additionally, introducing the new treatment report is expected to increase the percentage of reports containing at least one diagnosis from the current 10% to an even higher percentage in the future. Lastly, the flexible thresholds of the MLC model provide the practitioner with the ability to adjust the frequency of diseases that occur too often or not often enough.

> *How can XAI methods be used to support practitioners in their decision-making during the anamnesis?*

The analysis of global feature importance was conducted to identify the features with the biggest reductions in variance for the distinct ML models. The study found that the location variables were crucial for all models. Partial dependency analysis was utilized to determine the most important features for each diagnosis, and the results showed that the models produced similar findings. This underscores the significance of location in predicting diagnoses. Moreover, the local explainability output was assessed by creating four models and analyzing the preferences of practitioners. The findings were used to create a final model that offers insights into the rationale behind the final diagnosis. The final explainability model poses critical questions to practitioners and highlights incorrect symptoms and their alternatives. Furthermore, it indicates what inputted variables are related to the classified diseases. This explainability promotes practitioners' ability to critically analyze the ML model's output and make autonomous decisions on which diagnoses to further examine. The CICST Model implementation of LIME provides the underlying values for the final explainability model. LIME is a useful tool for examining model outputs during the modelling phase and understanding why certain results were generated. The final model serves as a good indication of what the final model could look like, which is visualized in Appendix J.

> *How to design and integrate a XAI based CDSS to identify the most influential variables and support practitioners in diagnosing patients?*

Overall, this study provides a detailed application of the Crisp-DM framework in developing a ML model for the anamnesis process, covering all phases from business understanding to evaluation. The current iteration has progressed to the stage of actual deployment, resulting in the creation of a prototype and an analysis of potential future developments. Although the present performance of the models is not yet sufficient for full implementation, a solid foundation has been laid. Each part of the process, from data gathering, data cleaning, and feature engineering, to modeling, has been systematically addressed, leading to the development of a prototype that demonstrates the system's current capabilities. The ML model shows promising

results but requires further refinement to be deployment-ready. Additional data collection is essential for making significant improvements, particularly in the feature engineering and modelling phases. This ongoing process will enhance the model's performance, ultimately making it more robust and effective for clinical use.

6.1. PRACTICAL AND SCIENTIFIC CONTRIBUTIONS

From a practical standpoint, this research has provided valuable insight to Company X on the feasibility of utilizing structured EHR data. The entire process, ranging from data retrieval, cleaning, and feature engineering for a machine learning model, has been demonstrated. Moreover, the modelled data applies to other types of CDSS reports that can be generated, resulting in a more comprehensive analysis for all possible report types at Company X, which can be found in Section 4.5. Lastly, the research showcases a possible XAI output that can be used for the final version of the ML model based on this anamnesis data, displayed in Section 5.3.3.

From a scientific perspective, this research examines the use of structured EHR data in the healthcare industry. Although structured data in healthcare is relatively new, there is a growing interest in using it to make reliable analyses and conduct further data science projects. This research highlights how structured EHR data can be utilized for a ML model, as can be seen in Chapter 4, and the challenges that may arise when creating or using such data. In terms of generalization, the MLC model performed better than the two SLC models while looking at the cross-validated performances from Table 5.1. However, the performance of the classification model appeared weaker for less common classes, suggesting room for improvement in handling imbalanced data. Following, the research showcases how global feature importance and dependency values could be handled for variables and their relation to diseases, showcased in Section 5.1.1.

Finally, the main contribution of this research focuses on explainability. First, a thorough literature review is conducted in Chapter 2, exploring the current activities of XAI in a healthcare setting. This review showcases other researchers' investigations into various XAI techniques in healthcare, highlights the challenges of implementing XAI, presents different XAI options, and examines the impact of XAI in healthcare. Building on the insights from the literature review, this research implements a custom XAI output. Currently, there is limited information in the literature on final designs validated by healthcare professionals. This research addresses this gap by introducing four different explainability options, showcased in Section 5.3. It emphasizes a user-centric implementation approach, involving practitioners in a questionnaire to review these options and utilizing a specialized focus group to develop the final explainability model, detailed in Section 5.3.3. This study underscores the importance of involving end-users in creating XAI models, particularly in the healthcare sector.

6.2. LIMITATIONS AND FUTURE RESEARCH RECOMMENDATIONS

Despite the valuable insights presented in the previous section, this research has several limitations that must be considered when evaluating the outputs and steps of this study critically. Furthermore, there is scope for further research that can build upon this study to make additional contributions to the literature and provide practical implications.

6.2.1. LIMITATIONS

The following are the limitations of this study that need to be taken into account when reviewing the results. These are areas where improvements can be made to enhance the insights in this study.

1. **Limitation to Subset of Diseases:** This study is focused only on a subset of diseases, leaving nine other disease categories unexplored. Additionally, many of the diagnoses within the included categories did not meet the threshold of 15 samples. Therefore, the set of 20 diseases could be expanded to include all 177 diagnoses.

2. **Amount of Data:** Currently, there is insufficient data to train a complete ML model for practical application. It is challenging to quantify the amount of data required for accurate predictions. However, looking at the results of the current model, it can not successfully predict diseases yet for a practical implementation.

3. **Data Quality:** The anamnesis reports used in this research are still in their early stages and are being improved and restructured to enhance their quality and usability. Additionally, the practitioners involved in the study were new to these reports, which could have impacted the data's quality. It is possible that the models were trained on data that was still in development, leading to inaccuracies and errors in the model's classification.

4. **Final Explainability Restrictions:** The final model had to be text-based to be compatible with the current system, which limited the model's ability to include graphical explanations. These explanations are commonly used by LIME and SHAP. This restriction on explainability could have limited the model's ability to fully utilize its explainable capabilities. However, this depends on the preferences of the end-users.

6.2.2. FUTURE RESEARCH RECOMMENDATIONS

Lastly, in light of the current research's restrictions and findings, additional research could be conducted to expand upon this study. This section highlights three potential areas for further exploration.

1. **Diagnostic Classification Categories:** Presently, the research only encompasses twenty of the potential one hundred seventy diagnoses. Establishing a reliable machine learning (ML) model for all one hundred seventy diagnoses presents a significant challenge. As a result, it is essential to explore strategies for possibly grouping these diagnoses prior to utilizing an ML model to make final determinations. The following are potential ap-

proaches to be investigated for creating these categories:

- **Cost Sensitive Thresholding:** During the research of Liu et al. [48], Cost sensitive thresholding was used. However, for this study, this aspect was not implemented, and only normal thresholds were implemented. Introducing cost sensitivity could be beneficial for diagnoses that exhibit strong similarities. Currently, each disease has its own threshold and one-vs-rest classifier. It would be advantageous if the model only returned the pertinent diagnosis based on the cost matrix, rather than presenting two highly similar diseases.
- **Clustering:** Instead of constructing the cost matrix, clustering or consolidating diagnoses could be explored as an alternative approach. Grouping diagnoses based on similarities and then developing classifiers for these groups is an effective means of accurate disease classification. However, implementing explainability in clustering models may prove to be more challenging.
- **Rule-Based Disease Separation:** Prior to allowing the ML model to make determinations, it may be feasible to employ three or four general questions to separate most of the diagnoses. Implementing rule-based separation methods beforehand could significantly reduce the number of predictions required by the model.

2. **Enhanced Features:** The report and feature engineering could be refined based on the outcomes of the global and local feature analysis. This would provide insights into the relative importance of different data points and suggest potential combinations or eliminations. Additionally, it could offer insights into which types of features might be introduced to better distinguish between different diagnoses.

3. **Autonomous Patient Diagnosis:** Once the model demonstrates improved performance, research could also be conducted on developing an autonomous system. In this scenario, patients could complete a questionnaire independently. Given the high pressure on healthcare resources and the growing demand for them, patients could potentially engage in certain exercises or preventive measures that would not worsen their condition during the wait for an appointment with a practitioner, based on the model's output. This would enable patients to prevent the progression of their condition while awaiting a consultation.

4. **Researching a bigger dataset:** Currenlty, the dataset is not large enough for ML predictions. However, these reports are currently still being used, and are probably expanded organization-wide. Therefore, after a certain timeframe, the dataset will have enough data for a follow-up research of the practical implementation of the ML models. Typically 5-10 samples per feature would be expected, meaning at least 255 samples per class would be necessary. However, even the most common class does not have this amount of data yet.

5. **Ontology and Linked Data:** Currently, the data is structured in a star schema to retrieve

additional information behind the questions in the reports. However, exploring the creation of an ontology or even a Linked Data set would be very beneficial. This approach could enhance explainability by highlighting the semantics behind the data, making it more understandable for outsiders. Additionally, it could facilitate the integration of data from different reports. For instance, Linked Data could connect symptoms to possible diseases and link these diseases to corresponding treatment and examination steps. This would enable the identification of treatment and examination steps practitioners use for various symptoms and diseases, providing valuable insights.

REFERENCES

[1] F. Di Martino, F. Delmastro. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artif. Intell. Rev. 56 (2023) 5261–5315.

[2] E. Kristoffersen, O. O. Aremu, F. Blomsma, P. Mikalef, J. Li. Exploring the relationship between data science and circular economy: An enhanced crisp-dm process model (2019) 177–189. doi:10.1007/978-3-030-29374-1_15.

[3] T. Chen, C. Shang, P. Su, E. Keravnou-Papailiou, Y. Zhao, G. Antoniou, Q. Shen. A decision tree-initialised neuro-fuzzy approach for clinical decision support. Artificial Intelligence in Medicine 111 (2021). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097572903&doi=10.1016%2fj.artmed.2020.101986&partnerID=40&md5=24441e2e89b50871fb078ee747a1e230. doi:10.1016/j.artmed.2020.101986, cited by: 27; All Open Access, Green Open Access, Hybrid Gold Open Access.

[4] T. Murari, L. Prathiba, K. K. Singamaneni, D. Venu, V. K. Nassa, R. Kohar, S. S. Uparkar. Big data analytics with oenn based clinical decision support system. Intelligent Automation and Soft Computing 31 (2022) 1241 – 1256. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116252009&doi=10.32604%2fiasc.2022.020203&partnerID=40&md5=1e256962fcc6dde559a87b80bf1a4b90. doi:10.32604/iasc.2022.020203, cited by: 0; All Open Access, Hybrid Gold Open Access.

[5] A. Khodadadi, N. Ghanbari Bousejin, S. Molaei, V. Kumar Chauhan, T. Zhu, D. A. Clifton. Improving diagnostics with deep forest applied to electronic health records. Sensors 23 (2023) 6571. URL: http://dx.doi.org/10.3390/s23146571. doi:10.3390/s23146571.

[6] J. Tapia-Galisteo, J. M. Iniesta, C. Perez-Gandia, G. Garcia-Saez, D. U. Puertolas, F. J. Izquierdo, M. E. Hernando. Prediction of cocaine inpatient treatment success using machine learning on high-dimensional heterogeneous data. IEEE Access 8 (2020) 218936 – 218953. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097373394&doi=10.1109%2fACCESS.2020.3041895&partnerID=40&md5=87856fd0ab8b212010e017e5f147fbb1. doi:10.1109/ACCESS.2020.3041895, cited by: 2; All Open Access, Gold Open Access.

[7] A. M. Antoniadi, M. Galvin, M. Heverin, L. Wei, O. Hardiman, C. Mooney. A clinical decision support system for the prediction of quality of life in ALS. J. Pers. Med. 12 (2022) 435.

[8] J. M. de Oliveira, C. A. da Costa, R. S. Antunes. Data structuring of electronic health records: a systematic review. Health and Technology 11 (2021) 1219–1235. URL: http://dx.doi.org/10.1007/s12553-021-00607-w. doi:10.1007/s12553-021-00607-w.

[9] S. I. Lambert, M. Madi, S. Sopka, A. Lenes, H. Stange, C.-P. Buszello, A. Stephan. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. NPJ Digit. Med. 6 (2023) 111.

[10] S. Mischos, E. Dalagdi, D. Vrakas. Intelligent energy management systems: a review. Artificial Intelligence Review 56 (2023) 11635–11674. URL: http://dx.doi.org/10.1007/s10462-023-10441-3. doi:10.1007/s10462-023-10441-3.

[11] R. Alsaigh, R. Mehmood, I. Katib. Ai explainability and governance in smart energy systems: A review. Frontiers in Energy Research 11 (2023). URL: http://dx.doi.org/10.3389/fenrg.2023.1071291. doi:10.3389/fenrg.2023.1071291.

[12] S. Bahoo, M. Cucculelli, X. Goga, J. Mondolo. Artificial intelligence in finance: a comprehensive review through bibliometric and content analysis. SN Business amp; Economics 4 (2024). URL: http://dx.doi.org/10.1007/s43546-023-00618-x. doi:10.1007/s43546-023-00618-x.

[13] A. J. Barda, C. M. Horvat, H. Hochheiser. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. BMC Medical Informatics and Decision Making 20 (2020). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092601699&doi=10.1186%2fs12911-020-01276-x&partnerID=40&md5=f34d0534504ebb2a21be35b7a1941497. doi:10.1186/s12911-020-01276-x, cited by: 37; All Open Access, Gold Open Access, Green Open Access.

[14] R. Bartels, J. Dudink, S. Haitjema, D. Oberski, A. van 't Veen. A perspective on a quality management system for ai/ml-based clinical decision support in hospital care. Frontiers in Digital Health 4 (2022). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134568360&doi=10.3389%2ffdgth.2022.942588&partnerID=40&md5=4d11abdacb05744c81c5f68af3951fad. doi:10.3389/fdgth.2022.942588, cited by: 2; All Open Access, Gold Open Access, Green Open Access.

[15] F. M. Calisto, C. Santiago, N. Nunes, J. C. Nascimento. Breastscreening-ai: Evaluating medical intelligent agents for human-ai interactions. Artificial Intelligence in Medicine 127 (2022). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127146666&doi=10.1016%2fj.artmed.2022.102285&partnerID=40&md5=12824bc276dcf4cee4d6aaeab8a5db58. doi:10.1016/j.artmed.2022.102285, cited by: 31; All Open Access, Hybrid Gold Open Access.

[16] N. Bienefeld, J. M. Boss, R. Lüthy, D. Brodbeck, J. Azzati, M. Blaser, J. Willms, E. Keller. Solving the explainable ai conundrum by bridging clinicians' needs and developers' goals.

npj Digital Medicine 6 (2023). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160021440&doi=10.1038%2fs41746-023-00837-4&partnerID=40&md5=71592c85edf5536584140161c980f0b8. doi:10.1038/s41746-023-00837-4, cited by: 4; All Open Access, Gold Open Access, Green Open Access.

[17] Y. Du, A. M. Antoniadi, C. McNestry, F. M. McAuliffe, C. Mooney. The role of xai in advice-taking from a clinical decision support system: A comparative user study of feature contribution-based and example-based explanations. Applied Sciences (Switzerland) 12 (2022). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140431929&doi=10.3390%2fapp122010323&partnerID=40&md5=2f93fa4551c2a74f8ca38e4ff434f871. doi:10.3390/app122010323, cited by: 4; All Open Access, Gold Open Access.

[18] ActiZ, D. N. ggz, F. M. Specialisten, InEen, N. F. van Universitair Medische Centra, N. V. van Ziekenhuizen, N. Zorgautoriteit, P. Nederland, V. van Nederlandse Gemeenten, V. . V. Nederland, Z. K. Nederland, Z. Nederland, Zorgthuisnl, Z. Nederland, W. e. S. Ministerie van Volksgezondheid, Integraal zorg akkoord: Samen werken aan gezonde zorg, 2022. URL: https://www.rijksoverheid.nl/documenten/rapporten/2022/09/16/integraal-zorgakkoord-samen-werken-aan-gezonde-zorg.

[19] D. M. Berwick, T. W. Nolan, J. Whittington. The triple aim: care, health, and cost. Health Aff. (Millwood) 27 (2008) 759–769.

[20] M. Wang, C. Lee, Z. Wei, H. Ji, Y. Yang, C. Yang. Clinical assistant decision-making model of tuberculosis based on electronic health records. BioData Mining 16 (2023). URL: http://dx.doi.org/10.1186/s13040-023-00328-y. doi:10.1186/s13040-023-00328-y.

[21] J. M. Pavon, L. Previll, M. Woo, R. Henao, M. Solomon, U. Rogers, A. Olson, J. Fischer, C. Leo, G. Fillenbaum, H. Hoenig, D. Casarett. Machine learning functional impairment classification with electronic health record data. Journal of the American Geriatrics Society 71 (2023) 2822–2833. URL: http://dx.doi.org/10.1111/jgs.18383. doi:10.1111/jgs.18383.

[22] Y. Raita, T. Goto, M. K. Faridi, D. F. M. Brown, C. A. Camargo, K. Hasegawa. Emergency department triage prediction of clinical outcomes using machine learning models. Critical Care 23 (2019). URL: http://dx.doi.org/10.1186/s13054-019-2351-7. doi:10.1186/s13054-019-2351-7.

[23] C.-C. Chiu, C.-M. Wu, T.-N. Chien, L.-J. Kao, C. Li, C.-M. Chu. Integrating structured and unstructured ehr data for predicting mortality by machine learning and latent dirichlet allocation method. International Journal of Environmental Research and Public Health 20 (2023) 4340. URL: http://dx.doi.org/10.3390/ijerph20054340. doi:10.3390/ijerph20054340.

[24] M. Naiseh, D. Al-Thani, N. Jiang, R. Ali. How the different explanation classes impact trust calibration: The case of clinical decision support systems. International Journal of Human Computer Studies 169 (2023). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139859152&doi=10.1016%2fj.ijhcs.2022.102941&partnerID=40&md5=87e55eeea60ef783e02b78186514493b. doi:10.1016/j.ijhcs.2022.102941, cited by: 11; All Open Access, Green Open Access, Hybrid Gold Open Access.

[25] Y. Jia, J. McDermid, T. Lawton, I. Habli. The role of explainability in assuring safety of machine learning in healthcare. IEEE Transactions on Emerging Topics in Computing 10 (2022) 1746 – 1760. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85132514307&doi=10.1109%2fTETC.2022.3171314&partnerID=40&md5=bd0b9ea9dcaedd51b49721341ef72ed8. doi:10.1109/TETC.2022.3171314, cited by: 13; All Open Access, Green Open Access, Hybrid Gold Open Access.

[26] G. V. Aiosa, M. Palesi, F. Sapuppo. Explainable ai for decision support to obesity comorbidities diagnosis. IEEE Access 11 (2023) 107767 – 107782. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85173024031&doi=10.1109%2fACCESS.2023.3320057&partnerID=40&md5=fbf58576de04a0e34e47cd360a7f4bc4. doi:10.1109/ACCESS.2023.3320057, cited by: 0; All Open Access, Gold Open Access.

[27] F. Giuste, W. Shi, Y. Zhu, T. Naren, M. Isgut, Y. Sha, L. Tong, M. Gupte, M. D. Wang. Explainable artificial intelligence methods in combating pandemics: A systematic review. IEEE Reviews in Biomedical Engineering 16 (2023) 5 – 21. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133767411&doi=10.1109%2fRBME.2022.3185953&partnerID=40&md5=b7fd325e1596b2fa6a51abe24ba3ec78. doi:10.1109/RBME.2022.3185953, cited by: 14; All Open Access, Green Open Access, Hybrid Gold Open Access.

[28] V. Blanes-Selva, S. Asensio-Cuesta, A. Doñate-Martínez, F. Pereira Mesquita, J. M. García-Gómez. User-centred design of a clinical decision support system for palliative care: Insights from healthcare professionals. Digital Health 9 (2023). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146179361&doi=10.1177%2f20552076221150735&partnerID=40&md5=251e63be8a3f975dc2938fc564688c76. doi:10.1177/20552076221150735, cited by: 2; All Open Access, Gold Open Access, Green Open Access.

[29] G. Jagadamba, R. Shashidhar, H. Gururaj, R. Vinayakumar, A. Meshari, A. Yasser. Electronic health record (ehr) system development for study on ehr data-based early prediction of diabetes using machine learning algorithms. Open Bioinformatics Journal 16 (2023). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174296026&doi=10.2174%2f18750362-v16-e230906-2023-15&partnerID=40&md5=e8a6720751de94620a2edb38da664c46. doi:10.2174/18750362-v16-e230906-2023-15, cited by: 0; All Open Access, Gold Open Access.

[30] L. Pumplun, F. Peters, J. F. Gawlitza, P. Buxmann. Bringing machine learning systems into clinical practice: A design science approach to explainable machine learning-based clinical decision support systems. Journal of the Association for Information Systems 24 (2023) 953 – 979. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85165725989&doi=10.17705%2f1jais.00820&partnerID=40&md5=6ee960b165d51988a5677896f7c9ab8e. doi:10.17705/1jais.00820, cited by: 0; All Open Access, Bronze Open Access.

[31] M. Nuutinen, R.-L. Leskelä. Systematic review of the performance evaluation of clinicians with or without the aid of machine learning clinical decision support system. Health Technol. (Berl.) 13 (2023) 1–14.

[32] J. Iqbal, D. C. Cortés Jaimes, P. Makineni, S. Subramani, S. Hemaida, T. R. Thugu, A. N. Butt, J. T. Sikto, P. Kaur, M. A. Lak, M. Augustine, R. Shahzad, M. Arain. Reimagining healthcare: Unleashing the power of artificial intelligence in medicine. Cureus 15 (2023) e44658.

[33] D. Barrera Ferro, S. Brailsford, C. Bravo, H. Smith. Improving healthcare access management by predicting patient no-show behaviour. Decision Support Systems 138 (2020). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089896434&doi=10.1016%2fj.dss.2020.113398&partnerID=40&md5=39fcd73f75d210e0a2d02fce00a605dd. doi:10.1016/j.dss.2020.113398, cited by: 19; All Open Access, Green Open Access.

[34] A. Choudhury. Factors influencing clinicians' willingness to use an ai-based clinical decision support system. Frontiers in Digital Health 4 (2022). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141185282&doi=10.3389%2ffdgth.2022.920662&partnerID=40&md5=ebc1afc000b9a3aec404f1492d23a06c. doi:10.3389/fdgth.2022.920662, cited by: 4; All Open Access, Gold Open Access, Green Open Access.

[35] M. M. Hasan, G. J. Young, J. Shi, P. Mohite, L. D. Young, S. G. Weiner, M. Noor-E-Alam. A machine learning based two-stage clinical decision support system for predicting patients' discontinuation from opioid use disorder treatment: retrospective observational study. BMC Medical Informatics and Decision Making 21 (2021). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119979571&doi=10.1186%2fs12911-021-01692-7&partnerID=40&md5=2283a6e57b92213e84162f40974ef61c. doi:10.1186/s12911-021-01692-7, cited by: 7; All Open Access, Gold Open Access, Green Open Access.

[36] G. Mahadevaiah, P. Rv, I. Bermejo, D. Jaffray, A. Dekker, L. Wee. Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. Med. Phys. 47 (2020) e228–e235.

[37] F. Magrabi, E. Ammenwerth, J. B. McNair, N. F. De Keizer, H. Hyppönen, P. Nykänen, M. Rigby, P. J. Scott, T. Vehko, Z. S.-Y. Wong, A. Georgiou. Artificial intelligence in clinical decision support: Challenges for evaluating AI and practical implications. Yearb. Med. Inform. 28 (2019) 128–134.

[38] L. Rundo, R. Pirrone, S. Vitabile, E. Sala, O. Gambino. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. J. Biomed. Inform. 108 (2020) 103479.

[39] D. van de Sande, M. E. van Genderen, J. Huiskens, D. Gommers, J. van Bommel. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med. 47 (2021) 750–760.

[40] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, 2016. arXiv:1602.04938.

[41] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017. arXiv:1705.07874.

[42] Gradient Boosting in ML - GeeksforGeeks — geeksforgeeks.org, https://www.geeksforgeeks.org/ml-gradient-boosting/, 2024.

[43] Random Forest Algorithm in Machine Learning - GeeksforGeeks — geeksforgeeks.org, https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/, 2024.

[44] Generalized Linear Models - GeeksforGeeks — geeksforgeeks.org, https://www.geeksforgeeks.org/generalized-linear-models/, 2024.

[45] D. Berrar, Cross-Validation, 2018. doi:10.1016/B978-0-12-809633-8.20349-X.

[46] D. Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation. Mach. Learn. Technol. 2 (2008).

[47] K. Tybjerg. Medical anamnesis. collecting and recollecting the past in medicine. Centaurus 65 (2023) 235–259. URL: http://dx.doi.org/10.1484/J.CNT.5.135348. doi:10.1484/j.cnt.5.135348.

[48] Y. Liu, Q. Li, K. Wang, J. Liu, R. He, Y. Yuan, H. Zhang. Automatic multi-label ecg classification with category imbalance and cost-sensitive thresholding. Biosensors 11 (2021) 453. URL: http://dx.doi.org/10.3390/bios11110453. doi:10.3390/bios11110453.

APPENDIX A:FULL KEYWORD QUERIES FOR LITERATURE SEARCH

Table A.1: Table containing the full queries used for the literature search

Website	Final Query Used
Scopus Query 1	TITLE-ABS-KEY (ai OR "Aritificial Intelligence" OR "Machine Learning" OR ml) AND TITLE-ABS-KEY (implementation OR design OR deployment) AND TITLE-ABS-KEY (cdss OR "Clinical Decision Support System" OR "Clinical Decision Aid" OR dss) AND TITLE-ABS-KEY (healthcare OR "Medical Care") AND NOT TITLE-ABS-KEY (mobile OR remote OR monitor) AND PUBYEAR ≥ 2019 AND PUBYEAR ≤ 2023 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (OA , "all"))
Scopus Query 2	TITLE-ABS-KEY (xai OR xml OR Explainable AI OR Transparent AI OR "Explainable Artificial Intelligence" OR Interpretable AI OR Explainable Machine Learning) AND TITLE-ABS-KEY (cdss OR "Clinical Decision Support System" OR "Clinical Decision Aid" OR dss OR Key Variables) AND TITLE-ABS-KEY (healthcare OR "Medical Care") AND NOT TITLE-ABS-KEY (mobile OR remote OR monitor) AND PUBYEAR ≥ 2019 AND PUBYEAR ≤ 2023 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "ENGI")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (OA , "all"))
Pubmed Query 1	("AI" OR "ML" OR "Artificial Intelligence" OR "Machine Learning") AND (Implementation OR Design OR Deployment) AND (cdss OR dss OR "clinical decision support systems" OR DSS OR "Clinical Decision Aid") AND (Healthcare OR "Medical Care") AND (English[Language]) AND ("2018/01/01"[Date - Publication] : "3000"[Date - Publication]) AND ("Review"[Publication Type] OR "Systematic Review"[Publication Type]) AND ("free full text"[sb])
Pubmed Query 2	("explainable artificial intelligence" OR XAI OR "Explainable AI" OR "Transparent AI" OR "Interpretable AI") AND (cdss OR dss OR "clinical decision support systems" OR DSS OR "Clinical Decision Aid") AND (Healthcare OR "Medical Care") AND (English[Language]) AND ("2018/01/01"[Date - Publication]: "3000"[Date - Publication]) AND ("Review"[Publication Type] OR "Systematic Review"[Publication Type] OR "Journal Article"[Publication Type])

APPENDIX B: LITERATURE FINDINGS

B.1. FINDINGS LITERATURE REGARDING ML IMPLEMENTATIONS IN HEALTH CARE

Table B.1: Summary of Technical Machine Learning Studies Applied in Health Care

Author	Data Type	Machine Learning Technique(s)	Metric of Evaluation	Main Findings
(Barrera Ferro et al. 2020)	Appointment and Patient Var. Data	LASSO Regression (LASSO), Layer-wise NN	effective visits, target lead time, %appointments, AUROC	Relation of Variables on No Show Behavior
(Murari et al. 2022)	Open Text, and structured text / variables datasets	Elman NN, K Nearest Neighbour (KNN)	Sensitivity, Specificity, Accuracy, F-score, Kappa	Better results using BDA-OENN model than common ML models
(Hasan et al. 2021)	structured Data: patient and treatment adherence	Decision Tree (DT), RF,XGBoost, LR,NN, SVM	Precision, Recall, F1 score, C-statistics, Receiver Operating Characteristic (ROC) curve	Model could help predict patients who risk discontinuing
(Jagadamba et al. 2023)	Structured and Unstructured EHR	RF, KNN, Naïve-Bayes (NB)	Accuracy, Precision, Performance (speed)	
(Aiosa et al. 2023)	Text datasets from Kaggle or surveys	MLP, XGBoost, LR, NN, RF, DT, Linear Support Vector (LSV)	Accuracy, Precision, Recall, F1, ROC AUC	Models to diagnose obesity types where MLP and XGB were best

(Tapia-Galisteo et al. 2020)	Unstructured data of healthcare reports	SVM, RF, LR, MLP	AUC, Recall, Specificity, F1, Matthews Correlation Coefficient	82% accuracy of prediction treatment success
(Pumplun et al. 2023)	Image data	Deep Neural Network with DenseNet Architecture	Sensitivity, Specificity, ROC Curve	Physicians perceiving the ML CDSS as more explainable and usable
(Chen et al. 2021)	Discrete Data about Glucose, BMI levels	Fuzzy DT	Accuracy per dataset	Understandable model with 90% accuracy
(Wang et al. 2023)	Structured and unstructured data	MSI-PTDM, Single Stream Tuberculosis Diagnosis Model (SS-PTDM), XGBoost, Text-Convolutional Neural Networks (CNN), RF, SVM, Bi-directional Long Short-Term Memory (Bi-LSTM),Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM)	sensitivity, specificity, accuracy and Area Under the Curve (AUC)	Model with 91% accuracy
(Khodadadi et al. 2023)	Structured EHR records	DRF	AUROC	AUROC Scores of 0.8
(Pavon et al. 2023)	EHR of activities/instruments	KNN + XGB	Accuracy, AUROC, AUPRC	Accuracy of 71%
(Raita et al. 2019)	Structured EHR data	LR, RF, GB Decision Tree, Deep NN	AUC, ROC	AUC values of 0.86

(Chiu et al. 2023)	Structured EHR data, Unstr. diagnosis data	AdaBoosting (AB), Bagging, Gradient Boosting (GB), Light GB model, LR, MLP, Support Vector Classifier (SVC), XGBoost	Precision, Recall, F1-Score, Accuracy, AUROC

B.2. Findings Literature regarding XAI

Table B.2: Overview of XAI related literature including ways of gaining trust, the used XAI techniques and findings of the final explanation model or display

Author	Trust Via XAI	XAI Techniques	XAI models, Displays
(Naiseh et al. 2023)	- Cognition-based trust and affect-based trust (HCT Gregor) - Explanations task-centred, familiar, simple, reliable, casually relevant - Ensure model is up to date	- Local, Example-based, Counter-factual, Global Explanations - Example-based and counter-factual had higher perceived understandability	Interpretable Local Surrogates, Occlusion Analysis, Integrated Gradients and Layerwise Relevance Propagation
(Barda et al. 2020)	- Verify model information - Assess model credibility - Include domain knowledge in model	- Global or instance level explanation - Type of techniques, Input, Output, Certainty, Why, Why not, What if, When - Reduce information processing effort	- Unit of explanation - organization of the units - Dimensionality, and size of expl. - Manner data is represented
(Bienefeld et al. 2023)	- Should rely on domain knowledge - Ability to explore knowledge, exploit existing knowledge	- What-if and Example-based are usefull	- shapely values were not helpful, clinicians are not mathematicians
(Aiosa et al. 2023)	XAI helps in evaluating reliable output, improving trust, revealing new insights, finding possible weaknesses, tuning	Global Explanation / Input	SHAP feature importance - Display needs ways to evaluate if output is reliable
(Jia et al. 2022)	Explainability and safety: performance, explainability, robustness, human UX, Data management, other safety controls	- Approximation - Example - Feature Relevance - Visual Explanation	-LIME, SHAP, Deeplift, Layer-wise Relevance Propagation (LRP) - Counterfactual can give clinicians the option to think about output
(Giuste et al. 2023)	XAI can improve model performance, instil trust, and provide the value needed to affect user decision-making	Perturbation, Activation, Gradient, Mixed, Attention	Consistency and interpretation are crucial for a good XAI output

(Du et al. 2022)	Edge between self-reliance and over-reliance on the CDSS	- Explanation by feature importance - Explanation by example	Clinician predict a GDM risk, the model displays similar patient and important features, Clinician prefered explanation by feature
(Di Martino and Delmastro 2023)	Unfamiliarity with ML features, lack of contextual information, need for cohort-level evidence	- Local and Global methods - Techniques, quality and frameworks differ because different data, models and explanations - Model-agnostic explanations	- Lime, SHAP - Clinical validation basic requirement - Similar predictions for sim. Data - Explanation based on user goals and expected reasoning

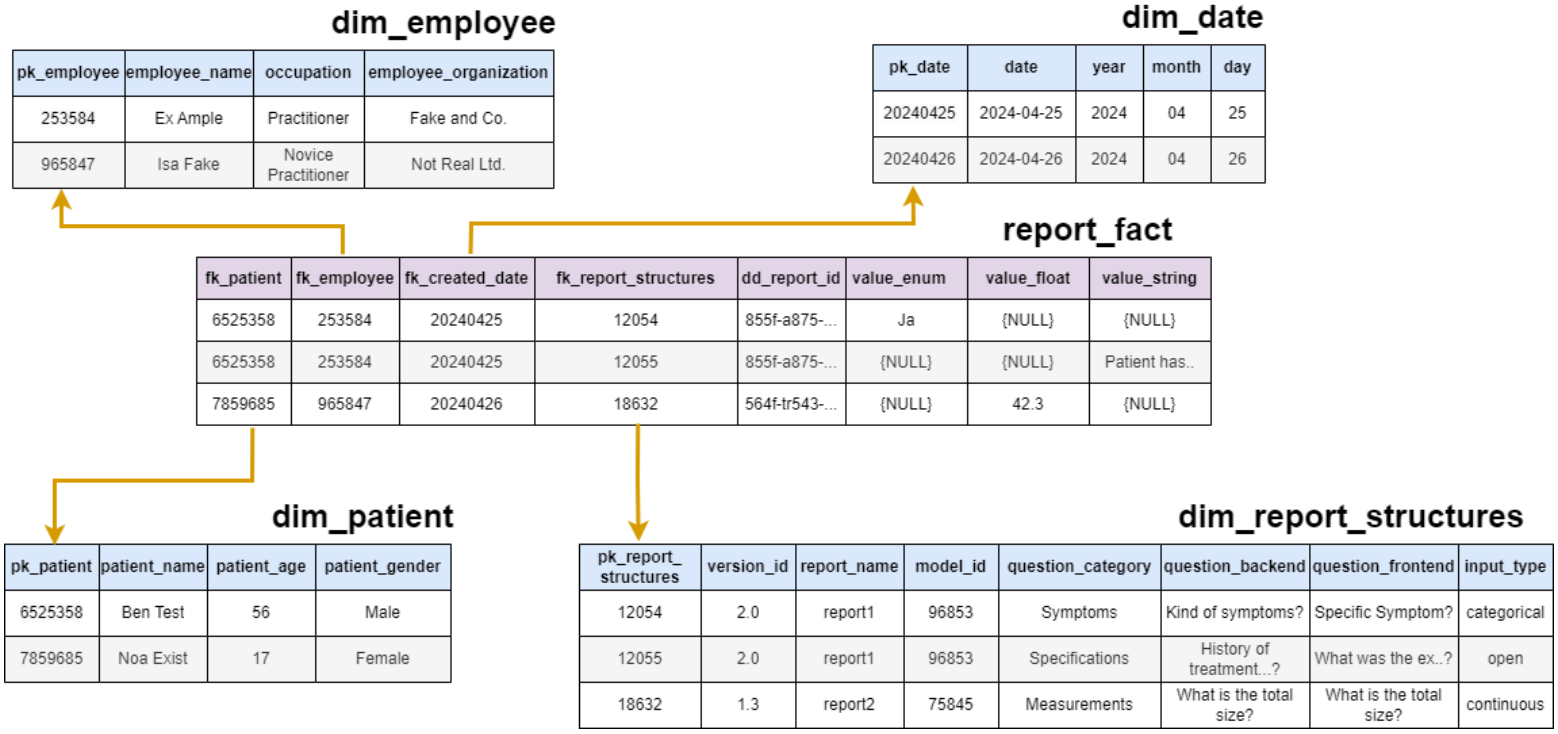
B.3. Findings Literature Regarding Implementation and Impact of CDSS

Table B.3: Summary implementation process and impact of cdss

Author	Implementation Process	Improve adoption rate	Impact on Triple Aim
(Bartels et al. 2022)	- Quality measures by manufacturer - User responsibility in the process - Automated testing and manageable code	- Helpdesk/Central Point for ML malfunctions - Instructions and guidance documentation - Training of end-users	Only a small fraction of ML has been implemented in CDS
(Choudhury 2022)	- Fear of being replaced, myths, reliability, resilience, inexplicability of AI, unfamiliarity with the technology	The Perception of AI and the expectancy of AI all had negative effects of the amount of risks practitioners saw in using AI models in practice - Reduce Risk concerns	- Practitioners mostly saw that AI could improve consistency - practitioners are very likely to prioritize the risk factors
(Blanes-Selva et al. 2023)	- Low adopt. Social factors and usability - Multidisciplinary team, especially for model and UX design - Think out loud method	-Loss of autonomy / Feeling of being replaced - Low computer literacy/lack of trust in AI - Don't see Need for system - Missallignment design and the need	- Increased patient safety by reducing advice against protocol - Diagnosis support/ workflow improvement - Improved service quality due better protocol
(Calisto et al. 2022)	- HAII - Check mental and physical demand	Ability to reject and accept gave positive impact and a higher feeling of control	- Diagnostic accuracy - Time performance - Ability
(Jia et al. 2022)	- Development: Data management, ML Algorithm, Model Learning, Model Comparison	- Local Feature importance and counterfactual explanations to explain - Safety Argument	Trade-off between performance and explainability is crucial in healthcare process
(Lambert et al. 2023)	- Training and inclusion of medical professionals is very important	Fear of loss of autonomy and difficulties integrating AI into clinical workflows were unanimously reported to be hindering factors	- Clinician satisfaction improvements - More information - Better practical training

(Mahadevaiah et al. 2020)	Selection, Acceptance testing, commissioning, implementation, quality assurance	- quality assurance (safe and effective CDSS) - rigorous selection process - acceptance testing - Ensure CDSS is maintained and problems solved quickly	Improve patient safety
(Nuutinen and Leskelä 2023)	Determine based on information and the people involved what the CDSS and UI should look like	Perception and Expactancy of AI should be handled for the end user (clinicians)	AI can improve consistency
(Magrabi et al. 2019)	Determine if it should look at all types of patients or specific types (discriminatory features)	- Look into data quality issues or model performance. - Define monitoring goals and identify key stakeholders	Helps evaluate the medical process
(Iqbal et al. 2023)	transparent communication with patients and comprehensive evaluation of AI's implementation is crucial	- Transparency in the model - privacy protection - Look at stakeholder interests	- Increase patient safety - Improve service quality - Reduce workflow against protocol
(Rundo et al. 2020)	- Physician-centered approach because some have difficulties using technology - think aloud method, focus groups, walkthrough with end-users (near-live)	-Task, User, Representation, Function (TURF) Framework - User-centred design	- Abundance of information could overwhelm capabilities clinicians during daily decision ma. - Reducing repetitive tasks - Improve various aspects of work
(Antoniadi et al. 2022)	Use XAI to validate the model during production, testing and development	- Safety, Fairness, Usability - Keep clinicians part of the process	patient safety, costs, prescription practices, preventive and optimal treatment, recommended standards

Appendix C: Example Starschema with data



dim_employee

pk_employee	employee_name	occupation	employee_organization
253584	Ex Ample	Practitioner	Fake and Co.
965847	Isa Fake	Novice Practitioner	Not Real Ltd.

dim_date

pk_date	date	year	month	day
20240425	2024-04-25	2024	04	25
20240426	2024-04-26	2024	04	26

report_fact

fk_patient	fk_employee	fk_created_date	fk_report_structures	dd_report_id	value_enum	value_float	value_string
6525358	253584	20240425	12054	855f-a875-...	Ja	{NULL}	{NULL}
6525358	253584	20240425	12055	855f-a875-...	{NULL}	{NULL}	Patient has..
7859685	965847	20240426	18632	564f-tr543-...	{NULL}	42.3	{NULL}

dim_patient

pk_patient	patient_name	patient_age	patient_gender
6525358	Ben Test	56	Male
7859685	Noa Exist	17	Female

dim_report_structures

pk_report_structures	version_id	report_name	model_id	question_category	question_backend	question_frontend	input_type
12054	2.0	report1	96853	Symptoms	Kind of symptoms?	Specific Symptom?	categorical
12055	2.0	report1	96853	Specifications	History of treatment...?	What was the ex..?	open
18632	1.3	report2	75845	Measurements	What is the total size?	What is the total size?	continuous

Figure C.1: Star schema after modelling with example data

APPENDIX D: PROCESS DATA EXTRACTION

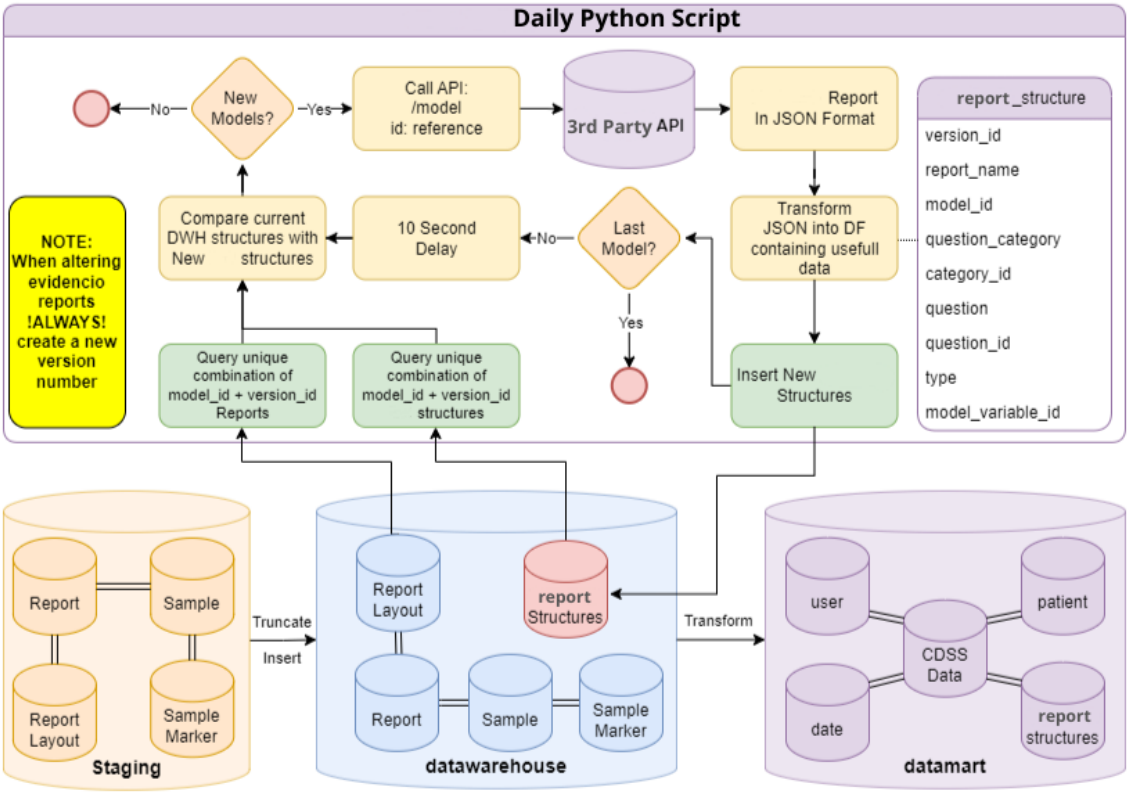


Figure D.1: The process for retrieving the data structures

APPENDIX E: FEATURE ENGINEERING

Total Features Before Selection	Total Features After Selection	Total Features After Selection And Feature Engineering
1153	299	54

When does pain occur

Old Features		New Features
After Rest	-->	After Rest
In Rest	-->	In Rest
During ADL-Activities	-->	During ADL-Activities
After ADL-Activities	-->	After ADL-Activities
During Warming-Up	-->	During Warming-Up
During Sporting	-->	During Sporting
After Sporting	-->	After Sporting
During Work/School	-->	During Work/School
After Work/School	-->	After Work/School
End of the Day	-->	End of the Day
At Night	-->	At Night
11	Per location	11
66	Totals	11*

Exact location of complaints

Old Features		New Features
Subtype-1-Exact-Location-1	-->	Exact-Location-1
Subtype-2-Exact-Location-1	-->	Exact-Location-1
Subtype-1-Exact-Location-2	-->	Exact-Location-2
Subtype-2-Exact-Location-2	-->	Exact-Location-2
Subtype-1-Exact-Location-3	-->	Exact-Location-3
Subtype-2-Exact-Location-3	-->	Exact-Location-3
Subtype-1-Exact-Location-4	-->	Exact-Location-4
Subtype-2-Exact-Location-4	-->	Exact-Location-4
Subtype-1-Exact-Location-5	-->	Exact-Location-5
Subtype-2-Exact-Location-5	-->	Exact-Location-5
Subtype-1-Exact-Location-6	-->	Exact-Location-6
Subtype-2-Exact-Location-6	-->	Exact-Location-6
12	Per location	8
32	Totals	8*

Symptoms

Old Feature		New Feature
Symptom-1	-->	Symptom-1
Symptom-2	-->	Symptom-2
Symptom-3	-->	Symptom-3
Symptom-4	-->	Symptom-4
Symptom-5	-->	Symptom-5
Symptom-6	-->	Symptom-6
Symptom-7	-->	Symptom-7
Symptom-8	-->	Symptom-8
Symptom-9	-->	Symptom-9
Symptom-10	-->	Symptom-10
10	Per location	10
60	Totals	10*

Other variables

Old Feature		New Feature
Trauma	-->	Trauma (0-1)
Recurrence	-->	Recurrence (0-1)
NRS	-->	NRS (0-10)
Course of Complaints	-->	Course of Complaints (0-4)
Gear Specific	-->	Gear Specific (0-4
Walking Activities	-->	Walking-Activities (0-4)
Daily Activities	-->	Daily Activities (0-4)
Intensive Activities	-->	Intensive Activities (0-4)
8	Per location	8
33	Totals	8*

Inflammatory symptoms

Old Feature		New Feature
Inflammatory-Symptoms-1	-->	Inflammatory-Symptoms-1
Inflammatory-Symptoms-2	-->	Inflammatory-Symptoms-2
Inflammatory-Symptoms-3	-->	Inflammatory-Symptoms-3
Inflammatory-Symptoms-4	-->	Inflammatory-Symptoms-4
Inflammatory-Symptoms-5	-->	Inflammatory-Symptoms-5
Inflammatory-Symptoms-NA	-->	Inflammatory-Symptoms-Total
5	Per location	6
30	Totals	6*

Duration of existence of the complaints

Old Feature		New Feature
<1 Week	-->	<2 Months (1)
<1 Month	-->	<2 Months (1)
<2 Months	-->	<2 Months (1)
2-6 Months	-->	2-12 Months (2)
>6 Months	-->	2-12 Months (2)
>1 Year	-->	>1 Year (3)
Unknown	-->	Unknown (0)
7	Per location	4
42	Totals	1*

Side of the complaints

Old Feature		New Feature
Subtype-1 Side of Complaints	-->	Side of Complaints (1)
Subtype-2 Side of Complaints	-->	Side of Complaints (1)
Subtype-3 Side of Complaints	-->	Side of Complaints (1)
Subtype-4 Side of Complaints	-->	Side of Complaints (1)
Subtype-5 Side of Complaints	-->	Side of Complaints (2)
5	Per location	2
30	Totals	1*

Location of the symptoms

Old Feature		New Feature
Diagnosis-Location-1	-->	Diagnosis-Location-1
Diagnosis-Location-2	-->	Diagnosis-Location-2
Diagnosis-Location-3	-->	Diagnosis-Location-3
Diagnosis-Location-4	-->	Diagnosis-Location-4
Diagnosis-Location-5	-->	Diagnosis-Location-5
Diagnosis-Location-6	-->	Diagnosis-Location-6
1	Per location	1
6	Totals	6

*For these totals, the answers per pain location have been combined.

Figure E.1: Selection of features that were used and altered to be usable for ML tasks

APPENDIX F: DESCRIPTIVE STATISTICS OF THE DATA AFTER CLEANING

Feature	Count	Mean	Std	Min	Max
At Night	1040	0.028	0.165	0	1
Diagnosis-Location-5	1040	0.338	0.473	0	1
Exact-Location-1	1040	0.012	0.107	0	1
Existence Period	1040	1.858	1.068	0	3
Symptom-5	1040	0.162	0.368	0	1
Symptom-8	1040	0.007	0.082	0	1
Daily Activities	1040	1.818	0.618	1	4
Diagnosis-Location-2	1040	0.138	0.345	0	1
Symptom-3	1040	0.076	0.265	0	1
Exact-Location-5	1040	0.163	0.370	0	1
End of Day	1040	0.067	0.251	0	1
Diagnosis-Location-6	1040	0.073	0.260	0	1
In Rest	1040	0.125	0.331	0	1
intensive Activities	1040	2.066	0.621	1	4
Exact-Location-11	1040	0.024	0.153	0	1
Exact-Location-9	1040	0.043	0.204	0	1
Exact-Location-10	1040	0.001	0.031	0	1
Gear Specific	1040	0.612	0.488	0	1
Symptom-9	1040	0.021	0.144	0	1
Exact-Location-4	1040	0.110	0.313	0	1
Exact-Location-8	1040	0.006	0.076	0	1
During Walking	1040	1.796	0.605	1	4
Exact-Location-3	1040	0.258	0.438	0	1
Exact-Location-7	1040	0.009	0.093	0	1
Diagnosis-Location-4	1040	0.139	0.347	0	1
After ADL Activities	1040	0.328	0.470	0	1
After Sporting	1040	0.100	0.300	0	1
After Work/School	1040	0.111	0.314	0	1
Diagnosis-Location-1	1040	0.012	0.107	0	1
NRS	1040	6.165	1.625	1	10

Continued on next page

Table F.1 – continued from previous page

Feature	Count	Mean	Std	Min	Max
Inflammatory-Symptom-4	1040	0.072	0.259	0	1
Exact-Location-6	1040	0.595	0.491	0	1
Exact-Location-2	1040	0.104	0.305	0	1
Inflammatory-Symptom-1	1040	0.131	0.337	0	1
Symptom-1	1040	0.481	0.500	0	1
After Rest	1040	0.285	0.451	0	1
Symptom-10	1040	0.032	0.175	0	1
During ADL Activities	1040	0.661	0.474	0	1
During Warming-Up	1040	0.030	0.170	0	1
During Work/School	1040	0.150	0.357	0	1
Symptom-4	1040	0.093	0.291	0	1
trauma	1040	0.040	0.197	0	1
Symptom-2	1040	0.037	0.188	0	1
Course of Complaints	1040	2.386	0.846	1	3
Inflammatory-Symptom-5	1040	0.021	0.144	0	1
Diagnosis-Location-3	1040	0.370	0.483	0	1
Inflammatory-Symptom-2	1040	0.065	0.247	0	1
Symptom_6	1040	0.482	0.500	0	1
Side	1040	1.387	0.487	1	2
Inflammatory-Symptom-3	1040	0.134	0.340	0	1
Inflammatory-Symptom-Total	1040	0.423	0.838	0	5

Table F.1: This table shows various features with their count, mean, standard deviation, minimum, and maximum values.

APPENDIX G: CONFUSION MATRICES

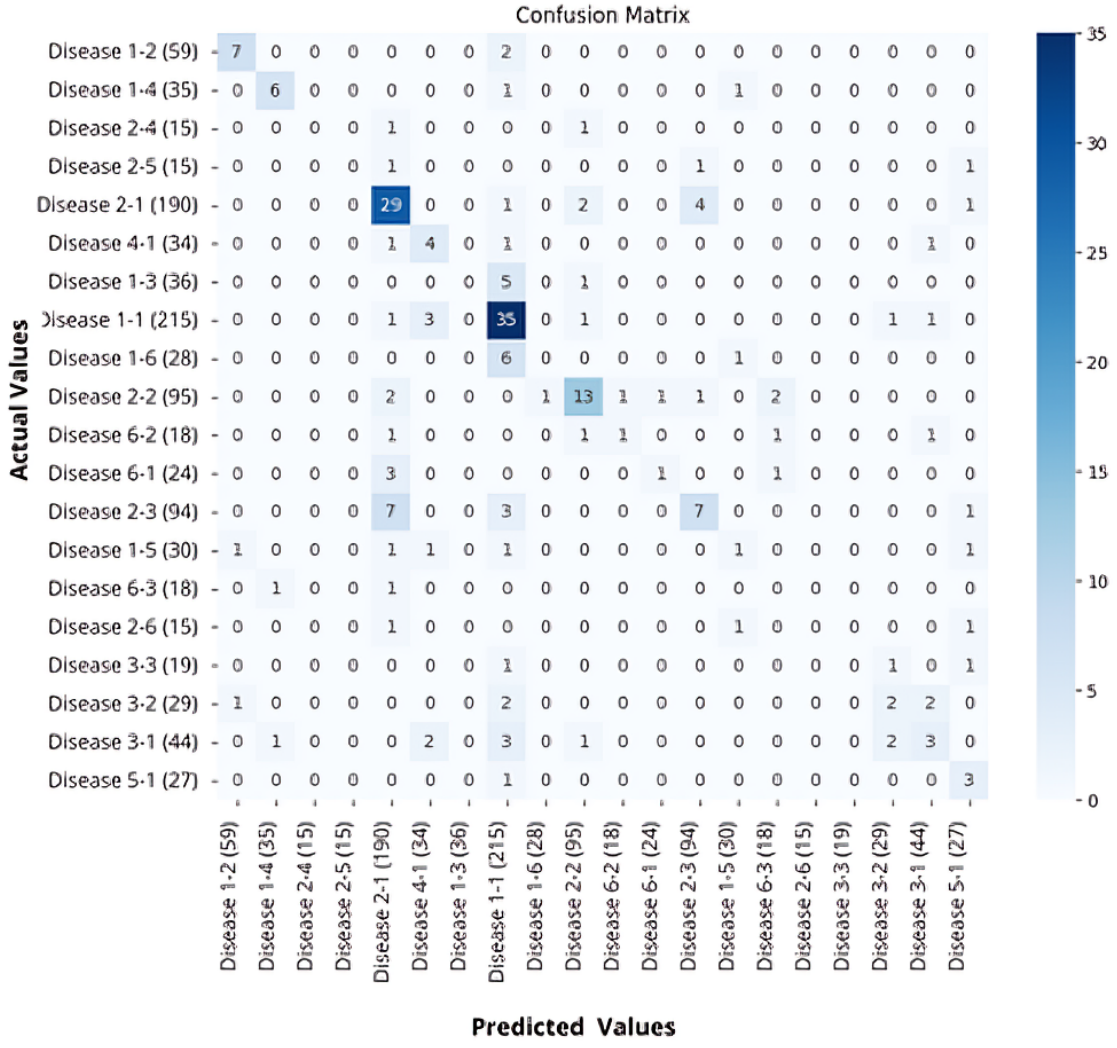


Figure G.1: Confusion Matrix for Model Baseline

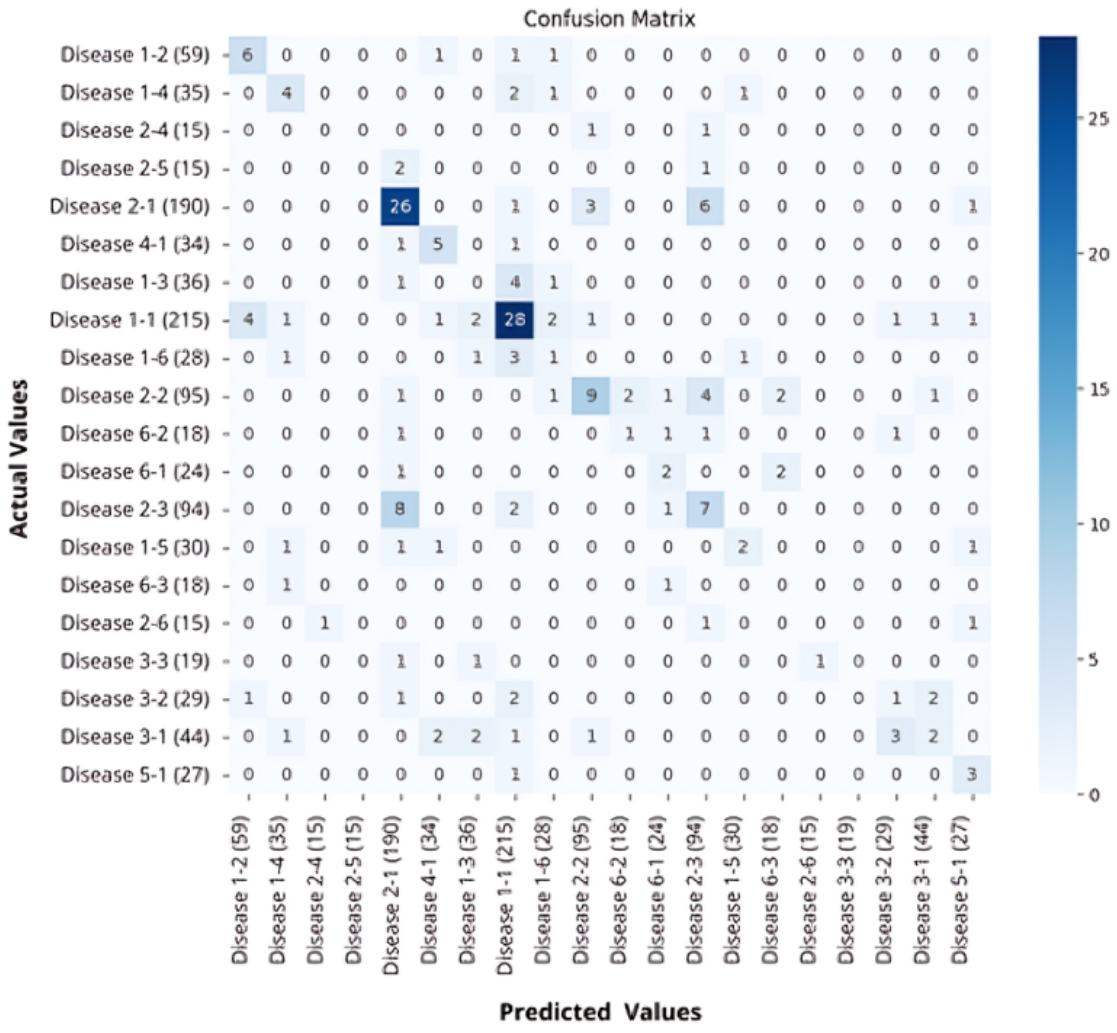


Figure G.2: Confusion Matrix for Model Oversampling

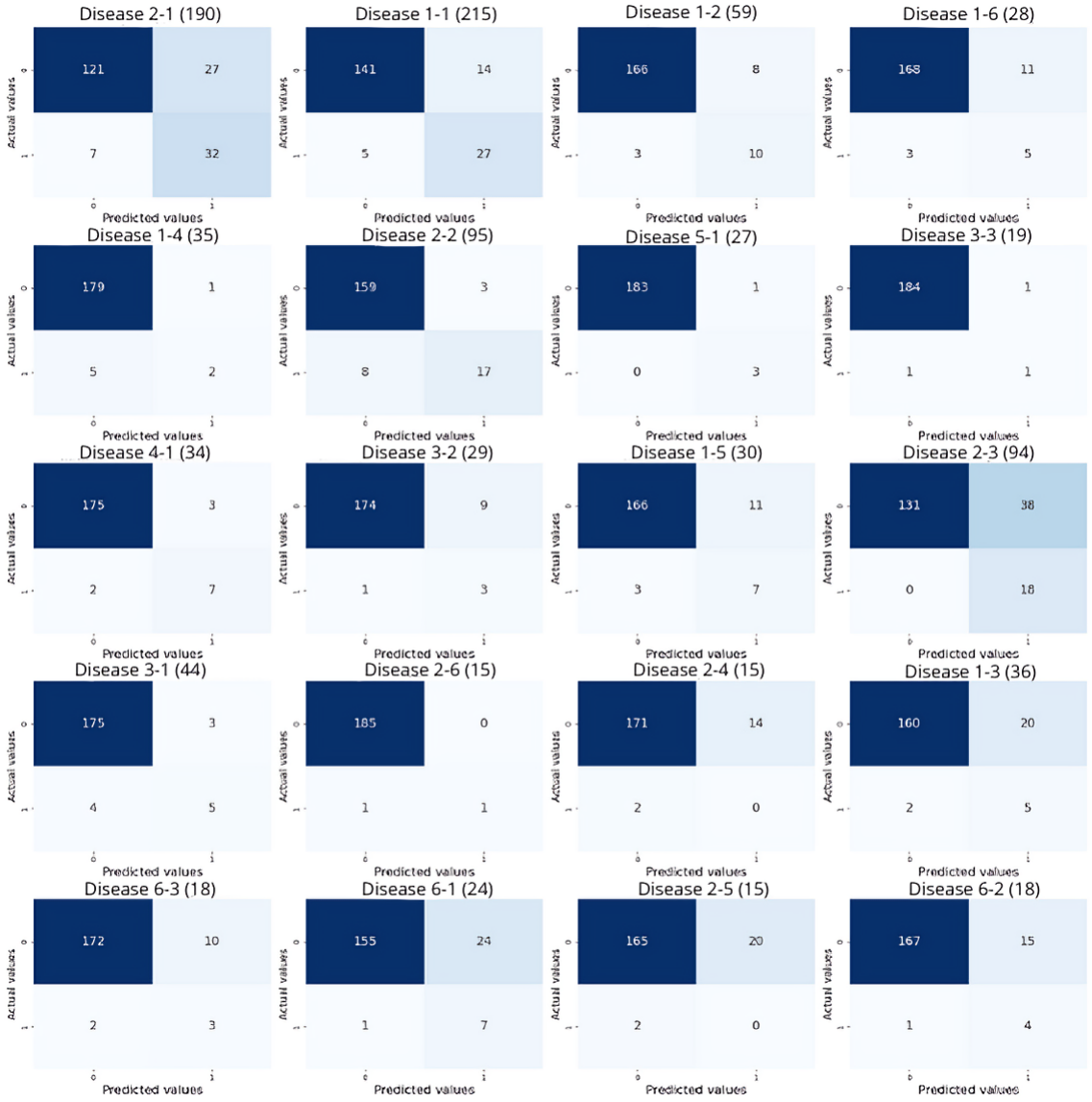


Figure G.3: Confusion Matrix for CICST Model

APPENDIX H: THRESHOLD VALUES FOR THE LABELS IN THE CICST MODEL

disease	threshold
Disease 2-1 (190)	0.26
Disease 1-1 (215)	0.37
Disease 1-2 (59)	0.52
Disease 1-6 (28)	0.19
Disease 1-4 (35)	0.1
Disease 2-2 (95)	0.24
Disease 5-1 (27)	0.24
Disease 3-3 (19)	0.18
Disease 4-1 (34)	0.36
Disease 3-2 (29)	0.23
Disease 1-5 (30)	0.1
Disease 2-3 (94)	0.23
Disease 3-1 (44)	0.14
Disease 2-6 (15)	0.05
Disease 2-4 (15)	0.05
Disease 1-3 (36)	0.34
Disease 6-3 (18)	0.37
Disease 6-1 (24)	0.05
Disease 2-5 (15)	0.05
Disease 6-2 (18)	0.28

Table H.1: Weight factors for the CICST Model

APPENDIX I: EXPLAINABILITY QUESTIONNAIRE ANSWERS

Questions asked during questionnaire:

- 0. Rank the models from best to worst
- 1-4a. Do you understand the information that the model is showing you?
- 1-4b. Does the model give enough information to determine which further examination you should do?
- 1-4c. Does the explanation of the model provide the information to determine if it is a right or wrong diagnosis prediction?
- 1-4d. Do you have any tips or remarks for this model?
- 5. Rank the models again from best to worst
- 6. Do you have any final remarks? Or would you prefer combinations of these models?

Id	Model Ranking Before Explanation	Underst. 1	Info. Sufficiency 1	Reliability Eval. 1	Underst. 2	Info. Sufficiency 2	Reliability Eval. 2	Underst. 3	Info. Sufficiency 3	Reliability Eval. 3	Underst. 4	Info. Sufficiency 4	Reliability Eval. 4	Model Ranking after Explanation
4	Model A;Model C;Model D;Model B;	Eens	Eens	Eens	Eens	Neutraal	Neutraal	Eens	Eens	Eens	Eens	Eens	Eens	Model A;Model C;Model D;Model B;
5	Model B;Model A;Model D;Model C;	Eens	Eens	Neutraal	Eens	Eens	Eens	Eens	Eens	Eens	Eens	Neutraal	Eens	Model A;Model B;Model C;Model D;
6	Model B;Model D;Model A;Model C;	Zeer mee eens	Eens	Eens	Zeer mee eens	Eens	Eens	Zeer mee eens	Eens	Eens	Zeer mee eens	Eens	Eens	Model B;Model D;Model A;Model C;
7	Model A;Model D;Model C;Model B;	Zeer mee eens	Zeer mee eens	Zeer mee eens	Eens	Neutraal	Eens	Zeer mee eens	Eens	Zeer mee eens	Eens	Zeer mee eens	Eens	Model A;Model D;Model C;Model B;
8	Model A;Model C;Model B;Model D;	Zeer mee eens	Eens	Zeer mee eens	Neutraal	Eens	Eens	Zeer mee eens	Eens	Zeer mee eens	Eens	Eens	Neutraal	Model A;Model C;Model B;Model D;
9	Model A;Model C;Model B;Model D;	Zeer mee eens	Eens	Eens	Zeer mee eens	Zeer mee eens	Zeer mee eens	Zeer mee eens	Neutraal	Eens	Zeer mee eens	Zeer mee eens	Zeer mee eens	Model A;Model D;Model C;Model B;
10	Model B;Model A;Model D;Model C	Eens	Eens	Eens	Zeer mee eens	Eens	Eens	Eens	Eens	Eens	Neutraal	Neutraal	Neutraal	Model B;Model A;Model C;Model D
11	Model A;Model C;Model D;Model B	Zeer mee eens	Eens	Eens	Oneens	Neutraal	Eens	Zeer mee eens	Eens	Eens	Eens	Eens	Eens	Model C;Model D;Model A;Model B
12	Model A;Model D;Model B;Model C	Zeer mee eens	Neutraal	Eens	Oneens	Neutraal	Oneens	Eens	Oneens	Oneens	Eens	Neutraal	Oneens	Model A;Model D;Model B;Model C
13	Model D;Model B;Model A;Model C	Eens	Oneens	Oneens	Eens	Eens	Eens	Eens	Eens	Eens	Oneens	Oneens	Oneens	Model C;Model B;Model A;Model D
14	Model B;Model C;Model A;Model D	Eens	Neutraal	Neutraal	Eens	Eens	Eens	Zeer mee eens	Neutraal	Neutraal	Eens	Oneens	Oneens	Model A;Model C;Model D;Model B;
15	Model C;Model A;Model B;Model D	Zeer mee eens	Eens	Neutraal	Zeer mee eens	Eens	Eens	Zeer mee eens	Eens	Eens	Zeer mee eens	Eens	Neutraal	Model C;Model A;Model B;Model D
16	Model A;Model D;Model B;Model C	Zeer mee eens	Neutraal	Eens	Zeer mee eens	Neutraal	Eens	Zeer mee eens	Neutraal	Neutraal	Zeer mee eens	Neutraal	Neutraal	Model B;Model C;Model A;Model D

Figure I.1: Filled in answers of the questionnaire

APPENDIX J: PREVIEW IMPLEMENTATION

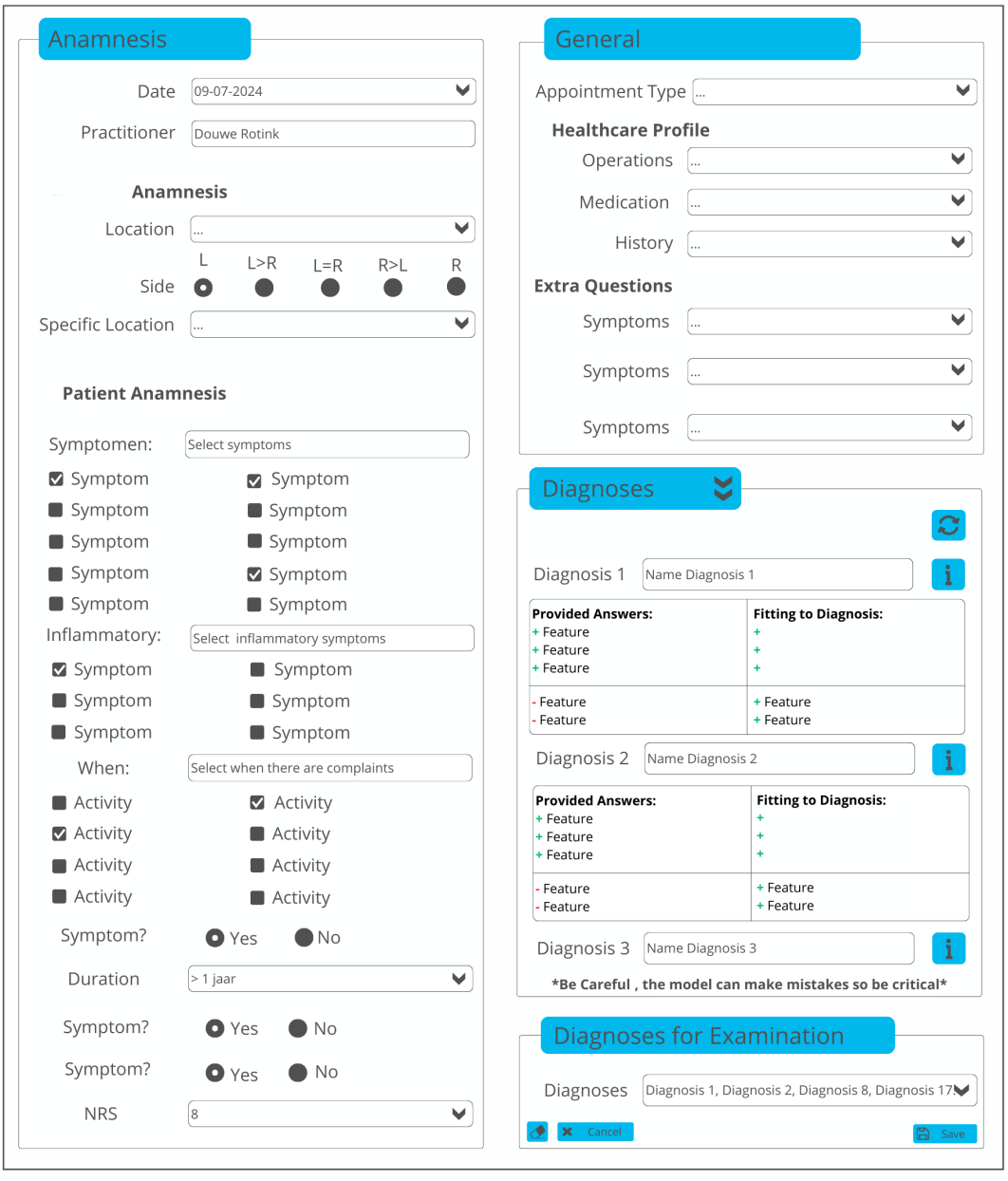


Figure J.1: Screenshots of what the final implementation could look like