UiA University of Agder

Master

Human-Centered Design (HCD) of Personal Decision Support System (PDSS) for understandable individual health management, using Natural Language Processing (NLP) and machine learning

MARTA HANCZAREK
VETLE ØRSTAVIK HOLLUND
FERDINAND GUMØ LØBERG

SUPERVISOR
Martin Wulf Gerdes

University of Agder, 2022
Faculty of Engineering and Science
Department of Engineering and Sciences

Obligatorisk gruppeerklæring

Den enkelte student er selv ansvarlig for å sette seg inn i hva som er lovlige hjelpemidler, retningslinjer for bruk av disse og regler om kildebruk. Erklæringen skal bevisstgjøre studentene på deres ansvar og hvilke konsekvenser fusk kan medføre. Manglende erklæring fritar ikke studentene fra sitt ansvar.

1.	Vi erklærer herved at vår besvarelse er vårt eget arbeid, og at vi ikke har brukt andre kilder eller har mottatt annen hjelp enn det som er nevnt i besvarelsen.	Ja
2.	**Vi erklærer videre at denne besvarelsen:** • Ikke har vært brukt til annen eksamen ved annen avdeling/universitet/høgskole innenlands eller utenlands. • Ikke refererer til andres arbeid uten at det er oppgitt. • Ikke refererer til eget tidligere arbeid uten at det er oppgitt. • Har alle referansene oppgitt i litteraturlisten. • Ikke er en kopi, duplikat eller avskrift av andres arbeid eller besvarelse.	Ja
3.	Vi er kjent med at brudd på ovennevnte er å betrakte som fusk og kan medføre annullering av eksamen og utestengelse fra universiteter og høgskoler i Norge, jf. Universitets- og høgskoleloven §§4-7 og 4-8 og Forskrift om eksamen §§ 31.	Ja
4.	Vi er kjent med at alle innleverte oppgaver kan bli plagiatkontrollert.	Ja
5.	Vi er kjent med at Universitetet i Agder vil behandle alle saker hvor det forligger mistanke om fusk etter høgskolens retningslinjer for behandling av saker om fusk.	Ja
6.	Vi har satt oss inn i regler og retningslinjer i bruk av kilder og referanser på biblioteket sine nettsider.	Ja
7.	Vi har i flertall blitt enige om at innsatsen innad i gruppen er merkbart forskjellig og ønsker dermed å vurderes individuelt. Ordinært vurderes alle deltakere i prosjektet samlet.	Nei

Publiseringsavtale

Fullmakt til elektronisk publisering av oppgaven Forfatter(ne) har opphavsrett til oppgaven. Det betyr blant annet enerett til å gjøre verket tilgjengelig for allmennheten (Åndsverkloven. §2).
Oppgaver som er unntatt offentlighet eller taushetsbelagt/konfidensiell vil ikke bli publisert.

Vi gir herved Universitetet i Agder en vederlagsfri rett til å gjøre oppgaven tilgjengelig for elektronisk publisering:	Ja
Er oppgaven båndlagt (konfidensiell)?	Nei
Er oppgaven unntatt offentlighet?	Nei

Abstract

This project addressed the challenges of patients to comprehend typical information (e.g. in the form of ”doctor letters” or ”patient journals”) about their health condition, and to get understandable and personalized recommendations in a user friendly way to improve their health and well-being, considering their individual and actual health status. The chosen approach has been to explore the possibilities of natural language processing technologies and use of machine learning (particularly random forests), integrated in a human centered design of a personal decision support system. The project followed a combination of Design Science Research Methodology and Human Centered Design Methodology. For the initial user needs assessment, 12 interviews with potential users of the target application were carried out, representing varying degrees of experience with the Norwegian health care system. The interviews showed a large gap in the comprehension of the information contained in personal medical journals, and resulted in the problem specification, needs assessment, and a PACT analysis. Further a series of requirements were produced with the use of Volere shells.

The project was divided into three main parts, Natural Language Processing (for the data extraction and content summarization), personal health suggestions (with use of machine learning and feature importance determination), and front end design (for an integrated target application).

The language processing was largely based on pre-trained transformer based models tuned for downstream tasks. Several models trained with biomedical and clinical text were evaluated on automated summarization, semantic search and assertion detection, and showed promising results. Areas of improvement and future work were summarization of short texts and formal evaluation of medical term explanations.

The front end design went trough three iterations, a low fidelity prototype in the for of a paper prototype, and two versions of the high fidelity prototype. The user testing showed an improvement in understanding of medical data in both high fidelity prototypes.

The health suggestion were based on the feature importance determination of random forests. Three different determination methods were tested, finding minor variations in results, but with Gini Importance gaining a major advantage in computational speed. The recommendation produced could not be tied to clinical results but would require further study to prove or disprove the effectiveness of the recommendations.

Acknowledgements

We would like to start the thesis off by thanking our supervisor, Martin Wulf Gerdes for the support, guidance and feedback trough out the project. We would also like to thank all the participants of out interviews and usability tests, that took their time and aswered our questions and provided us with valuable data. It would not be possible for us to complete the project without their help. Finally we would like to thank both friends and family for supporting us though many years of education.

University of Agder, 2022
Marta Hanczarek, Vetle Hollund, Ferdinand Løberg

Table of Contents

List of Figures

List of Tables

This page is intentionally left blank.

Abbreviations

BMI	Body Mass Index
CAEC	Consumption of food between meals
CALC	Consumption of alcohol
CH2O	Consumption of water daily
DSRM	Design Science Research Methodology
FHWO	Family history with overweight
FAF	Physical activity frequency
FAVC	Frequent consumption of high caloric food
FCVC	Frequency of consumption of vegetables
HCD	Human-Centered Design
MTRANS	Transportation used
NCP	Number of main meals
NLP	Natural Language Processing
NSD	Norwegian Centre for Research Data
PIMP	permutation based feature importance
SCC	Calories consumption monitoring
TUE	Time using technology devices
UI	User Interface
UX	User Experience
W3C	World Wide Web Consortium

Chapter 1

Introduction

With modern health services diving deeper into the world of technology, the lack of research data is not the major problem it once was. Medical data has become more accessible with the use of patient portals and available in large quantities through sites like kaggle.com. In the initial interview of this project (appendix H), it was found that it was rather the understanding of the data that was a problem, not necessarily the access. Usually, you would consult a health professional to ask for an understandable explanation of your health condition. However, with time becoming a scarce resource for doctors and other health experts to spend with each patient [10], this is not a simple solution.

1.1 Problem and Significance

With multiple different systems to manage personal health information, there is a significant variation in usability. The effectiveness in terms of increasing the patients' understanding of their condition is a volatile variable. From test sets provided for the master's project, there are multiple examples of the use of abbreviations and medical terms (mostly Latin words), with each of the fore-mentioned elements obfuscating the meaning to a layperson within health care.

In order to get a better understanding of the problem space, interviews of patients and health care professionals were conducted. The only requirement of the interview subjects was having previous experience with the specialized Norwegian health care system. Some subjects reported that they received a lot of information that was too dense to easily parse. It was also found that most of the interview subjects, especially on the patient side, found medical terms and abbreviations difficult to understand. They had to either ask their doctor or look them up in a dictionary or with the use of a search engine.

We asked whether our interview subjects felt they got enough information about their condition from their doctor. Several reported that they either did not get enough information in general, or that they wanted more information about something specific.

When asked about what kind of advice or recommendation they would like to receive to improve their health condition given a certain diagnosis, most subjects responded that they wanted specific and applicable advice with some explanation as to why. Medical professionals generally wanted to know more of the underlying reasoning, whereas laypeople wanted a greater focus on what to do and what not to do.

Computer literacy may also impact the understanding of medical data[34], as such, there is a need for a platform that will help users understand their personal medical information with relatively low amounts of computer literacy.

As some countries still do not use patient portals to provide health information to patients, not all users have access to a digital version of their medical journals. This indicates that it would also be necessary to allow users to specify the condition they would require feedback on. This was not a prevalent issue in the interviews, as all interview subjects had access to their medical journal in a digital format.

1.2 Research Questions

This section presents the research questions and sub-questions for the thesis. The questions

RQ1: What issues do patients have when trying to understand medical documents, and are there NLP methods that can alleviate these issues?

RQ1.1: Can current state-of-the-art Natural Language Processing (NLP) models be used effectively in combination to realize the functional requirements of our system?

RQ2: Can specific and significant correlations between health features and medical diagnosis be found using selected Machine Learning techniques?

RQ2.1: Can we extrapolate causation from correlation provided from a machine learning algorithm?

RQ3: How can Human-Centered Design increase usability of a Personalized Decision Support System?

RQ3.1: How can a Human-Centered Design help increase understanding of medical terms and documents?

RQ3.2: How to illustrate personalized advice and recommendations in a user friendly matter?

1.3 Limitations of Scope and Focus

The scope of the study is limited at different parts. The initial interviews for the needs assessment have been limited to 12 people. Further, the usability test pertaining to each iteration was limited to 5 users.

These limitations can be boiled down to the same base criteria: the limitation of time available vs. the intended scope of aspects to be considered and covered. The project has a hard deadline *June 3. 2022*. With the limit of just under six months to complete the project, we have chosen to limit the number of participants to have a balanced focus on the different parts of the project.

The gathering of medical data is highly regulated, and as such, all data used is anonymized data, most of which was taken from `kaggle.com` or other openly available sources.

The investigation of NLP methods were done with English-language models and data because of the limited availability of Norwegian data and models. Because of this, we limited the use of NLP methods to ones that are likely to be suitable for Norwegian models in the future.

1.4 Report structure

The different stages of Design Science Research Method (DSRM) [11] are tied to chapters in the report as shown in fig. 3.1. The *problem identification and motivation* are described in chapter 1, the introduction of this thesis. The *Objectives of the solution* was found through multiple different steps and as such chapter 3, chapter 2 and chapter 4 is part of this stage. The *design and development* of both the back-end software and the user interface is shown and described in chapter 5, and the data and any relevant preprocessing used for the back-end is described in chapter 6. When it comes to the *demonstration* this can be considered both for the usability tests, where the user interfaces were tested on users, but also with the proposed integrated solution in chapter 5. The Key Performance Indicators (KPI) underlying the evaluation are described in section 3.5. The front-end design was evaluated over three iterations, with the back-end using continuous development and, as such, only shows one iteration as described in chapter 7 and chapter 8.

Chapter 2

Background

This chapter will start with a short overview of existing solutions that can be compared to our solution in different aspects. It will then introduce the reader to the theory for the technologies used in this thesis. That is split into three main parts. There was a natural divide between NLP, Data Analysis (including Random Forests), and the theory of Human-Centered Design focused on usability. The chapter closes with some performance measures used within the different subjects.

2.1 Existing solutions for patient communication

The subject of e-health is a vast and quickly expanding area of research. This leads to multiple solutions for the same problems. This section will mention a few different well-established options for our proposed solution.

2.1.1 WebMD

WebMD [71] offers multiple services to users, including pharmacy options, drug information, storage for personal medical information, and symptom checklists[72]. The symptom checklists attempt to let a patient get a better overview of what conditions they might be suffering from based on the symptoms they are experiencing. With WebMD's symptom checker being a proprietary product, it is not clear how it works. According to PRNewswire, the symptom checker contains existing correlation relationships between different symptoms and conditions and, from the information provided by the user, generates a list of potential conditions[49].

2.1.2 Patient Portals

Patient portals are being offered by most health care providers[30]. A patient portal is used as an interface for electronic health records (EHR), also called patient journals. `helsenorge.no` is a good example of a patient portal. Allowing users to safely view their own health records with varied functionality. Examples of other functionality are renewing prescriptions, booking an appointment with their primary physician, or applying for travel cost reimbursement.

The data provided in the patient portal by the EHR is relatively raw, containing test results and information the physician has produced. The purpose of a patient portal was to increase health knowledge, self-efficacy, and medication adherence, and according to a 2019 review on patient portals from the Johns Hopkins University, it has managed to do so. The same review also showed that the clinical results, such as blood pressure, cholesterol, and weight loss, were not consistent and concluded with a lack of evidence to support the improvement of clinical outcomes [25].

2.1.3 Clinical Decision Support System (CDSS)

The idea of a CDSS is not new, and its use can be found as early as the 1970s. It was used in a varying capacity in EHR as of 2013 [61], with an increasing amount of institutions using them. The functionality of a CDSS is extensive, with alerts, drug management, disease management, and more. The cost of the development and maintenance high, but can results in extensive savings in other areas [61]. CDSS are split between two different architectures, namely, knowledge-based and non-knowledge-based. The non-knowledge-based systems are data-driven applications using machine learning to provide statistical pattern recognition in clinical data. The knowledge-based systems use if-then rules, using known connections between symptoms and conditions to provide decision support [8].

A series of systematic reviews show that multiple studies have displayed a positive impact of CDSS[8]. There are downsides, however, such as the levels of computer literacy might impact the efficacy of CDSS [34].

2.2 Natural Language Processing

Natural Language Processing (NLP) has been described as the "intersection of artificial intelligence and linguistics"[40]. NLP problems are concerned with having computers understand natural language and extract useful information from it. High-level NLP tasks include translation, summarization, and text generation.

Early approaches to these problems relied on handwritten sets of rules, which limited scalability and robustness because of the complex and inconsistent rules of natural language.

As computing power increased, simpler and more robust models based on probabilities came into use. These models use machine learning and large text corpora to learn broad rules, which are used to calculate the most likely meaning of a phrase. These methods produce relatively reliable results, given that the analyzed language does not deviate too radically from the training corpora. [40]

Over the last decade, NLP models based on deep learning with artificial neural networks have quickly gained popularity. This is in large part because of the increase in computing power provided by graphical processing units specifically designed for deep learning. Several of the current state-of-the-art NLP models are based on the Transformer architecture, most notably BERT and its derivatives.[42][67][67]

2.2.1 NLP Concepts and Tasks

Searching is the act of finding specific information in a text corpus. Searching has many levels of complexity, depending on the type of searching one wants to do. One of the simplest methods is exact text matching, where exact instances of a query are found in the target document. This method is computationally efficient and easy to implement. More complex queries can be constructed using search patterns such as regular expressions[64], but they correspond poorly to how humans generally reason about language.

$$\text{run} \rightarrow \text{He grunts when he runs.}$$

A more complex search method may account for aspects of grammar, for example by matching words with different inflections. This can be done by reducing an inflected word to its word stem in a process called stemming. Approaches to stemming are generally rule-based, statistical methods, or a combination of the two [28].

$$\text{thinks} \xrightarrow{think-} \text{They are thinking about it.}$$

Even more complex are search methods based on semantic similarity. These methods attempt to find terms in the target text with similar meanings to a query, regardless of their syntactic similarity. Approaches can be divided into four categories: knowledge-based, corpus-based, deep neural network-based, and hybrid models [14].

pet → I have a dog and a cat.

Some approaches to semantic similarity searching also account for the context in which a word is written, which can be used to disambiguate words with several possible meanings.

go for a run → I went jogging yesterday.
I run a company.

Text summarization is the act of distilling text down to its most important or relevant parts. There are two main approaches to summarization: extractive and abstractive summarization.

In extractive summarization, words and phrases are extracted directly from the source text without changing them. The extracted phrases can be used to automatically generate keywords or key-phrases, or they can be combined to create a summary. The main disadvantage to this form of summarization is that the output may seem unnatural, as no new text is generated to connect the extracted phrases.

In abstractive summarization, an abstract representation of the most important concepts and their meaning is created. Text is then generated from these concepts, which results in a more cohesive output than extractive summarization. The main disadvantages of this approach is that it requires more computational power to perform, and a model with a deep understanding of the domain the source text discusses. This is especially challenging in fields with little access to training data.[32]

Named entity recognition is the problem of identifying and categorizing named entities. This is done by finding words or phrases in a text that refers to a single instance of a category. For example, given a category "school" and the sentence "I attend the University of Agder", one would consider "University of Agder" to be an instance of a school. Although it is a type of school, the word "University" would not be considered a named entity as it can refer to multiple schools.[39]

Word embeddings are numerical representations of words and their meaning, typically in the form of vectors. These are often easier to work with than raw text as they can be manipulated with mathematical operations, and reduce complexity by omitting unimportant details from natural language. Additionally, several machine learning methods require consistently sized inputs, which word embeddings can facilitate.[31]

2.2.2 Neural Network Concepts

Neurons are interconnected cells in the human nervous system that that can be activated by an some input and send signals to other neurons. They are integral part of the human brain.

Artificial Neural Networks (ANN) were inspired by this structure. ANN consist of a large number of interconnected simple nodes called artificial neurons. These neurons can receive inputs and produce outputs in the form of real numbers. All of a neuron's inputs have an associated weight, which can be seen as the strength of the connection between two neurons. When training an ANN, these weights are adjusted to achieve a desired final output. The output of a neuron is calculated by passing the weighted sum of its inputs through an activation function.

ANN are typically structured by organizing neurons into several layers. The first layer, called an input layer, receives inputs from external data. The last layer, called an output layer, outputs the result. The layers between the input and output layers are called hidden layers. The simplest network architecture is one where neurons in a layer are only connected to neurons in the next layer. These are called Feedforward Neural Networks (FNN), as seen in fig. 2.1. [62]

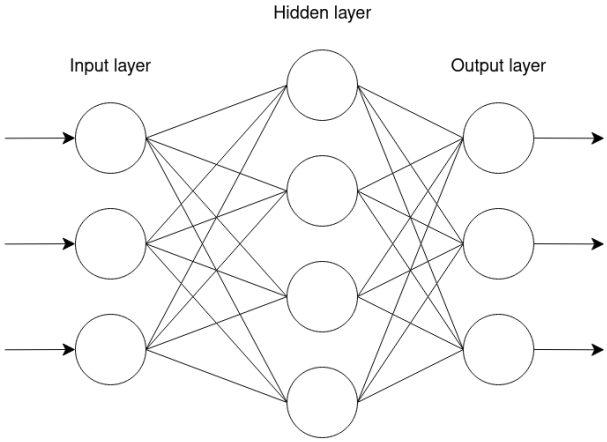


Figure 2.1: Feedforward Neural Network architecture

Recurrent Neural Networks (RNN) use a more complex architecture in which neurons can be connected within a layer or to previous layers. This allows the network to extend over time through cyclical connections.[1]

2.3 Data Analysis

The data processing will be split into two subsections. First is the way the data was processed using the feature importance in random forests. Second is the background of factor analysis, which is commonly used to find the relevance of any given feature.

2.3.1 Random Forest

The random forest is an ensemble algorithm that can be used for both classification and regression, first presented by Tin Kam Ho in 1995[26]. The proposed setup of the random decision forest by Kam Ho included the random component to increase the accuracy of the decision tree. It consists of creating trees in randomly selected subspaces of the feature space.

The base of the random forest remains more or less unchanged, though where variations have occurred is how data was split for training. The addition that was made by Breiman in 2001 was the use of bagging[9]. Using subsets of the entire dataset to create trees, and feature bagging, excluding different features to achieve measurable importance of each feature in the dataset. Breiman also mentioned the use of a random forest to determine feature importance. Allowing it to be used to find the importance of a given feature in classifying a given entry.

2.3.2 Feature importance

The importance of features is a score representing how important a feature is in classifying the data entry and producing a prediction. A higher feature importance score means the feature will have a higher effect when it comes to classification [57].

Using random forests there are multiple ways to calculate feature importance. The first is the built-in feature importance for the random forest; the mean decrease impurity(MDI), also referred to as Gini importance, is calculated from the structure of the forest itself. The Gini importance is calculated on the decrease of the impurity of a split. This can be done using gain and entropy. The entropy can be calculated with the formula $E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$. It can further be expanded with multiple variables to the expression: $E(X,T) = \sum_{c \in x} P(c)E(c)$. The gain is then calculated by using the entropy functions: $Gain(x,t) = Entropy(x) - Entropy(x,t)$. The gain/entropy formulas are mainly used to calculate the structure, though they have the side benefit of classifying the importance of any given feature. This is the default importance method used by *sklearn's* random forest[55].

Another method to calculate feature importance is the use of Shapley values from game theory[37]. The use of Shapley value interpretation is model agnostic and can be used to find the feature importance of any machine learning model. An example can explain the concept of Shap values. Imagine a random forest created by training different trees with different subsets of the features. Examining the misclassification rate of any given tree can show the importance of the features that the tree lacks. Concretely if the misclassification rate is low, the features present have an increased importance. In contrast, if the misclassification rate is high, either the missing features have a higher importance or the present features have a decreased importance. Repeat this example a number of times, and compare the results.

Finally, there is the permutation-based feature importance (PIMP) [6]. The method shuffles a given feature in the test set and compares the results of the different permutations. The reduction in the score is indicative of the importance of the given feature. The permutation-based feature importance is a very computationally heavy operation, and as such other methods may be preferred.

2.4 Usability

Usability is an essential part of the field of Human-Centered Design. The International Organization for Standardization describes usability as *"The extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"*[27]. As the definition implies, one of the main priorities of usability is to provide a system, product, or service with a well functioning and elaborate design.

2.4.1 Accessibility

As an aspect of usability, accessibility relates to the user being able to use the system, product, or service. One could argue that for usability to be a challenge at all, the intended user needs to be able to have access to the system, product, or service to begin with. People being left out based on their disabilities has long been a legal and ethical issue [7, p. 107].

When talking about accessibility regarding software technologies, the goal of accessibility is to remove barriers for people with some form of disability and give all users access to the full potential of the system, product, or service. When designing a new system, product, or service, there are some universal principles the designer should try to implement. These principles are:

“

1. *Equitable use.* The design does not disadvantage or stigmatize any group of users.
2. *Flexibility in use.* The design accommodates a wide range of individual preferences and abilities.

3. *Simple, intuitive use.* Use of the design is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level.
4. *Perceptible information.* The design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities
5. *Tolerance for error.* The design minimizes hazards and the adverse consequences of accidental or unintended actions.
6. *Low physical effort.* The design can be used efficiently and comfortably and with a minimum of fatigue.
7. *Size and space for approach and use.* The design can be used efficiently and comfortably and with a minimum of fatigue.

The 7 Principles [63] "

Even though the overall goal of a designer is to include everyone, regardless of disabilities or not, the reality is that total inclusion is unattainable [7, p. 107]. However, this does not change the designer's responsibility regarding implementing accessibility into their system, product, or service.

In order to increase user experience (UX), one should create visually pleasing design solutions. The five visual-design principles shown in fig. 2.2 help to achieve this.

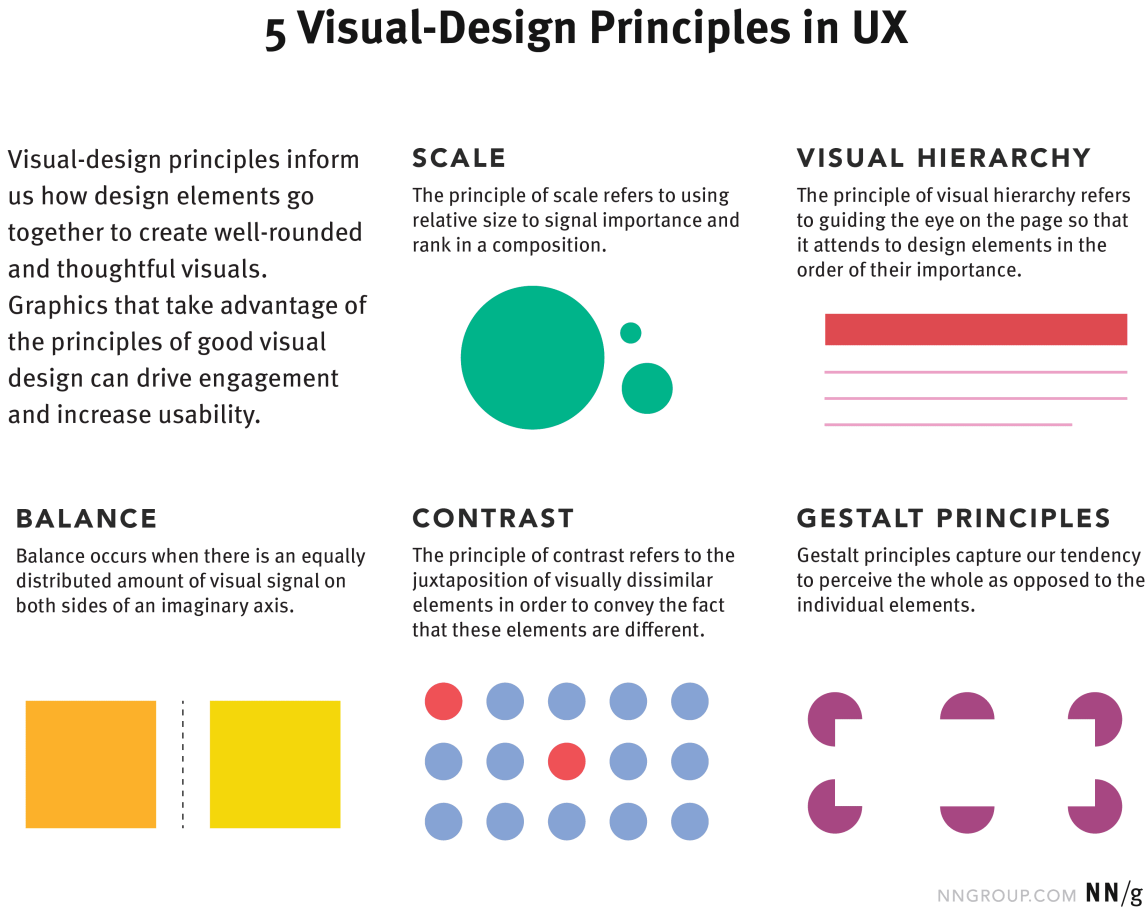


Figure 2.2: 5 Visual-Design Principles in UX, as presented by nngroup.com [24]

Chapter 3

Methodology

Considering the size of the project in addition to the different technologies, a mixed-method was used. The DSR methodology is constructed from six different stages as described by Carstensen et al. [11]. Based on the DSR methodology, a prototype of the project's information system was produced. As a major part of the project was to increase the users' understanding of their own medical data, a Human-Centered Design methodology [27] was used to include user needs and feedback in the iterative design and development of the application.

3.1 Problem Identification & Motivation

The project was based on two different master thesis proposals, though with a larger group of three people, merged into a single combined project. Initially, it was proposed as a clinical decision support system. However, with a time limitation of slightly less than six months, the project proposal was changed to a Personal Decision Support System (PDSS). A series of interviews with health care professionals and patients with a mixed amount of experience with medical journals and the Norwegian health care system were conducted. It was found that there was a need for something to increase understanding and decrease the gap between health care professionals and those not educated in the subject. The feedback from these interviews showed a significant variance in understanding. Since both health care professionals and patients with a long history with health care services had problems understanding medical journals, it was clear that a simplification was needed. The interviews also showed that patients mainly were happy with their primary physician, though they

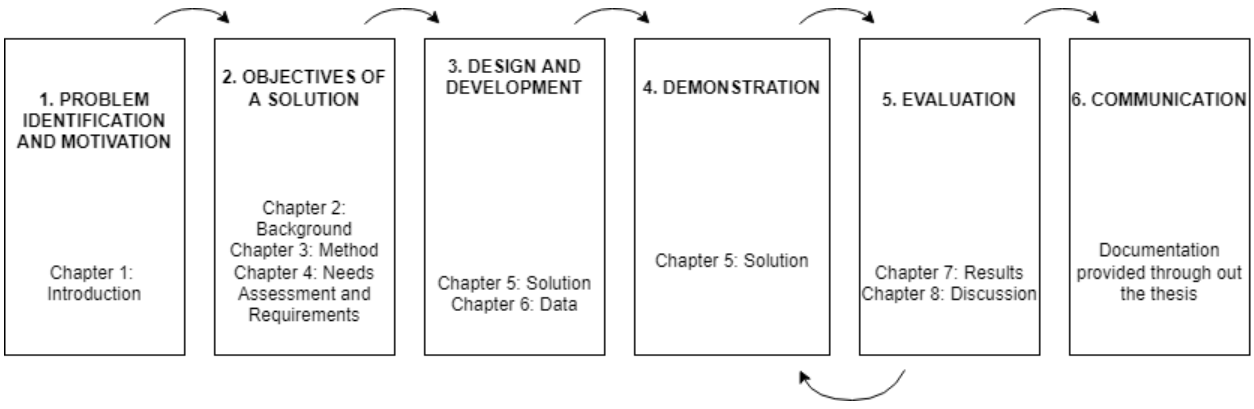


Figure 3.1: Our adaption of the DSR methodology, based on DSRM [11], and inspired by Peffers et al. [45]

would often like a more expanded explanation of their condition and what - if anything - they could do to improve it. Health care professionals showed a wish to do just that, but it was not possible with the available time per patient. The result of the problem identification and motivation stage can be seen in chapter 1.

3.2 Define Objectives of a Solution

The objectives of the solution were initially found using information gathered through qualitative interviews. The data was further compressed and analyzed, finally leading to the Volere requirement shells[54].

3.2.1 Data gathering

The three project authors recruited the participants. Recruitment was made with a broad set of criteria to maximize the user base for the data gathering. There was no upper age limit, with the lower limit being 18 years of age. Further, it was preferred that the interviewee had some prior experience with Norway's medical health care system, either in written form or physically. Lastly, some of the interview subjects were asked additional questions since they were educated in a health care profession. In all, 12 interviews were conducted. The questions asked in the interviews are listed below, and a compressed table of results can be found in appendix H.

- Age Group: (18-25,26-35,36-45,46-55,56-65,66+)
- Education/profession:
- How satisfied are you with the dissemination of information from your health contacts? What's good? What could be better?
- Have you read your own medical journal or other written communication from your health contacts?
- Do you find it easy to understand what is written in you medical journal?

Only asked to medical personnel:

- *Do you read the medical journals of other people?*
- Have you gotten enough information regarding concrete actions you can do to improve your long term health?
- What type of feedback do you prefer? Concrete actions or the theory behind them? Example: "Do 30 minutes of physical activity every day", or "A lower resting heart rate and weight will have a positive effect on patients with diabetes".

The interviews were conducted by first describing how the data would be used and collecting a signed consent form outlining the same information described by the interview holder. It is recognized that sampling bias might occur as there is no random component to the recruitment of participants, though this was considered while creating the interview questions.

The recruitment of participants for the user testing followed the exact requirements as the interviews but had a higher focus on getting a broader demographic. The user testing started with an explanation of the consent form, as well as a short introduction to the project. The exact instructions were used in both of the user tests. In each test, 5 participants tested the application. A more profound explanation of the usability testing can be found in section 5.4.

Privacy Concerns

The information gathering for the project was approved by the Norwegian Centre for Research Data (NSD)[41]. Reference number 542357. Participants of both interviews and user tests were presented with a verbal and written explanation of the consent form. The consent form contained information on how the data would be stored and used and, if necessary, how and whom to contact for removal or edits of the information they provided. All data entered into the application were test data prepared beforehand, unrelated to the user.

3.2.2 Needs assessment and requirement specification

The interviews mentioned in section 3.1 and chapter 1 where transcribed. A needs assessment was written as a summarization for the interviews, outlining the problem significance as experienced by users in different target groups. The raw data from the interviews and the processed needs assessment were used to produce multiple solution requirements that were defined in Volere shells[54]. Volere shells are templates for defining requirement specifications.

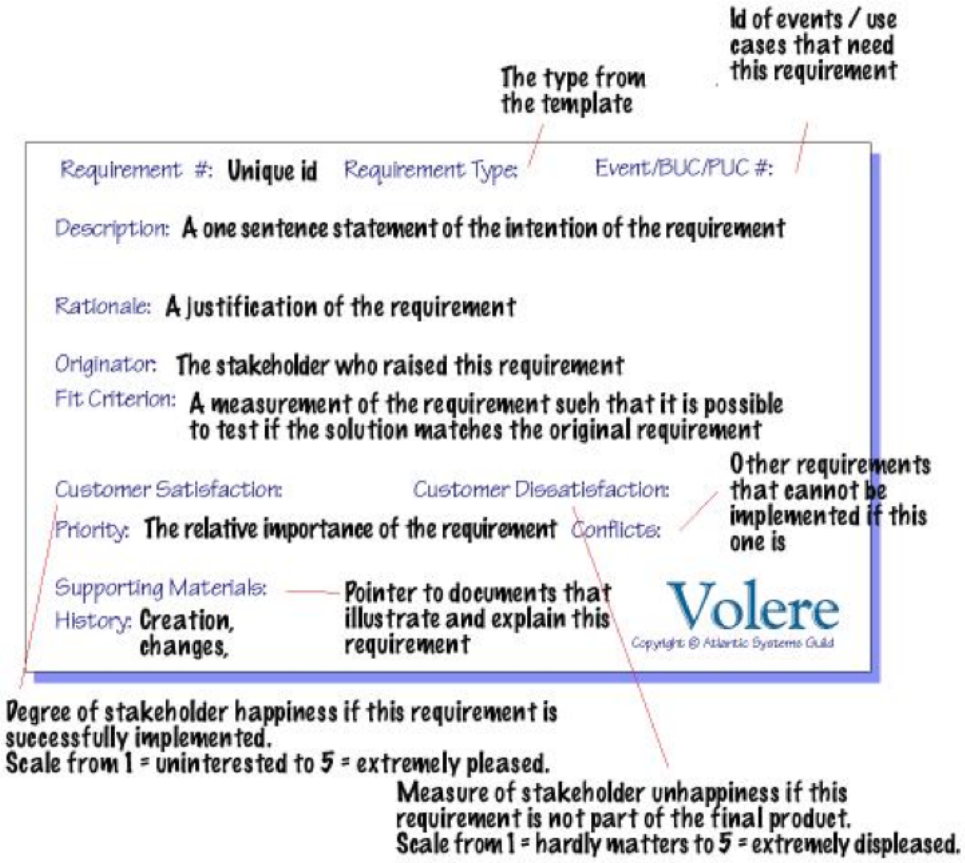


Figure 3.2: The Volere shell template with explanations[54]

3.3 Design & Development

The design and development followed the Human-Centered Design process.

Human-Centered Design is an iterative process that concentrates on tailoring a system to the end-users needs. The process considers end-users needs in all aspects of the design process resulting in a design with high user satisfaction. HCD holds four main activities, also visualized in fig. 3.3;

1. Understanding and specifying the context of use.

2. Specifying the user requirements.
3. Producing design solutions.
4. Evaluating the design.

Understanding and specifying the context of use was the process of interviewing users in the specified target group and testing a paper prototype representing the idea of a potential solution. Thereafter, a PACT analysis was conducted based on the interviews and paper prototype results. This step was only conducted once, as the iterative part of the process did not require a new context of use. This part is further explained under chapter 4.

Specifying the user requirements included creating user stories based on the needs conducted from the context of use specification. Further, both functional and non-functional requirements were created; these requirements are presented in the Volere requirement shells, and the results can be seen in section 3.2.2.

Producing design solutions consisted of creating a design solution that fulfilled most of the user requirements specified in the previous step. This solution was presented in a high-fidelity prototype. This solution is further explained under in section 5.4.

Evaluate the design the evaluation of the design was done by user testing. The solution created in the previous step was evaluated in an iterative process. The results provided from the first user test lay the foundation for the improvement made to the second prototype, which was evaluated in a second iteration. This process is explained in section 3.4, section 3.5, and section 5.5.

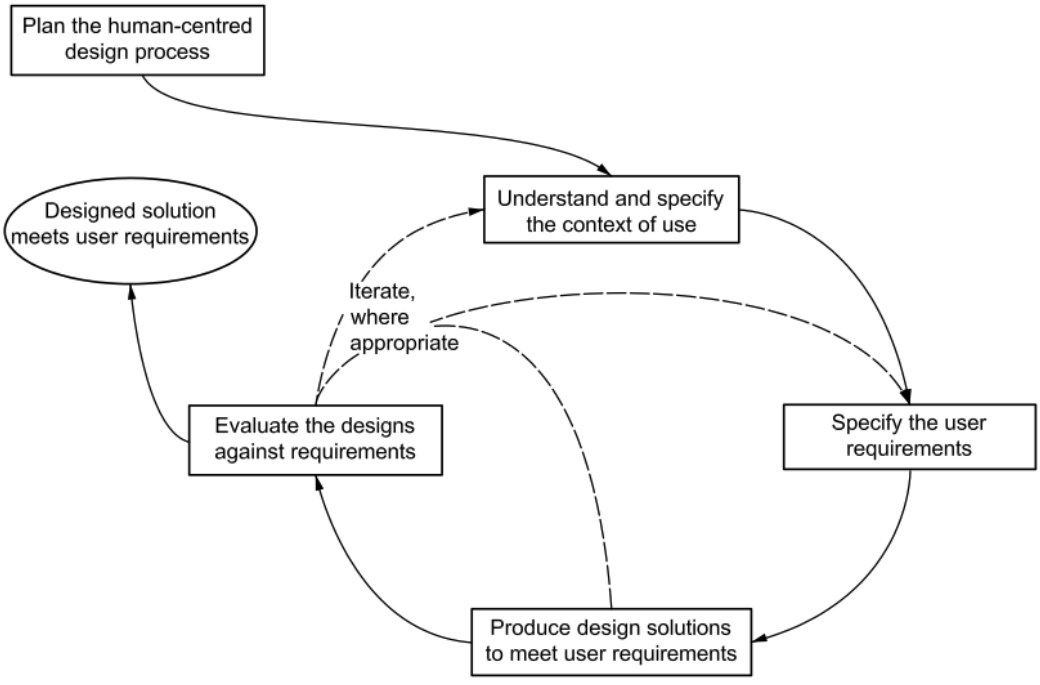


Figure 3.3: Interdependence of human-centered design activities

3.4 Demonstration

In order to create an application that helps improve the explainability of medical journals and suggest personalized health measures; a high-fidelity prototype was developed. High fidelity prototype simulates an actual user interface and allows the designers to observe the

user's behavior during the use of the prototype. This enables the designer to make changes to the product before implementing it, saving a lot of time and effort.

In order to create the high fidelity prototype, the interactive process of HCD seen in fig. 3.3 was followed. The first two steps of the process; understanding and specifying the context of use and specifying the user requirements, laid the foundation for the design choices used in the third step of the process being the production of the design solution.

3.5 Evaluation

The technical evaluation of the project was done through measurements of key parameters corresponding to the different technologies used. We will first introduce the performance metrics used, and then elaborate on how they were used for each part.

3.5.1 Metrics

Precision: The precision score indicated how many of the classified positive values are correct or, in short, the correctness rate of positive classifiers. The result will vary between 0.0 and 1.0, with the latter being a perfect score. A completely random binary classification should produce a precision of 0.5.

$$Precision = TP/(TP + FP)$$

Recall: Precision indicates the number of true positives, but it does not consider the amount of false negatives; because of this, we use recall. The recall score will produce results in the same range as *precision*, with 1.0 being a perfect score. Much like *precision*, the recall shows the rate of positives classified correctly out of positives present in the data.

$$Recall = TP/(TP + FN)$$

F1-score: Finally, a combination of both precision and recall is used, namely f1-score(also called F-measure). The F1 score has the same score range as both Precision and Recall.

$$F1 = (2 * Precision * recall)/(Precision + Recall)$$

Correlation coefficient: Correlation coefficients are numerical measures of correlation between two numbers. Whereas precision, recall, and F1-score are used to evaluate systems with discrete values such as classification, correlation coefficients are used to evaluate systems with continuous values.

The Pearson Correlation Coefficient (PCC) is used to measure linear correlation. Spearman's rank correlation measures the Pearson correlation between the ranking of two variables. Given the rankings IR and ER, where IR is the model ranking and ER is the human evaluated ranking, it is defined as

$$\rho(IR, ER) = \frac{cov(IR, ER)}{{}^{\sigma}IR{}^{\sigma}ER}$$

where $cov(IR, ER)$ is the covariance and ${}^{\sigma}IR$,${}^{\sigma}ER$ are the standard deviations of the variables IR and ER. [51]

ROUGE: ROUGE is a set of measures used to compare a generated summary to a human-created gold standard summary. ROUGE-N is one of the ROGUE measures and uses n-gram overlap to evaluate a candidate summary to reference summaries. [36]

3.5.2 Application of metrics for evaluation

Patient recommendations

The diagnosis of the diseases in scope (diabetes and cardiovascular condition) mainly consists of random forest classification. With the diagnosis classification being binary in terms of "either the patient has a given disease, or they do not", the evaluation consists of f1-score, recall, and precision. The importance of these scores varies from what is important in the project. We want to minimize false negatives (results we predict as negatives but in reality are positive), as this could lead to a patient not seeking medical attention for a potentially harmful condition.

With the extraction of feature importance, it is difficult to prove the importance of any given activity without further testing or, if possible, a clinical trial. The feature importance will, therefore, only be evaluated in the form of comparing the different feature importance algorithms described in section 2.3.2.

NLP

Evaluation of NLP problems requires a variety of metrics depending on the task. Semantic similarity is generally evaluated using a set of sentence pairs with a human-evaluated similarity score. The model is then evaluated using a correlation coefficient, which will be Spearman's in this project as recommended by Reimers et al. [51]. Summarization in this project is evaluated using ROUGE-1, ROUGE-2 and ROUGE-L scores. The summaries in this project are both short and single-document summaries, for which Lin [36] recommend these measures. For NER, precision, recall, and F1 is used. In this task, precision is based on the proportion of correctly and incorrectly identified diseases, recall is based on the proportion of identified diseases to diseases identified by humans in the test data, and F1 is the harmonic mean of the two. Assertions are also evaluated using precision, recall, and F1.

HCD

The fourth step in the HCD process shown in is fig. 3.3 is the evaluation of the design. This was done by evaluating the results obtained from user testing. The usability aspects evaluated during the user test were the principles of universal design which focuses on creating a design fit for all. These principles are:

1. *Equitable use.*
2. *Flexibility in use.*
3. *Simple, intuitive use.*
4. *Perceptible information.*
5. *Tolerance for error.*
6. *Low physical effort.*
7. *Size and space for approach and use.*

The evaluation was done by observing the users during the user test and analyzing the captured video recordings of the test in addition to opening and closing interviews.

The design solution was evaluated through a user tests. The test was divided into three parts:

- *Opening interview* determining aspects that could affect the test.
- *Observation* which showed how efficiently the users solved given tasks.

- *Closing interview* that collected general feedback about the product.

The test results was evaluated in the following matter: The results from the opening interviews were analyzed in relation to each other to determine the differences between the users. The observation's success criteria were evaluated using a scale that determined how well users managed to solve given tasks. The scale consisted of the following divisions;

- *completed the task without any difficulties* this criterion was met when the users solved the given task in an expected way, as described in chapter 5.
- *completed the task with some difficulties* this criterion was met when the users managed to solve the given task on their own within a descent time range.
- *completed the task with many difficulties* this criterion was met when the users did not manage to solve the given task without additional help from the test leader or used a long time to solve the given task.

The results from the closing interviews provided feedback about the system. That, together with the observation part, was evaluated by how well they fulfilled the principles of universal design.

These principles are:

1. *Equitable use.*
2. *Flexibility in use.*
3. *Simple, intuitive use.*
4. *Perceptible information.*
5. *Tolerance for error.*
6. *Low physical effort.*
7. *Size and space for approach and use.*

The 7 Principles [63]

3.6 Communication

After completing the project, this thesis will be made available through the University of Agder academic institution. Further, the project meetings were held regularly with the project supervisor and internal meetings in the group. The result of the communication stage is represented in this thesis, written with the LaTeXeditor Overleaf.

Chapter 4

Needs assessment and requirements

The needs assessment and requirement chapter introduced the reader to the early actions completed in the project, done to build a set of requirements for development. These requirements were to produce all parts of the project.

4.1 Needs assessment

A needs assessment was produced from the results of the qualitative interview performed early in the project. A PACT analysis was developed from the needs assessment and was a key part in the production of the Volere shell requirements as described in section 3.2.2.

A full overview of the preliminary needs assessment results can be found in appendix G. The needs assessment showed a significant variance in both experience, knowledge, and understanding, without any prevalent correlation. The target audience significantly lacked an understanding of the language used in medical journals. The most surprising information from the interviews was the lack of understanding in health care professionals, often resorting to a third-party explanation of medical expressions. This is both natural and understandable, but it could give the impression of the medical field becoming so specific that a lot of meaning is lost in semantics.

Patients often found the feedback from their primary physician to be difficult to understand, as Latin words and unknown terms were often used in the written explanation of both the condition and the actions the patients themselves could do. We found that the gap between the different target groups (doctors, nurses, and patients) was larger than expected, and there was a concrete need for explanations of different experience levels.

PACT analysis

PACT is an acronym for People, Activities, Context, and Technology. It is considered a tool to support Human-Centered design. Describe the application/product users, how the users will use the application, where and why the user will use the application, and finally, what technology is necessary for using the application.

PEOPLE	The leading target group of users are people interested in health, understanding their medical journals, or improving their health. The primary users of the system are Norwegian speaking people that have medical journals written in Norwegian. The system is designed to consider both novice and expert technology users. The user should have basic computer skills. People who struggle with color blindness, dyslexia, or visual impairment should not struggle to use the system.
ACTIVITIES	The system can be used often or rarely, depending on users' needs. Therefore it is designed to be easy to learn, easy to remember, and easy to use. Any given task will not require many clicks. The uploaded data will stay in the system even when the user gets interrupted, as long as the user or other factors have not shut the system down. When the user comes back, they can find their place again and pick up where they left off without any problems. The system will have a response time of less than 0.1 seconds for navigational purposes and less than 5.0 seconds for heavy operations such as; text summarization and data analysis. In the event of mistakes and errors, the user will always have an opportunity to correct their mistakes.
CONTEXT	Physical environment: The system shall be used in an environment suitable for computer usage, considering that the software is displaying personal medical information. If the user uses the system outdoors, the weather must be taken into consideration. Internet connection is not required as long as the users have their medical journals available digitally offline. Social context: There is no need for prior experience with the system before using it. The system is self-explanatory, and information within the system helps better understand the system itself.
TECHNOLOGY	For the system to work as intended the users should use Windows or macOS. The computer's hardware specification may affect the system's response time. Input: The users will use a mouse, touchpad, or touchscreen to navigate the system and a keyboard, touchscreen, or mouse to input specific information. Output: The system output will be displayed on the computer monitor.

4.1.1 Low fidelity prototype

A low-fidelity prototype was conducted to illustrate the idea of the application. The prototype shows the layout of the application. The created low fidelity prototype was a paper prototype that describes the functions the application performs (see Fig. 4.1).

The paper prototype was tested on five users. Users testing the application were in the age

range of 22-28, with different technical abilities. During the test, the users were explained the idea behind the application and a description of how the application worked using the paper prototype. After the prototype was explained, the users shared their thoughts and suggested improvements.

Description of the paper prototype
The paper prototype consists of three main sections; a landing page, a journal extension, and a decision system (Fig. 4.1). Starting on the landing page. The landing page has three main components. A summary component will show a summary of the journal and show generated keywords after the journal has been uploaded. While the journal is not uploaded, the summary will remain empty. The two remaining components are buttons that will take the users to the two remaining sections.

Journal Extension contains three main components. Here, user can upload their medical journal by pressing the "upload journal" button. After the journal is uploaded, the application will provide the user with keywords generated from the medical journal. The keywords will be presented in the "keyword" box. The last thing the users can do in this section is to search for specific keywords in the medical journal.

The decision system also contains three main components. Here the user can select diseases and factors to get suggested health measures. The health measures will be presented on the right side of the application. The most important health measure will appear at the top and present the importance of the measure with percentage. The users can also use the wand-button to fetch keywords generated in the medical journal section. This will select relevant diseases and factors automatically based on the information provided in their journal.

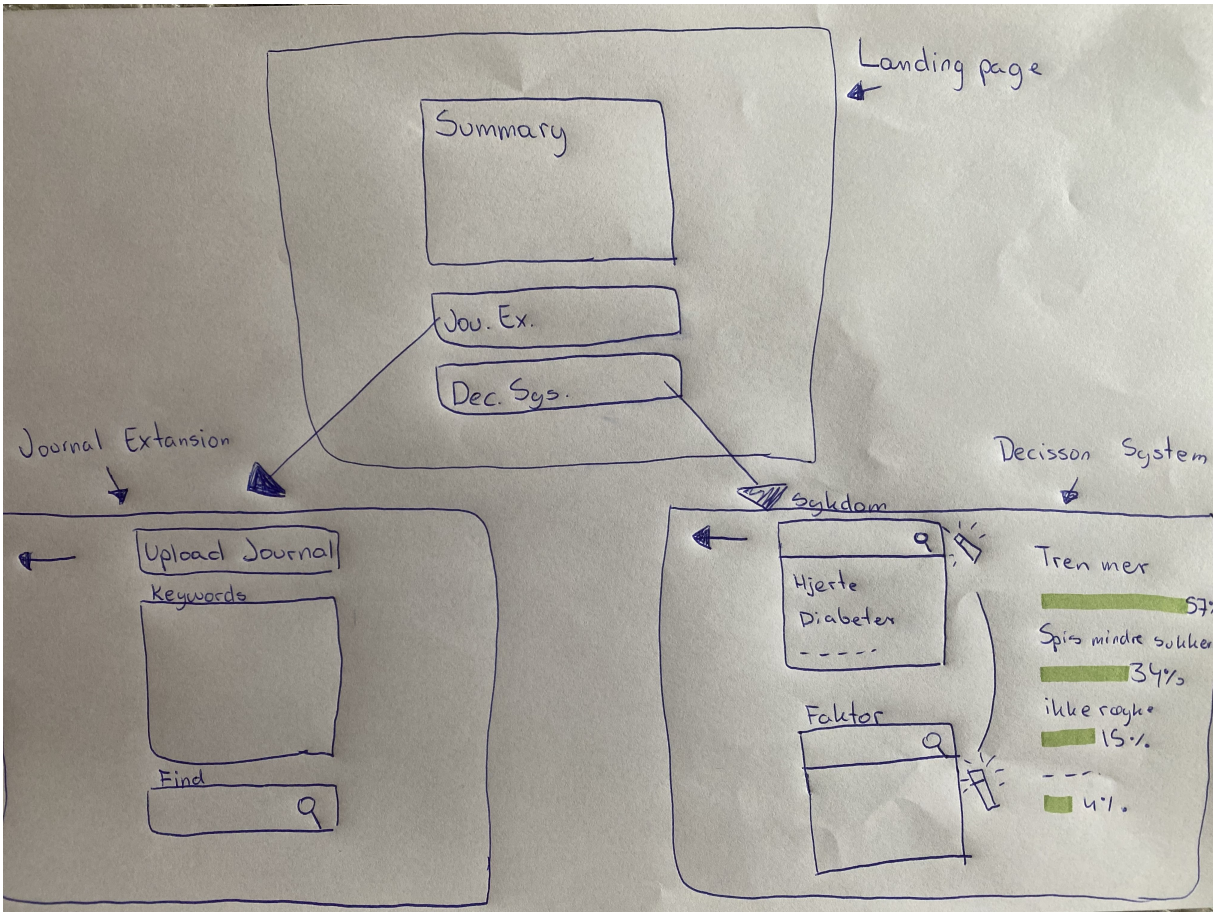


Figure 4.1: Paper prototype

4.1.2 Summary of the findings from paper prototype

Looking at the landing page, the users did not like that the summary was the largest element on the page as it was not the most important one. Some users did not understand why the summary was placed on the landing page; it made more sense to the users to have it in the journal extension section.

Journal extension: The users did not understand how to access their journal; they suggested having a log-in section to retrieve the medical journal directly from Helsenorge. It seemed inconvenient to download the medical journal in order to upload it to the application. The users did not understand what they were supposed to upload and suggested describing this part more. The keywords were not precise for the users; they did not understand why they were displayed and what happened after the keywords were found. The users asked what the application meant by "Find." They suggested specifying what they could find in the application. The users did not like that all elements were placed in the middle of the page and suggested distributing the elements more. Users indicated that it would make more sense to upload the medical journal and then have the possibility to see the uploaded journal, have the possibility to search in the medical journal, and then possibly get keywords.

Decision system: The users suggested having a set of preselected questions relevant to the health measures to get more precise health measures. The users should then get a choice between answering these questions or not. While answering the questions, the user should have the possibility to skip questions they did not want to answer. When a user chooses to synchronize information from their journal, the factors and diseases selected for the user should appear at the top of the list. So the users could get an overview of what has been selected and change the chosen factors if needed. The users suggested having a function that allowed to filter health measures based on the diseases. Users wanted to be provided with additional information regarding the suggested health measures, how to perform the suggested measures, how often, etc.

General feedback for the system was that the names on the buttons and titles throughout the application were not very descriptive. The users suggested having everything in Norwegian. They wanted a more suitable front page. Overall the users did not find the application very intuitive and suggested having more visible buttons and more descriptions throughout the application.

Feedback from testing the paper prototype can be seen in appendix A.

4.1.3 User stories

The following list contains the user stories produced from the preliminary interviews(appendix H), the needs assesment(appendix G) and the feedback of the paper prototype. They represent the desires of the users including a justification for the desire. The user stories do not necessarily represent the final product, but more so a request of functionality. Some of the requirements may the same functionality, while others may be discarded as to limit the workload.

1. As a user, I want the advanced medical terms in the medical journal explained so that I could manage to understand my medical journal on my own.
2. As a user, I want to read my journal without having to look up additional information so that I can understand it better and spare time.
3. As a user, I want the doctors to avoid using abbreviations in my medical record so that the medical journal becomes easier to read and understand.
4. As a user, I want to get both specific health-related measures and an explanation of

why to follow them so that I can choose which one I want to look at depending on the situation.

5. As a user, I want to get an additional explanation of suggested health measures so that I know how and why I should follow them.

6. As a user, I want a structured and organized way of finding specific medical information so that I succeed in finding the correct medical information efficiently.

7. As a user, I want a summary of my medical journal so that I can easily get an overview of my earlier doctor appointments.

8. As a user, I want my doctor to have more time to answer all questions about my medical problems and use more straightforward language to convey the information so that I can better understand what I have to do to improve my condition.

4.2 Requirement specification

The data for the requirement specifications was gathered from interviews with possible users, the needs assessment, and from stake holders, such as project owner, project developers and project designer. They are also based on the user stories, though some have been discarded as to limit the scope of the project. The requirement types use the Volere requirement specification template[54] to determine the type of requirement. A summary of the requirements can be seen in table 4.1. The PUC refers to the user stories in section 4.1.3, with PUC matching up with the enumerated list.

Table 4.1: Table for requirement types

ID	**Requirement type**	**Description of the requirement**
1	Functional requirement	It shall be possible to upload a medical journal into the system
2	Functional requirement	The system shall provide an explanation of challenging medical terms found in the medical journal
3	Functional requirement	The system shall provide a summary of the medical journal
4	Functional requirement	The system shall provide keywords that allows to search in medical journal, the keywords are generated from the medical journal
5	Functional requirement	The system shall search for relevant parts in the medical journal using generated keywords
6	Functional requirement	The system shall search for relevant parts in the medical journal based on a search word provided by the user
7	Functional requirement	The system shall provide health measures based on the information provided by the user
8	Functional requirement	The system shall extract data from a medical journal provided by the user
9	Functional requirement	The user shall be able to provide medical data to the system manually
10	Functional requirement	The system shall provide an explanation of the presented health measures
11	Usability requirement	The users interaction with the system shall be easy and intuitive
12	Usability requirement	The system shall be easy to use and easy to learn
13	Performance and scalability requirement	Heavy operations shall be frontloaded so that navigation remains responsive
14	Performance and scalability requirement	The system shall run as a native desktop application
15	Portability and compatibility requirement	The system shall be used an a personal compute
16	Security requirement	The system shall not rely on the internet for data transmit
17	Localization requirement	The system shall not rely on the internet for data transmit

4.2.1 Functional requirements

Requirement #: **1** Requirement Type: **9** Event/BUC/PUC #: **7**

Description: **It shall be possible to upload a medical journal into the system**
Rationale: **So that the system can use the medical journal to provide the user with intended functionality of the system**
Originator: **Project designers**
Fit Criterion: **The system shall be able to use data provided by the user and the data shall be an exact copy of what the user intended to provide**

Customer Satisfaction: **4** Customer Dissatisfaction: **5**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **2** Requirement Type: **9** Event/BUC/PUC #: **1**

Description: **The system shall provide an explanation of challenging medical terms found in the medical journal**
Rationale: **To ease reading of the medical journal and make it more understandable**
Originator: **Anonymous interview subject**
Fit Criterion: **The challenging words from the medical journal shall be highlighted on hover. The explanation of the word shall appear when the user clicks on the word**

Customer Satisfaction: **5** Customer Dissatisfaction: **5**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **3** Requirement Type: **9** Event/BUC/PUC #: **7**

Description: **The system shall provide a summary of the medical journal**
Rationale: **To gather an overview of the most important content from the medical journal**
Originator: **Project designers**
Fit Criterion: **The summary shall appear in the summary section within 5 seconds after the journal has been uploaded. The summary shall contain the most important parts from the journal**

Customer Satisfaction: **5** Customer Dissatisfaction: **4**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **4** Requirement Type: **9** Event/BUC/PUC #: **6**

Description: **The system shall provide keywords that allows to search in medical journal, the keywords are generated from the medical journal**
Rationale: **To present more relevant information**
Originator: **Project designers**
Fit Criterion: **The generated keywords shall appear in the keywords section, the keywords shall be sorted alphabetically**

Customer Satisfaction: **2** Customer Dissatisfaction: **3**

Priority: **Medium** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **5** Requirement Type: **9** Event/BUC/PUC #: **6**

Description: **The system shall search for relevant parts in the medical journal using generated keywords**
Rationale: **To ease the search for specific diseases and events in the medical journal**
Originator: **Project designers**
Fit Criterion: **By selecting one of the generated keywords the system shall present the most important parts from the journal sorted by semantic similarity**

Customer Satisfaction: **3** Customer Dissatisfaction: **2**

Priority: **Medium** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **6** Requirement Type: **9** Event/BUC/PUC #: **6**

Description: **The system shall search for relevant parts in the medical journal based on a search word provided by the user**
Rationale: **To ease the search for specific diseases and events in the medical journal**
Originator: **Project designers**
Fit Criterion: **By searching for a specific word the system shall present the most important parts from the journal sorted by semantic similarity**

Customer Satisfaction: **3** Customer Dissatisfaction: **4**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **7** Requirement Type: **9** Event/BUC/PUC #: **4**

Description: **The system shall provide health measures based on the information provided by the user**
Rationale: **To inform the users of the diseases or health risks they might have based on the information and suggest health measures to improve those**
Originator: **Project designers**
Fit Criterion: **The relevant diseases shall be selected and appear at the top of the list, the relevant measures shall be presented and sorted by relevance. The once with most health impact will be placed at the top**

Customer Satisfaction: **5** Customer Dissatisfaction: **5**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **8** Requirement Type: **9** Event/BUC/PUC #: **7**

Description: **The system shall extract data from a medical journal provided by the user**
Rationale: **To make it easier for users to provide data to the system**
Originator: **Project designers**
Fit Criterion: **The extracted data shall correspond with the data provided by the user**

Customer Satisfaction: **5** Customer Dissatisfaction: **3**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **9** Requirement Type: **9** Event/BUC/PUC #: **6**

Description: **The user shall be able to provide medical data to the system manually**
Rationale: **To enable user to provide data to the system without using an automatic method or provide data that is not present in the medical journal**
Originator: **Project designers**
Fit Criterion: **The data available to the system shall correspond to the user intention**

Customer Satisfaction: **5** Customer Dissatisfaction: **5**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **10** Requirement Type: **9** Event/BUC/PUC #: **5**

Description: **The system shall provide an explanation of the presented health measures**
Rationale: **To explain the health measures more accurately**
Originator: **Anonymous interview subject**
Fit Criterion: **A question mark placed behind each health measure shall highlight on hover and provide the needed explanation when the user clicks on the question mark**

Customer Satisfaction: **4** Customer Dissatisfaction: **4**

Priority: **Medium** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

4.2.2 Non-Functional requirements

Requirement #: **11** Requirement Type: **11** Event/BUC/PUC #: **2**

Description: **The users interaction with the system shall be easy and intuitive**
Rationale: **To make the system enjoyable to use and save time**
Originator: **Anonymous interview subject**
Fit Criterion: **The users solve intended task without failure**

Customer Satisfaction: **5** Customer Dissatisfaction: **5**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **12** Requirement Type: **11** Event/BUC/PUC #: **2**

Description: **The system shall be easy to use and easy to learn**
Rationale: **So the users can use the system without problems even when using the system rarely**
Originator: **Anonymous interview subject**
Fit Criterion: **The users manage to solve tasks on first try without failure**

Customer Satisfaction: **5** Customer Dissatisfaction: **5**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **13** Requirement Type: **9** Event/BUC/PUC #: **2**

Description: **Heavy operations shall be frontloaded so that navigation remains responsive**
Rationale: **To make the system feel more satisfying to use**
Originator: **System designers**
Fit Criterion: **The navigation shall take less than 0,1 sec and heavy operations less than 5sec**

Customer Satisfaction: **3** Customer Dissatisfaction: **4**

Priority: **Medium** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **14** Requirement Type: **12** Event/BUC/PUC #: **-**

Description: **The system shall run as a native application**
Rationale: **To maximize the performance by avoiding the overhead of a web browser**
Originator: **System designers**
Fit Criterion: **The system runs as a native application**

Customer Satisfaction: **2** Customer Dissatisfaction: **3**

Priority: **Low** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **15** Requirement Type: **13** Event/BUC/PUC #: **-**

Description: **The system shall be used an a personal computer**
Rationale: **As the smart phones have smaller screens and tablets are not in wide spread use compared to personal computers**
Originator: **System designers**
Fit Criterion: **The system is used on a personal computer**

Customer Satisfaction: **3** Customer Dissatisfaction: **4**

Priority: **Low** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **16** Requirement Type: **15** Event/BUC/PUC #: **-**

Description: **The system shall not rely on the internet for data transmit**
Rationale: **To avoid the possibility of data breach**
Originator: **System designers**
Fit Criterion: **The system runs without internet**

Customer Satisfaction: **2** Customer Dissatisfaction: **2**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Requirement #: **17** Requirement Type: **16** Event/BUC/PUC #: **1**

Description: **The system shall use Norwegian as its primary language**
Rationale: **To target Norwegian speaking users**
Originator: **Anonymous paper prototype tester**
Fit Criterion: **The primary language of the application is Norwegian**

Customer Satisfaction: **4** Customer Dissatisfaction: **5**

Priority: **High** Conflicts: **0**

Materials: **Needs assessment**
History: **Created march 15th, 2022**

Chapter 5

Solution

This chapter will describe our approach in solving the problem of simplifying medical journal data, extracting key data, processing the data to provide decision support and finally presenting everything mentioned to the user.

5.1 Conceptual Model

The initial development started with a conceptual model. The model shown in fig. 5.1 is meant as a psychological representation of the task execution and flow of the application. By creating a conceptual model early, it was shown how the application was intended to be used. fig. 5.1 shows that the user can interact with the application without providing any personal data, with the exception of the natural language processing part. The dotted lines are the only actions where the users need to input any data, such as "Import journal" or "Type query".

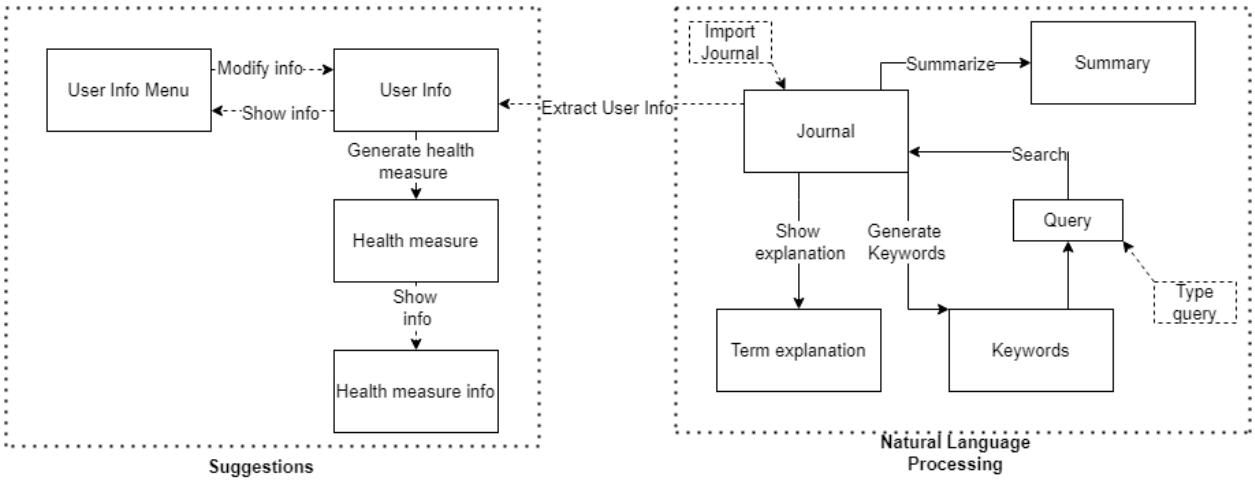


Figure 5.1: Conceptual model of the application functionality

5.2 Natural Language Processing

From the requirement specification presented in section 4.2 we identified six requirements that could be satisfied using NLP methods. We also identified the NLP tasks the requirements corresponded to. We developed a set of desired attributes for our solution based on

our research on contemporary NLP methods, available clinical data, and the identified requirements. We were able to evaluate solutions for summarization, searching and assertion prediction.

Nr.	Description	Task
1	The system shall provide an explanation of challenging medical terms found in the medical journal	NER
2	The system shall provide a summary of the medical journal	Summarization
3	The system shall provide keywords that allows to search in medical journal, the keywords are generated from the medical journal	Topic mining
4	The system shall search for relevant parts in the medical journal using generated keywords	Search
5	The system shall search for relevant parts in the medical journal based on a search word provided by the user	
6	The system shall extract data from a medial journal provided by the user	Assertion detection

We chose to base a large part of our approach on pre-trained transformer-based models. These models are computationally expensive to pre-train, but require less data and resources to fine tune for specific tasks. This is beneficial for the given problem, as clinical data is difficult to access because of its sensitive nature. [67]

For the same reason, English was chosen as the model and data language. Although there is ongoing development of Norwegian-language models, it has not progressed nearly as far as English-language models [68]. By choosing to use pre-trained models for our solution, the methods employed in this project can likely be adapted to models for other languages as their development progresses.

Finally, pre-trained models can be designed to target specific domains[13][33][46]. This requires a large amount of domain-specific data when pre-training, but can increase performance on downstream tasks without additional training data. This means that one group of researchers can use sensitive data to train the underlying model, which can then be used by other researchers, thus reducing the potential risk and difficulty of distributing medical information.

The following sections outline our approach to each problem. Information about the datasets used can be found in chapter 6.

Searching

For medical conditions, it is common to have different names for the same condition at various levels of formality. For example, *myopia*, *nearsightedness*, or simply *bad eyesight* can all refer to the same condition. Optimally, the system should be able to determine that these terms all refer to the same condition, which is why searching based on semantic similarity was chosen.

Three pre-trained base models were evaluated: MiniLM[69], BioBERT[46] and BlueBERT[46]. MiniLM is a general purpose model and was used to provide a baseline comparison to the domain specific models. BioBERT is a BERT-based model that has been pre-trained for biomedical text on a collection of PubMed abstracts and articles. BlueBERT is also a BERT-based model for biomedical text. It was pre-trained on PubMed abstracts and the MIMIC-III dataset. [50][29]

These models were tuned using the sentence-transformers library[52] to produce sentence embeddings with the ability to compare semantic similarity of inputs using the cosine simi-

larity of the output vectors. The BioBERT and BlueBERT models were tuned on datasets for natural language inference and semantic textual similarity. The MiniLM-based model was trained on a large collection of question-answering datasets . The models for BioBERT and BlueBERT were provided by Deka et al. [17], and the MiniLM-based model was provided by the sentence-transformers library through the HuggingFace model hub [56][47][48].

Our approach takes a text query produced by the user and compares it to a list of sentences derived from the target text. The target text is split into sentences using the English language spaCy sentencizer pipeline[60]. The query and all the sentences are processed by the model, and the produced sentence embeddings are compared to the query embedding and ordered by similarity.

5.2.1 Summarization

Extractive summarization was chosen over abstractive summarization. Abstractive summarization requires more data and resources for training, which we want to avoid as Norwegian clinical text is a very narrow domain with limited data. Additionally, abstractive summarization models may accidentally rephrase medical terminology in an awkward or inaccurate way.

Our approach to summarization relies on similar methods to the ones used for semantic search. The text to be summarized is split into sentences, and embeddings of all the sentences are produced. As proposed by the sentence-transformers library, LexRank[22] is then used to calculate the centrality of each sentence and a sorted list is returned.

5.2.2 Named Entity Recognition

To provide explanations for challenging medical terms, we first need to establish what a challenging medical term is. For the purposes of this project, we chose terms for which we had available NER training data. We identified two datasets with annotations for diseases, the NCBI-diseases and the BC5CDR-diseases corpus, and a BioBERT-based model fine tuned to identify them.[12][35][21]

Like the previous task, the target document is split into sentences. Each sentence is processed by the NER model, which produces a list of tokens. Each token is labeled with its entity group, e.g. whether it is a disease or a chemical. It also has a marker to identify if it is part of a named entity or not. Tokens representing the beginning of a named entity is marked with a *B*, and the following tokens that are part of the same entity are marked with *I*. Tokens that are not part of a named entity are marked with *O*. Based on this, the tokens can be combined to form complete words and phrases, which can then be used with an external medical dictionary to provide explanations of the terms.

5.2.3 Information Extraction

To extract information from a medical journal, we first need to identify the type of information we wish to extract. We chose to focus on evaluating detection of the presence of a set of predetermined conditions, such as diabetes or heart disease. We evaluated a pre-trained model based on the version of ClinicalBERT trained on MIMIC-III discharge summaries with BioBERT as its base model. A combination of datasets containing annotated assertions about medical conditions was used to train a ClinicalBERT model. The model takes text with a tagged condition and attempts to classify whether the tagged condition is present, absent, or possible.[18]

5.3 Recommendations for patient-controlled change

The goal of the recommendations-part is twofold: Firstly, be able to tell from available data if the patient likely suffers from a given disease, despite not being explicitly diagnosed with it. This point becomes null and void if the data extraction provides a diagnosis on the given disease, but it can still be linked to the other disease and/or conditions we have data on. A side benefit of the errors in diagnosis, is that we do not only obtain those who do in fact have the diagnosis, but also those with a risk of obtaining it based on known factors.

Secondly, the data processing should be able to tell the patient what specific actions and/or lifestyle changes they can do to either reduce risk of or improve quality of life while living with a given condition. The diseases are in this case limited to non-communicative lifestyle diseases, such as diabetes, hearth disease, etc.

The diagnosis is the relatively simple step of training a random forest model with the available information and let the classification run its course. Assuming the information from the data extraction step contains explicit diagnosis of a given condition, the classification itself is not needed. There is however still the need for the factors of this given condition, and the training is still necessary to determine feature importance, with features being aspects of the patient such as Body mass index(BMI), blood pressure, resting hearth rate, etc. What we are currently finding is the feature importance of these features and the feature importance of the lifestyle choices that may affect the aspects of the patient, showing an indirect correlation to the condition.

5.3.1 Layer 1

The diagnosis is found by training a random forest for each condition/disease we have, where we have any overlapping data from the patient. Explaining fig 5.2 further, the input provides the known data provided by the data extraction. In the model the data shown is limited to the diabetes part of the data processing. The importance of the factors can be viewed in the diagnosis feature importance panel. The data is split in to tags we have data on, such as bmi, and tags that we do not have available data on such as glucose or insulin. A forest is trained for the bmi tag, returning the importance of the actionable values from the bmi dataset.

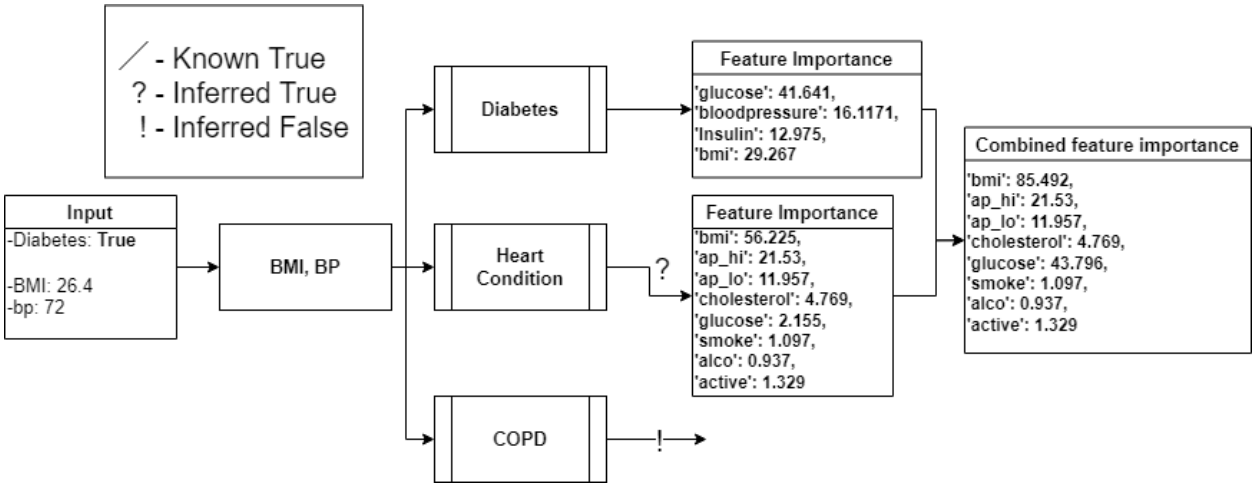


Figure 5.2: Layer 1: Visual representation of input data, and output of the first layer in the Data processing

5.3.2 Layer 2

The feature importance of the first layer are the importance of specific features, but contains no actions that the patient can perform to improve their prognosis, therefor another layer is used. The second layer takes the individual features from the first layer and if available trains another forest with available data to produce the feature importance directly tied to executable actions. An example of this can bee seen in fig 5.3.

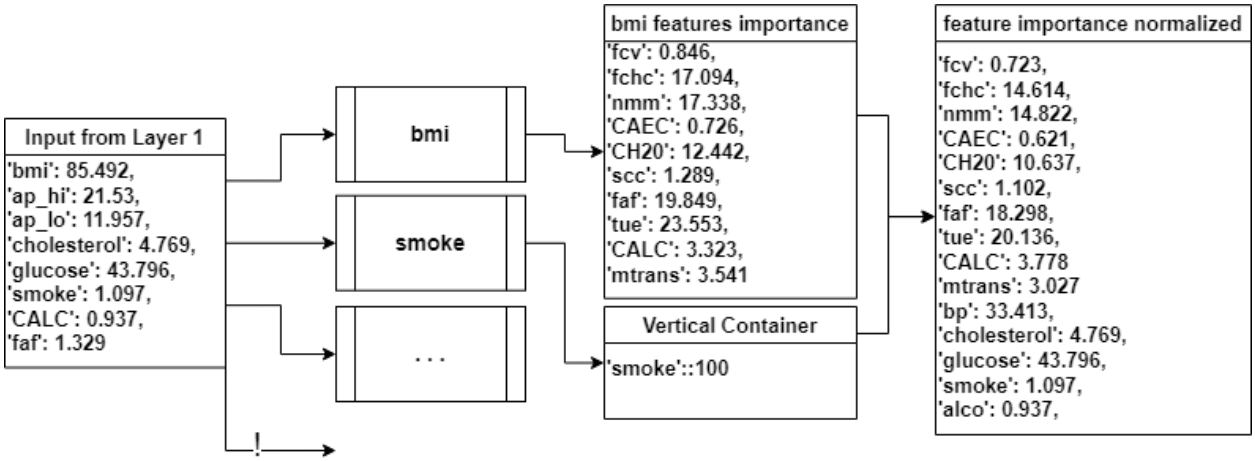


Figure 5.3: Layer 2: Visual representation of input data, and output of mixed actionable results

The example provided in fig. 5.3 also shows the limitations of the solution. Without available data connecting a given feature to any action and/or activity, we are unable to give a complete list of feature importance. The suggestions model was developed in python with the use of a series of modules to handle separate responsibilities. flask and flask_restful was used to create a single API endpoint to allow communication with the frontend. Using a web API for communication between the different parts something something. Further sklearn was used for the creating the random forests together with pandas for handling datasets.

5.4 User Interface (UI)

The user interface is the part of the software application that the users sees and interacts with. This is the first thing the user the user sees when being introduced to the application, and has to interact with for the duration of using the application. Because of this the UI is a vital part of the system.

5.4.1 High fidelity prototype

The application's name "HIPPO" is an abbreviation of Hippocrates, also referred to as the "Father of Medicine,". Hippocrates was a significant part of medical history. The chosen logo is, therefore, a hippopotamus in a medical scrub. The logo (fig. 5.4) was created in Adobe Photoshop [3].

Figure 5.4: The hippo logo

The high-fidelity prototype was created using Adobe XD [2]. It is a clickable prototype that is not fully functional. Adobe XD lets the user choose the size of the window it will be showing so that the prototype matches the intended device for the application. The selected device for this application is desktop and the resolution used in the prototype is therefore 1400x800.

5.4.2 Design choices

The main goal of the design was to design an application fit for most people of different ages and technical abilities. To accomplish this goal the design follows principles of universal design [63].

The primary usability goals focused on during the design were that the application was intuitive, easy to use, and easy to learn. To accomplish these goals, the following UI principles were followed;

Transparency lets the users focus on solving goals in the application without being distracted by the interface. Transparency avoids adding unnecessary elements to the interface, such as; graphics, buttons, windows, and attachments. The application uses transparency throughout the whole application by having few elements, and avoiding distractions in the design.

Clarity lets the user have a meaningful and understandable interaction. The application uses clarity throughout the design by providing the users with a clear purpose and clear choices in the application. There is no hidden information in the application. fig. 5.5 shows a good example of the use of clarity in the application. The figure shows the landing page with two main buttons; this provides the user with few options, reducing confusion.

Friendliness warns the users if they are about to delete important data or damage the system; it lets users correct their errors. The application uses friendliness by allowing users to correct the choices made in the application. fig. 5.10 shows a question where users can go back and forth as they please in order to correct their answers.

Consistency lets users solve the problems faster by saving time on understanding the differences in using different elements and functions. The application uses consistency by making similar elements like buttons and windows look and work similarly. fig. 5.7 shows two navigational buttons in the top left corner and bottom right corner. The buttons are designed in a similar matter and act as navigational buttons.

In addition to the previously mentioned UI principles the application uses visual design principles to increase usability fig. 2.2

Scale The application uses the scale principle by making the most important elements the largest, so it's easier to spot them. fig. 5.5 shows the landing page, where it is easily presented that the most important elements on this page are the journal and health measure buttons.

Visual Hierarchy is used throughout the design to guide the users from the most important part to the least important in order.fig. 5.7 guides the user to the medical journal, which is the most important part of this section. From there, the user's attention is guided to the summary of the journal and, end-wise, to the search section in the journal.

Balance The application uses balance by distributing elements symmetrically relative to the central imaginary axis. As seen in fig. 5.6 on the left side of the axis there is an large element and to balance it out there are two smaller elements on the right side of the axis.

Contrast The application uses contrast to show the difference between elements.

Gestalt Principles are used to show that elements similar to each other have a relation to each other and perform similar tasks. The application uses gestalt principles in several ways. fig. 5.11 shows that similar elements are grouped together. The diseases and factors are designed similarly, and both contain buttons that can be selected and unselected. These buttons serve the same purpose.

In order to create an application accessible for all users regarding age, size, gender, ability, or disability, principles of universal design were used during the design process. In order to create a design that is equitable to use by most, the design follows World Wide Web Consortium [73] (W3C) guidelines[70]. Therefore, a large font size and a high contrast ratio suitable for visually impaired users are used throughout the design. According to W3C small text should have a contrast ratio of at least 4.5:1[70], where the contrast ratio of small text in the application is 21:1. W3C also recommends having a contrast ratio of at least 3:1 for large text[70], where the designed application uses at least 4.21:1 on large elements, like buttons. The application uses a font type called Lato, a type of Open Sans font. According to Morales [38] "Open Sans is a highly readable, neutral, and minimalist font to choose from. This sans-serif font is one of the best fonts for user experience (UX) and readability. Open Sans is a safe option for most experiences and works best for businesses that value quality control and reliability. Some of the best websites of 2020 are designed in Open Sans. "

5.4.3 Application flow

The horizontal prototype approach was chosen to showcase the majority of the application's functionality. The horizontal approach lets users explore significant parts of the application.

When the users open the application, the first that meets them will be the logo shown in fig. 5.4, with a quote saying "easier way to health." After a few seconds, the page will transition to the landing page shown in fig. 5.5.

The users are asked what they want to do in the application on the landing page. The users can choose between two buttons: look at the journal or the health measures.

Figure 5.5: First high-fidelity prototype showing the landing page

If the user selects the journal part, the application will display the journal part shown in fig. 5.6. The journal part consists of three main components; the medical journal, summary, and find in the journal.

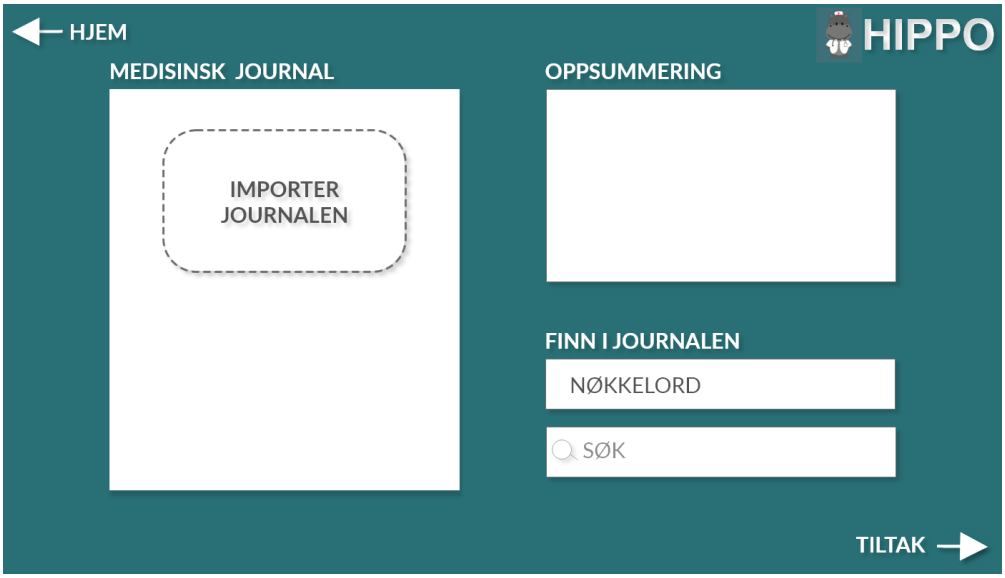


Figure 5.6: First high-fidelity prototype showing the journal section

The first thing the users have to do is to import the journal; it is done by clicking on the import journal button below the medical journal. Once the journal is imported, the application will display the medical journal in the medical journal section and automatically display the medical journal summary under the summary section. The user can hover over difficult medical terms while reading the medical journal. The difficult words will light on hover and provide the user with an explanation if the user clicks on the words as shown in fig. 5.7.

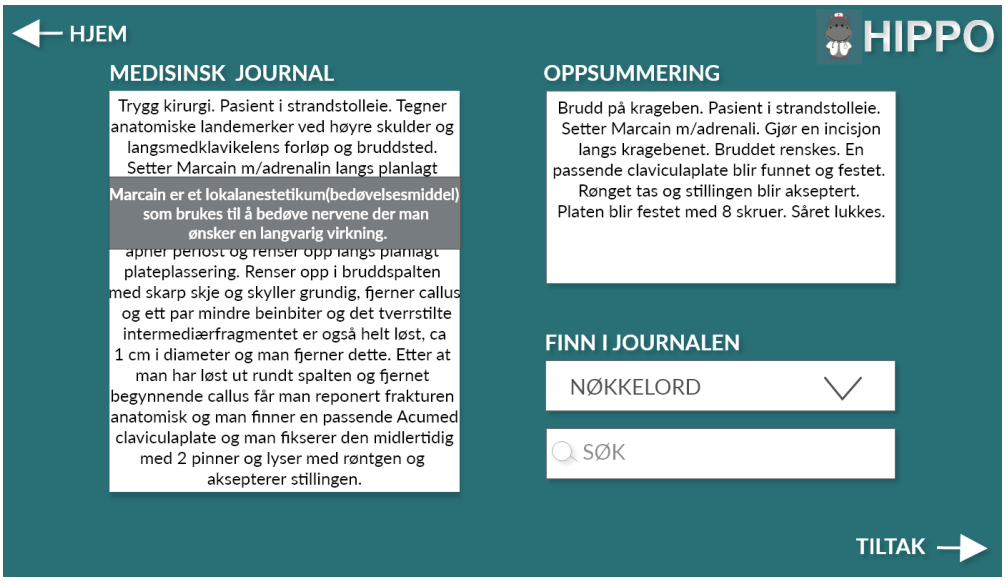


Figure 5.7: First high-fidelity prototype showing medical term explanation

To ease the search process, the user can search in the journal in two ways; they can search by the keywords provided by the application, which will appear if the users click on the keywords drop-down as shown in fig. 5.8. If the user clicks on one of the generated keywords, the application will display the most important sentences from the journal containing the keyword. The sentences will appear in the place of the medical journal and will be sorted by relevance.

Figure 5.8: First high-fidelity prototype showing keywords

If the user clicks on one of the provided sentences, the application will display the part of the medical journal containing the selected sentence. This searching process can also be done by searching for own words in the search bar below keywords.

If the user selects the health measure part, the user will be asked if they want to answer questions to get more precise health measures, alternatively if they want to skip that part as shown in fig. 5.9.

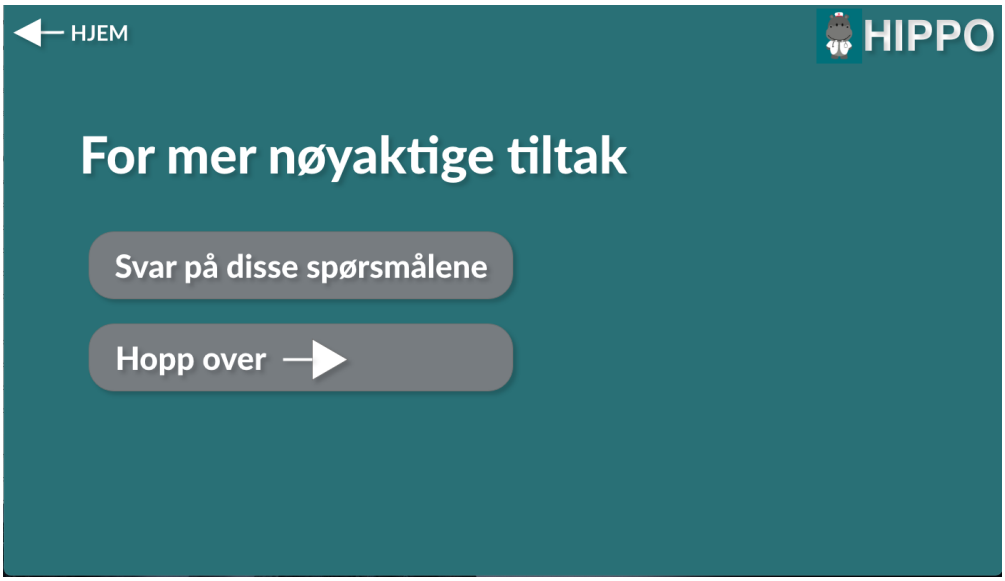


Figure 5.9: First high-fidelity prototype showing choice for more precise health measures

If the users select the more precise health measures, they will be asked if they want to synchronize their journal. The user can answer yes or no, after answering these questions. The user will have to answer a few questions that will help to provide the user with precise health measures. fig. 5.10 shows an example of these questions. The questions asked are the following:

- How tall are you?
- How much do you weigh?
- Do you smoke?
- How much water do you drink daily?
- How often do you work out each week?

Figure 5.10: First high-fidelity prototype showing questions for more precise health measures

The user can choose to skip questions they do not want to answer; they can also go back and forth to correct their answers. After all questions, the user will be presented with

the health measures as shown in fig. 5.11. If the users chooses to answer the additional questions, some relevant factors will be preselected. If the users choose to answer questions and synchronize their journal, the application will preselect information found in the journal and the additional factors from the questions. If the user does not wish to answer any questions or synchronize the journal, none of the factors or diseases will be selected.

The health measure part shown in fig. 5.11 consists of three main components: Diseases and factors which are selectable for the users, and efficiency degree that shows the efficiency of the presented health measures.

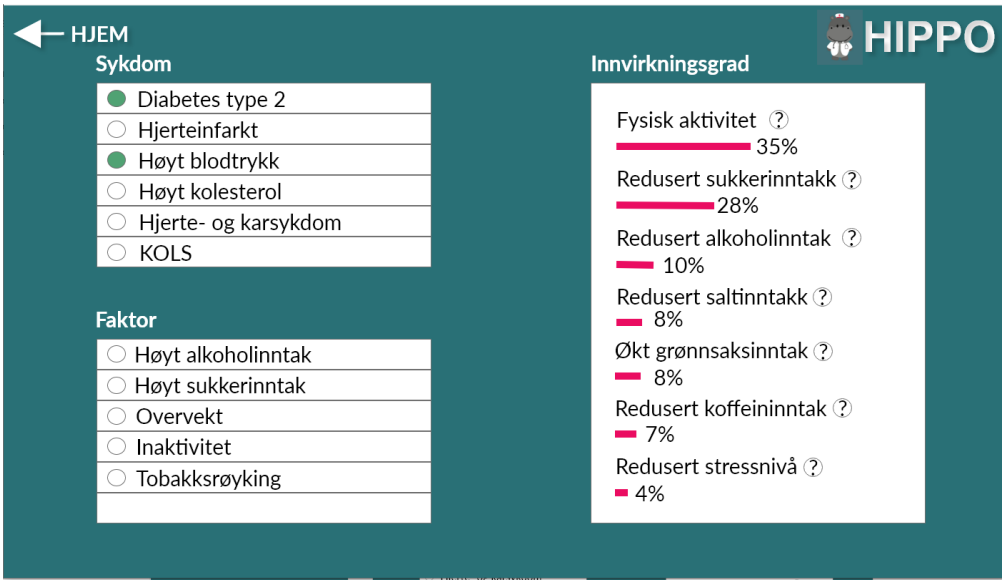


Figure 5.11: First high-fidelity prototype showing suggested health measures

5.5 User tests

User tests were used to evaluate both of the high-fidelity prototypes. Results from the first user test were used to improve the second prototype. Results from the second prototype were listed as future work due to time limitations. The user test took place at the University of Agder's usability lab (Fig 5.12). The users used a MacBook pro 13" (2560 × 1600), an external mouse, and a mousepad. The test team consisted of three people: two test leaders and one main observer.

Both tests were executed in the same way. The test was divided into three main parts: an opening interview, an observation, and a closing interview.

During the opening interview, a test leader greeted the users, explained what the test would look like, explained the idea behind the system, provided the users with privacy information, collected the signed consent forms, and asked the interview questions. The first interview was composed of questions that would help discover aspects of the users that could influence the test. The following questions were asked:

Opening interview questions

- How familiar are you with health concepts, what they are, and what they represent?
- How often do you use a computer, what system, and how knowledgeable would you say you are?
- Age range (18-25, 26-35, 36-45, 46-55, 56-65, 66+)

- Do you have reduced vision or any form of color blindness, mobility difficulties, dyslexia, ADHD, or other atypical neurology?
- Gender (male, female, other)

During the observation part, a test leader explained to the users that the system they will test is a prototype and that it, therefore, is not entirely functional. The test leader urged the users to think aloud and informed them that it is utterly optional to answer the questions and that they can end the test whenever they want without any explanation. After the necessary information was provided, the observation began. During the observation, the users were asked to perform several tasks in the system. The tasks the users received were created to test most of the system's functionality and check if it was intuitive, efficient, and easy to use. The following tasks were given to the users:

Tasks during the observation

1. Can you find a summary of your journal?
2. Are there any words in the journal that you don't understand? Is there any way for you to find out what the words mean?
3. Can you find information in the journal about diabetes? Is there another way you could find this information?
4. Can you find as precise health measures for you as possible?
5. Can you find health measures for a person that smokes and has high cholesterol levels, without knowing anything else about that person?
6. Can you find out how many active minutes the application means by being "physically active"?

During the closing interview, the users were asked several questions about the system. These questions helped gather an understanding of how well the users understood the system and if they enjoyed using it and provided general feedback. The following questions were asked:

Closing interview questions

- What do you think the percentage means?
- Was there anything that stood out as more challenging during the test?
- Was it intuitive to navigate the application?
- What feelings are you left with after using the application?
- Would you use the application again?
- Was the placing of the various buttons understandable?
- Was the naming of the various buttons and titles understandable?
- Do you have any general feedback about the system?

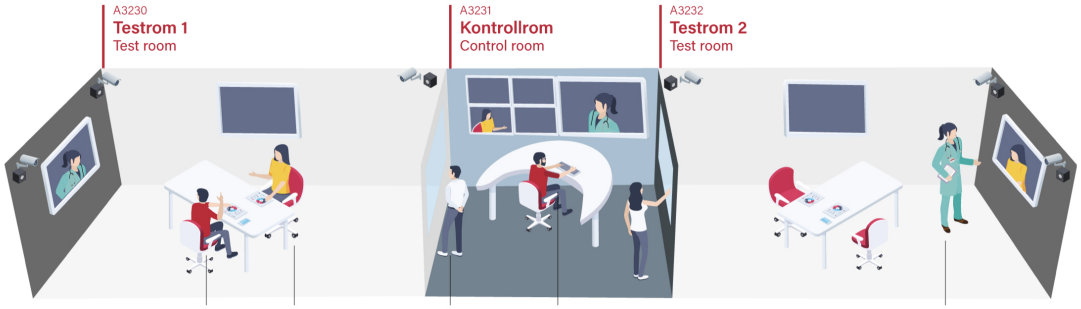


Figure 5.12: Visual representation of the usability lab

5.5.1 First user test

Figure fig. 5.13 and fig. 5.14 show demographical differences of the users participating in the first user test regarding gender, age, health terminology knowledge and computer skills. The exact results can be seen in C.

Figure 5.13: User test demography presenting age and gender

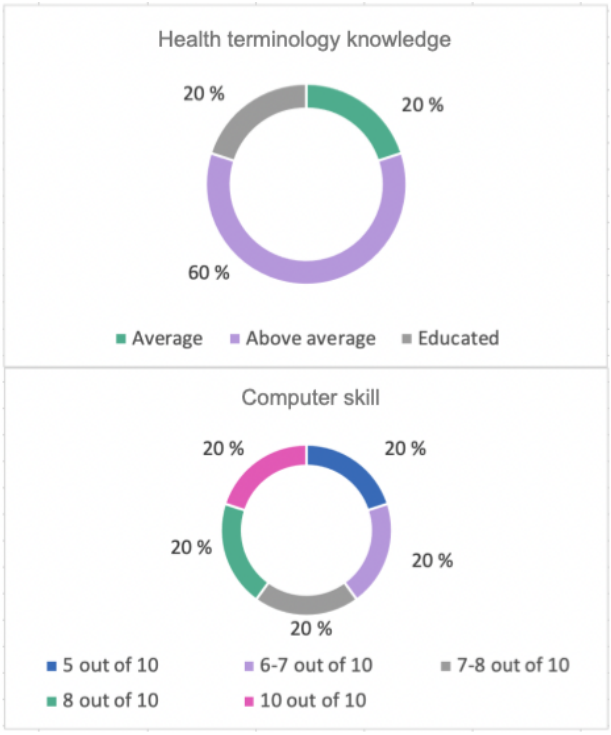


Figure 5.14: User test demography presenting health terminology and computer knowledge

Task 1: *Can you find a summary of your journal?*

The expected way of solving the problem:
Starting on the landing page

1. Select "journal" on the landing page
2. Press “import journal” button placed under the "medical journal" section
3. The summary will be presented to the right in the “summary” section

Four out of five users struggled with understanding that they had to first import the journal

in order to get a summary from the journal. Two out of the four that struggled needed to get help from the test leader in order to complete the task.

Task 2: *Are there any words in the journal that you don't understand? Is there any way for you to find out what the words mean?*

The expected way of solving the problem:
Starting in the medical journal section

1. Hover over a medical term one does not understand in the medical journal
2. Click on the medical term that lights on hover
3. The explanation of the word will appear beside the selected word

Majority of the users completed the task as intended without any problems. One of the users did not hover over the words while reading the journal and did use a bit more time to figure out that the medical terms lighted on hover. As soon as the user saw that, the user also completed the task. The users were very positive to this function, they found it very helpful and intuitive.

Task 3: *Can you find information in the journal about diabetes? Is there another way you could find this information?*

The expected way of solving the problem:
Starting in the medical journal section. There are two ways of solving this problem, one can start with either

First option:

1. Click on keywords drop-down placed below the "find in journal" section, to get a selection of searchable keywords
2. Select diabetes from the presented keywords and click on the keyword
3. Select one of the sentences containing diabetes that are now presented below "medical journal" section
4. The information about diabetes containing the selected sentence will be presented below "medical journal" section

Second option:

1. Click (which in this case will simulate input from the keyboard) in the search-bar placed below "find in journal" section
2. Diabetes will appear in the search-bar, click on the search-bar again to search
3. Select one of the sentences containing diabetes that are now presented below "medical journal" section
4. The information about diabetes containing the selected sentence will be presented below "medical journal" section

All of the users completed the task without problems.

Task 4: *Can you find as precise health measures for you as possible?*

Starting on the landing page
The expected way of solving the problem:

1. Select "health measures"
2. Select "answer these questions" on a question if a user wishes more precise health measures
3. Select "yes" on a question to whether the user wishes to synchronize the journal
4. Answer all of the questions presented
5. Select "view measures"
6. The suggested health measures are now presented below "efficiency degree" section

Three out of five users managed to complete the task as intended. However, they were not sure what the application meant by synchronizing the journal. Two of the users chose not to synchronize their journal. The users did not fully understand what the application meant by "synchronizing the journal". One of the users even thought that by synchronizing the journal the user would change some information in his original journal.

Task 5: *Can you find health measures for a person that smokes and has high cholesterol levels, without knowing anything else about that person?*

The expected way of solving the problem:
Starting on the landing page

1. Select "health measures"
2. Select "skip" on a question if a user wishes more precise health measures
3. Select "tobacco smoking" from the factors and "high cholesterol" from diseases
4. The suggested health measures are now presented below "efficiency degree" section

None of the users completed the task as intended. Three of the users solved the task by answering more precise questions but managed to skip synchronizing the journal, while two of the users answered the questions and synchronized the journal.

Task 6: *Can you find out how many active minutes the application means by being "physically active"?*

The expected way of solving the problem:
Starting in the health measure section

1. Click on the question mark placed behind the "physical activity" measure in the efficiency degree section
2. The information will appear below the question mark

Three out of five users solved this task without any problems. However, two of the users struggled with this part. One of the two users struggled to spot the button with the question mark but managed to spot it after a long while. The other user did not manage to complete this task without help from the test leader. It was noted that both of the users who struggled with this part had some visual impairment.

Summary of the findings from the first user test

In the last part of the user test, the users received questions about the application. The answers can be seen in appendix E. The summary of the feedback from both observation and the interviews will be presented in this part. Starting with the landing page, users liked the layout and the colors and that there were only two options to choose from as this made it easier to navigate the application. The users liked the hippo logo. The only suggestion for improving the landing page was to change the button name from "Journalen" to "Journal."

Moving on to the Journal part. Users did not find the importing part of the journal intuitive. It was suggested to import the journal as an isolated step to clarify this part. To make this action more understandable, some users suggested using a more uncomplicated word than "import." The summary part of the journal was well received by the users. The users quickly found the summary of their journals and were optimistic about this function. Looking at the medical journal, users found the text a bit messy; it was suggested to use more margins and not use centered text to make the journal easier to read. The medical term explanation was very well received; many users liked this part very much, as it made the journal understandable. The searching part in the journal was also well received by the users. They suggested having a search button by the search bar to make it more straightforward. Nevertheless, all users managed to search in the journal using both keywords and the search bar. Additionally, some users commented that the application was very square, and suggested also to change the "home" and "health measures" buttons on this page to buttons and not just text.

The health measures were the most challenging part for the users. On the question where users were asked if they wished to receive more precise health measures, it was not clear where the application would take the users if they chose to skip the questions. One of the users stated that they were drawn to "skip" questions probably because it was a big arrow on the button in addition to the text. Another user suggested changing the button's name from "skip" to a more descriptive button that tells what will happen if the users choose this path. For the question about synchronizing the journal, all of the users were confused. The users who chose to synchronize their journals were uncertain what that meant. One of the users thought that that could change the original content of their journal. Some users chose to synchronize their journals when they tried to find health measures for a foreign person. This part was not intuitive and created much confusion. Some users suggested having a pop-up that asked if the users were sure that they did not wish to synchronize their journals as it would give them tailored health measures. The users suggested using a more understandable term than "synchronize," which created confusion. The users did not have any problems answering the questions that gave more precise health measures. They could go back and forth to correct their answers and managed to complete this part without any difficulties. The part that displays suggested health measures was partly clear for the users. The users understood the selection of diseases and factors and managed to see that the health measures changed regarding selected diseases and factors. The part that confused the users was how the health measures were presented. The users did not understand completely where the percentage came from and what it meant. The users did not find the "efficiency degree" very explanatory either; it was cryptic and confusing for them. The suggestion was to change the title to a more descriptive one and to add an explanation to that part. For the extended information one could receive by clicking on the question mark, some users found it intuitive, while others struggled more. The question mark is a bit hard to spot, and the user that struggled with this part had some visual impairment which might be the case for why they struggled.

Other than the challenges users faced mentioned above; the majority was positive towards the application. They found it easy to navigate, that the buttons were mainly naturally placed throughout the application. All of the users liked the idea behind the application;

they found it constructive, informative, and important for society. The majority claimed that they would use the application; the frequency of the use would depend on how often their medical journal would change. The users liked the layout of the application, majority of the users liked the colors. The application was not stressful for the users, and one of the users claimed that this application was easy to learn by having used it once before, it would be very easy to use it again. The users liked that there was not much hidden information. However, there is room for improvement in the application, as some of the users were left with a feeling that the application was inconsistent and a bit chaotic.

Second prototype

Based on the feedback received from the first user test, several changes were made to improve the second version of the prototype.

The only change made on the landing page was to change the name on one of the buttons from "Journalen" to "Journal."

The journal part is now changed. Beginning with importing the journal. This step is now an isolated part as shown in fig. 5.15. Furthermore, the name of the button is now changed from "Import the journal" to "Click here to get the journal."

Figure 5.15: Second high-fidelity prototype showing an isolated section for getting the medical journal

The journal section is changed as well as shown in fig. 5.16. Starting with the medical journal, the text is not centered anymore. There are added more margins as well. There is now added a "search" button beside the search bar. The "home" button is now not only text but a visible button, and the text is changed from "home" to "back". Moreover, the shape of all buttons and sections have now rounded corners. When the users go back, they can either upload a new journal, go to an already uploaded journal, or go to the landing page.

Figure 5.16: Second high-fidelity prototype showing medical journal section

The health measure part is changed a lot as shown in fig. 5.17. The application now explains in more detail what it does while asking the users what they wish to do. It says, "With the help of artificial intelligence, we can gather useful information from your medical journal automatically. Do you wish to do that?" and the users can choose between answering "yes" or "no". It also shows the health measure part in the background of the questions.

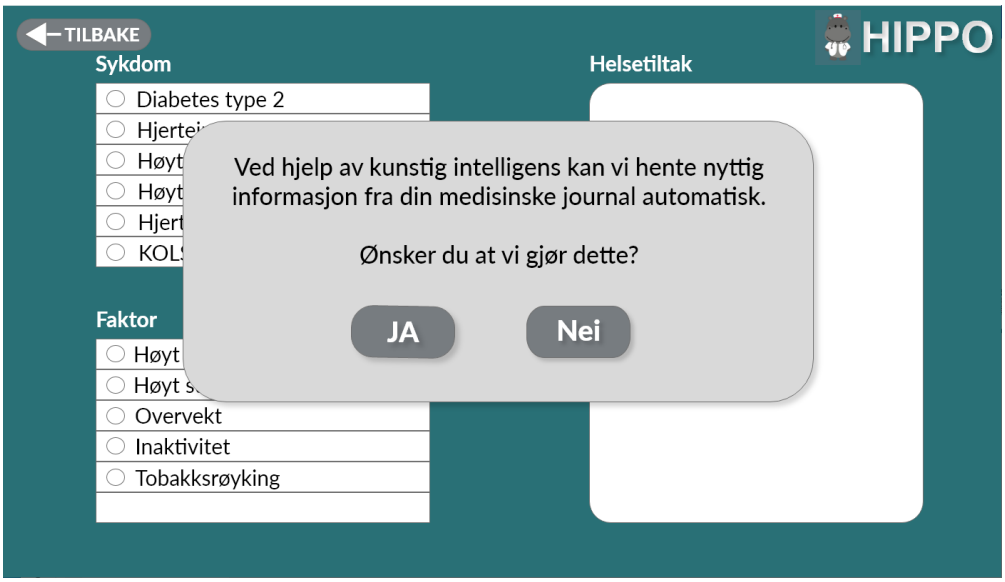


Figure 5.17: Second high-fidelity prototype asks if user wishes to use AI in order to use their data for health measures

Regardless of what the users answer, the application now asks a second question which is "We chose a set of questions that can contribute to more precise health measures. Do you wish to answer these questions?" and the users can choose between "yes" or "no". The presentation of the health measures part has also been slightly modified as shown in fig. 5.18. The title is changed from "efficiency degree" to "health measures". Besides the title, a button with a question mark is now added that explains how the efficiency degree is measured. The color of the question mark button beside the suggested health measures is now changed from white to gray. The corners of the health measure section are now rounded.

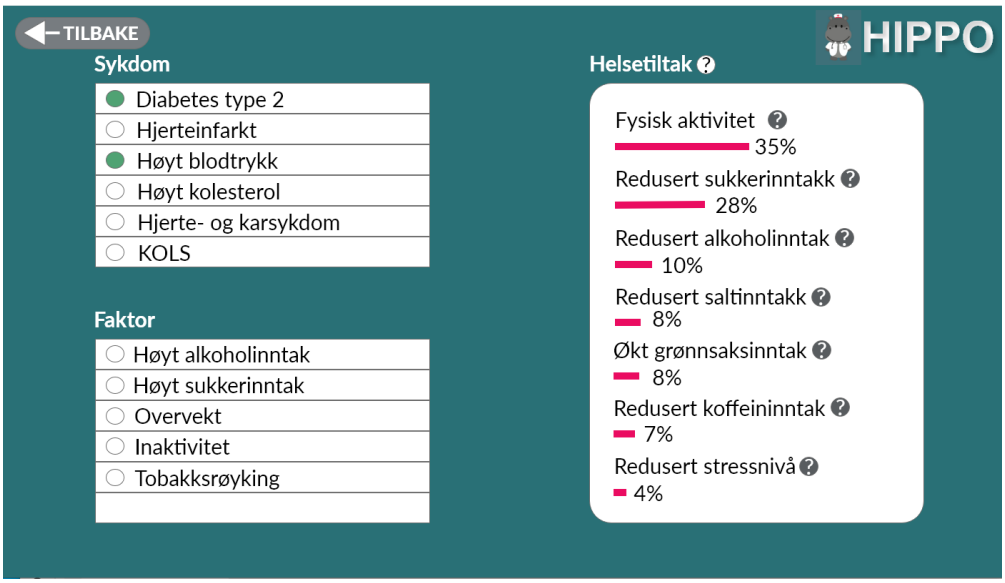


Figure 5.18: Second high-fidelity prototype showing health measure section

5.5.2 Second user test

Figure fig. 5.19 and fig. 5.20 show demographical differences of the users participating in the second user test regarding gender, age, health terminology knowledge, and computer skills. More precise demography can be seen in

For three out of five users the task was very easy to solve, two of the users struggled more. Both of the users that struggled with this task have visual impairment and might have struggled to see the question mark as it was not very visual/easy to spot.

Figure 5.19: User test demography presenting age and gender

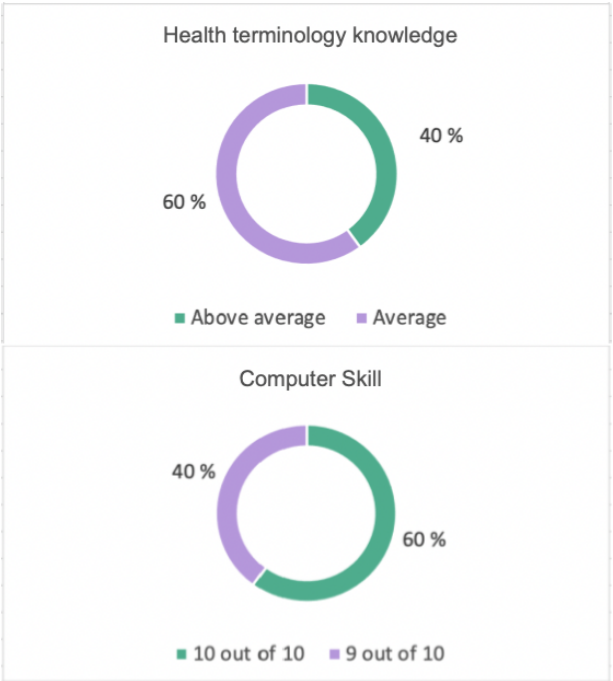


Figure 5.20: User test demography presenting health terminology and computer knowledge

Task 1: *Can you find a summary of your journal?*

The expected way of solving the problem:
Starting on the landing page

1. Select "journal" on the landing page
2. Press “get your journal here” button
3. In the medical journal section. The summary will be presented to the right in the “summary” section

All of the users managed to solve this task without any challenges.

Task 2: *Are there any words in the journal that you don’t understand? Is there any way for you to find out what the words mean?*

The expected way of solving the problem:
Starting in the medical journal section

1. Hover over a medical term one does not understand in the medical journal
2. Click on the medical term that lights on hover
3. The explanation of the word will appear beside the selected word

The majority of the users solved this task without any challenges; one of the users did not hover over the text while reading and did not notice that the text lighted on hover. However, the user spotted this after a short while and managed to solve the task. The user suggested that the words with an explanation should be underlined to show that they contain an explanation.

Task 3: *Can you find information in the journal about diabetes? Is there another way you could find this information?*

The expected way of solving the problem:
Starting in the medical journal section. There are two ways of solving this problem, one can start with either

First option:

1. Click on keywords drop-down placed below the "find in journal" section, to get a selection of searchable keywords
2. Select diabetes from the presented keywords and click on the keyword
3. Select one of the sentences containing diabetes that are now presented below "medical journal" section
4. The information about diabetes containing the selected sentence will be presented below "medical journal" section

Second option:

1. Click (which in this case will simulate input from the keyboard) in the search-bar placed below "find in journal" section

2. Diabetes will appear in the search-bar, click on the "search" button beside the search-bar
3. Select one of the sentences containing diabetes that are now presented below "medical journal" section
4. The information about diabetes containing the selected sentence will be presented below "medical journal" section

All of the users managed to solve this task without any challenges.

Task 4: *Can you find as precise health measures for you as possible?*

Starting on the landing page
The expected way of solving the problem:

1. Select "health measures"
2. Select "yes" on a question to "With the help of artificial intelligence, we can gather useful information from your medical journal automatically. Do you wish to do that?"
3. Select "yes" on a question to "We chose a set of questions that can contribute to more precise health measures. Do you wish to answer these questions"
4. Answer all of the questions presented
5. Select "view measures"
6. The suggested health measures are now presented below "efficiency degree" section

All of the users managed to solve this task without any challenges.

Task 5: *Can you find health measures for a person that smokes and has high cholesterol levels, without knowing anything else about that person?*

The expected way of solving the problem:
Starting on the landing page

1. Select "health measures"
2. Select "no" to the question "With the help of artificial intelligence, we can gather useful information from your medical journal automatically. Do you wish to do that?"
3. Select "no" to the question "We chose a set of questions that can contribute to more precise health measures. Do you wish to answer these questions"
4. Select "tobacco smoking" from the factors and "high cholesterol" from diseases
5. The suggested health measures are now presented below "health measures" section

Two out of three users managed to solve this task. However, three of the remaining users chose to synchronize their journals. The users who chose to synchronize their journals said they did not read the pop-up question precisely enough and thought they agreed to use the AI to get health measures. The part that the AI gathered information from their medical journal was not clear enough.

Task 6: *Can you find out how many active minutes the application means by being "physically active"?*

The expected way of solving the problem:

Starting in the health measure section

1. Click on the question mark placed behind the "physical activity" measure in the health measure section
2. The information will appear below the question mark

All of the users managed to solve this part without any challenges.

Summary of the findings from the second user test

The modifications made to the second prototype have made the application more intuitive.

The journal part was intuitive for the user. Importing the journal in an isolated step and changing the button's name to a more understandable language eliminated the previous confusion and resulted in a hundred percent success. The explanation of the words in the medical journal was very successful this time. One of the users suggested underlining the words that contained an explanation. Searching part in the medical journal was also very successful this time. The users commented that they liked that the corners of the boxes were rounded as this made the application more welcoming.

When it comes to the health measure part, improvements were made. All users managed to get more precise health measures for themselves this time. However, the part with finding health measures for å foreign person that smokes and has high cholesterol was now more clear but still somehow confusing. Even though the pop-ups explained in more detail what the application was doing, the users did not understand the whole message. Feedback from users on this part was that they read this part a bit fast and thought that they agreed to use AI to show health measures but did not understand that the application used AI to get data from their journal. A user suggested making it more apparent when looking for health measures for oneself and others. The health measure presentation was more clear this time. However, it was not entirely clear what the percentage meant, and some users said they were not motivated to do the measures that contained a low percentage. They suggested viewing data differently than percentages, e.g., boxes or a scale like this; important, more important, very important.

Other than that, this prototype received much positive feedback. All of the users thought that the application was intuitive to navigate. That was nothing that made them unsure about how to get to places or what the buttons did. The users claimed that the naming of the buttons was natural and understandable. After testing the prototype, they feel that the application was easy to read and useful and that much work has been done to make the users understand what happens in the application. They liked that there was additional information in the application but that it was tucked away in a good way so one could only see it if needed. The users said that even though the application was a health care application, it was not stressful or difficult to use. They said that the application was explanatory and nice. Furthermore, the hippopotamus helped make the application seem less scary than other healthcare applications sometimes appear to be. Users liked that the application explained things in simple words. The users liked the design, the colors, and the round corners. They liked that the application uses AI.

The users liked the idea of the application and thought that it might be useful, and they might use the application. However, they were not sure how often that would be used since they rarely went to the doctors. Nevertheless, they could see themselves using the application when new medical journals arrived.

Chapter 6

Data

This chapter will describe the different datasets used in the project. It will include the source and pre-processing where relevant. Due to the nature of the project, attempting to both extract any diagnosis from written text, as well as attempting to diagnose different conditions from the extracted data, multiple different datasets are needed.

6.1 NLP

Sourcing text for clinical NLP is a substantial challenge because of the sensitive nature of the data. For this reason, we have largely relied on a combination of biomedical research text and medical transcription samples. We believe the results produced on these datasets will transfer, at least partially, to real clinical text.

6.1.1 Pre-training

The PubMed dataset is a large set of scientific biomedical literature from the PubMed website compiled by Cohan et al. [16]. It consists of abstracts and texts that have been cleaned of non-text information such as figures and tables. It is used in the pre-training of BioBERT and BlueBERT [33][46].

The MIMIC-III dataset is a large collection of anonymized patient data. This includes a number of discharge summaries and radiology and cardiology reports. It has been used for pre-training language models, including BlueBERT and ClinicalBERT. [29][46][5]

The i2b2/VA 2010 workshop consisted of three NLP tasks, one of which was assertion classification. This dataset was used to fine tune the model used for assertion detection. [66][4]

6.1.2 Semantic Similarity

To evaluate semantic similarity, we utilize two datasets specific to biomedical and clinical semantic similarity estimation.

The BioSSES corpus is a dataset for estimation of semantic similarity in biomedical texts. It consists of 100 sentence pairs with similarity ratings ranging from 0 to 4 determined by five separate human experts. The sentences were picked from the dataset produced for the Text Analysis Conference 2014 biomedical summarization track. Preprocessing consisted of converting the data from .docx to .csv format and averaging the five similarity rankings [59].

The MayoSRS Reference Standard consists of 101 clinical term pairs with relatedness annotations ranging from not related (1) to very closely related (4). This set of 101 term pairs were annotated by medical coders, and a subset of 30 pairs were annotated by medical coders

and physicians. The dataset was distributed as three text files, so preprocessing consisted of combining the scores into one file in the .csv format [44].

6.1.3 Summarization

A subset of 200 articles from the PubMed dataset by Cohan et al. [16] was used to evaluate summarization.

MTSamples is a collection of freely available medical transcription samples. They are categorized by medical specialty and consist of a name, description, transcription and keywords. The data is provided as a website(https://www.mtsamples.com/), and a scraped version is available as a Kaggle dataset (https://www.kaggle.com/tboyle10/medicaltranscriptions). The Kaggle version is the one used in this project. For summarization, samples from the General Medicine medical speciality were used.

6.1.4 Assertions

Sentences from the MTSamples dataset was used for assertions as well. Sentences indicating the presence, absence or possibility of a diabetes diagnosis were extracted and manually evaluated.

6.2 Data Analysis

With any given condition having different affecting features, there were two options when it comes to datasets. The first would be a major dataset containing all variables pertaining to a person, this would include both objective, examination, and subjective features, including any disease we would attempt to diagnose. Considering that collecting this data could be a project of itself, the second option is far more viable, namely using multiple datasets. This section will therefore describe the details around the datasets used for this project.

6.2.1 Diabetes

The dataset used for diabetes is originally form the National Institute of Diabetes and Digestive and Kidney diseases. The main limitations of the dataset were that all patients are female, at least 21 years old, and of pima indian heritage [19][58]. The data was a subset of a larger database though we were unable to determine what database, though this could have provided better results, as the inclusion of only one gender could have skewed the results.

Table 6.1: Snippit of original diabetes dataset

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Preprossesing

The diabetes dataset was a well structured csv-file (see Table 6.1) where there were relatively few changes needed. For the purpose of diagnosing patients only the pregnancies feature was removed, though for the purpose of factor importance in relation to the improvement of prognosis, a few variables such as age, pregnancies, diabetes pedigree function (DPF),

and skin thickness were removed as shown in table 6.2. This was due to the nature of the measurement, were the patient would have no way of changing either of the two removed factors.

Table 6.2: Snippit of diabetes dataset after pre-processing

Glucose	Bloodpressure	Insulin	BMI	Outcome
148	72	0	33.6	1
85	66	0	26.6	0
183	64	0	23.3	1
89	66	94	28.1	0
137	40	168	43.1	1

6.2.2 Obesity

The obesity dataset was downloaded again from kaggle [20], though it was originally created by Palechor et al. [43], 23% of the dataset entries are collected for real patients with remaining 77% of the dataset entries are synthetic data. Synthetic data is generated from the real data, but not belonging to real people. The complete dataset has a total up to 2111 entries. The obesity dataset needed, opposed to the diabetes dataset, signifficant pre-processing. Predicting if a person is obese was not something that needed AI to function, as there is a well established formula to calculate it. However, the dataset contained many lifestyle features, such as time spent using technology, or physical activity frequency. Due to the amount of features, the table sample was transposed to fit in the report.

Table 6.3: Snippit with 3 samples from the original obesity dataset

gender	Female	Female	Male
age	21	21	23
height	1.62	1.52	1.8
weight	64	56	77
fhwo	yes	yes	yes
FAVC	no	no	no
FCVC	2	3	2
NCP	3	3	3
CAEC	Sometimes	Sometimes	Sometimes
SMOKE	no	yes	no
CH2O	2	3	2
SCC	no	yes	no
FAF	0	3	2
TUE	1	0	1
CALC	no	Sometimes	Frequently
MTRANS	Public_Transportation	Public_Transportation	Public_Transportation
NObeyesdad	Normal_Weight	Normal_Weight	Normal_Weight

See Abbreviations for an explanation of the terms used in table 6.3.

Pre-processing

Changing the target from obesity to a BMI prediction from a set of lifestyle features or subjective features made it possible to determine the importance of each individual user controlled action. The obesity dataset would not be used for diagnosis but rather as a tool to extract the importance of certain actions. Using the same reasoning as the diabetes dataset, the objective features that was out of the patients direct controll was removed,

such as family history, age and gender. Height and weight was combined to BMI using the formula: $BMI = kg/m^2$, there after set at the new target of the dataset. This left the data as seen in table 6.4.

Table 6.4: Snippit with 3 samples from the obesity dataset after pre-processing

fcv	fchc	nmm	CAEC	CH20	scc	faf	tue	CALC	mtrans	bmi
0	3.0	1	0	2.0	0	0.0	1.0	0	2	24.39
0	3.0	1	1	3.0	1	3.0	0.0	1	2	24.24
0	3.0	1	0	2.0	0	2.0	1.0	3	2	23.77

6.2.3 Heart Condition

The cardiovascular disease dataset was chosen simply for its size, containing 70 000 records, the downside being the lack of a source. The data was found on kaggle[65], but without any further sourcing, the data is not verifiable. The approach for the dataset remains similar to the previously mentioned sets. With the dataset containing subjective features such as

Table 6.5: 5 sample snippit of the original heart condition dataset

age	gender	height	weight	ap_hi	ap_lo	chol	gluc	smoke	alco	active	cardio
18393	2	168	62.0	110	80	1	1	0	0	1	0
20228	1	156	85.0	140	90	3	1	0	0	1	1
18857	1	165	64.0	130	70	3	1	0	0	0	1
17623	2	169	82.0	150	100	1	1	0	0	1	1
17474	1	156	56.0	100	60	1	1	0	0	0	0

if a person is active (given as a binary measure), the set has a wide range of how accurate data is when comparing to the age of a patient being provided in days to the subjective measurements.

The target of the dataset described as **cardio** is a binary measurement of the patient having been diagnosed with any cardiovascular disease.

Pre-processing

The heart condition dataset had the highest rate of change, with multiple variables being removed or changed. The measurements of cholesterol ($chol$) and glucose ($gluc$) were removed due to the possibility of inconsistency in its measurement. Systolic blood pressure (ap_hi) and diastolic blood pressure (ap_low) were combined to create a single index as described by Franklin et al. [23]. Height and weight were combined as BMI. Finally, age was changed to years instead of days, with the formula $age = floor(days/365)$.

Table 6.6: 3 sample snippit of heart condition dataset post pre-processing

bmi	bp	cholesterol	gluc	smoke	alco	active	cardio
21.9671201814059	190	1	1	0	0	1	0
34.927679158448385	230	3	1	0	0	1	1
23.507805325987146	200	3	1	0	0	0	1

Chapter 7

Results

The artefact, i.e. the designed and implemented prototype, has not been evaluated as a whole, due to the different objectives and evaluation criteria of each module. The evaluation results will be split into categories referring to their corresponding solution section in chapter 5. The evaluation criteria and method(s) can be found in chapter 3.

7.1 NLP

The semantic similarity methods were evaluated against two datasets, BIOSSES and MayoSRS, as presented in section 6.1.2.

Table 7.1: Spearman correlation for BIOSSES and MayoSRS using MiniLM-, BioBERT- and BlueBERT-based models

Model	BIOSSES	MayoSRS
MiniLM	0.8125	0.4555
BioBERT	0.8163	0.4979
BlueBERT	0.8407	0.6583

As expected, the general use model based on MiniLM performed the worst. The BioBERT model performed slightly better, and the BluBERT performed significantly better. It performed especially well on the MayoSRS dataset, which is likely due to it being trained on clinical text data.

The summarization methods were evaluated against two datasets,PubMed and MTSamples, as presented in section 6.1.3.

Table 7.2: ROUGE F1 scores using MiniLM-, BioBERT- and BlueBERT-based models

Model	PubMed			MTSamples		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
MiniLM	0.3812	0.1574	0.3430	0.1735	0.0896	0.1676
BioBERT	0.3850	0.1624	0.3465	0.1661	0.0860	0.1606
BlueBERT	0.3830	0.1578	0.3451	0.1635	0.0898	0.1578

The assertion model was evaluated against 100 sentences from the MTSamples dataset relating to diabetes, with human evaluation performed by us.

	True Positive	True Negative
Predicted Positive	74	16
Predicted Negative	0	10

The model was able to correctly classify all sentences in which the presence of a diabetes diagnosis was explicitly denied. It incorrectly predicted a positive diagnosis for sentences that asserted a family history of diabetes.

7.2 Diagnosis

The accuracy of the random forest diagnosis was as shown in table 7.3b at 0.73, with heart condition falling a little behind with 0.65 (shown in table 7.4b). The diagnosis of the two example conditions provided expected results, with no exceptional scores. The focus when considering the confusion matrix, as show in table 7.3a and table 7.4a, was the reduction of false negatives(Predicted negatives, but the prediction is incorrect). As stated in section 5.3, false positives are not necessarily a negative thing, as we can extrapolate from the positive diagnosis that a patient either has, or is at risk of getting a condition. This point again enhances the importance of reducing the number of false negatives, as it could lead to the completely opposite effect of what is desired.

Table 7.3: Classification results for the Diabetes diagnosis

	True diagnosis		
	Positive	Negative	Total
Positive	164	30	194
Negative	53	60	113
Total	213	94	307

a: Confusion matrix for diabetes diagnosis

	precision	recall	f1-score	support
Positive	0.76	0.85	0.80	194
Negative	0.67	0.53	0.59	113
Accuracy			0.73	

b: Classification report for diabetes diagnosis

Table 7.4: Result for the heart condition classification

	True diagnosis		
	Positive	Negative	Total
Positive	1326	636	1962
Negative	699	1312	2011
Total	2025	1948	3973

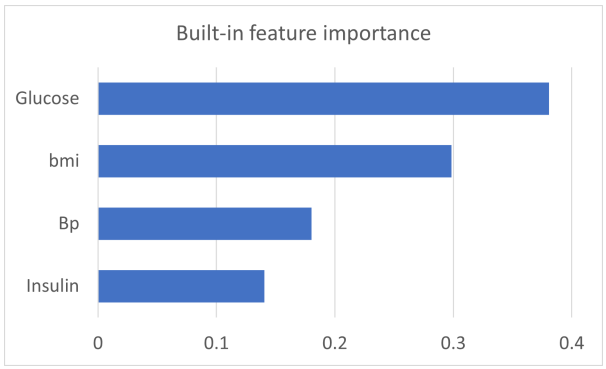
a: Confusion matrix for the heart condition diagnosis.

	precision	recall	f1-score	support
Positive	0.65	0.68	0.67	1962
Negative	0.67	0.65	0.66	2011
Accuracy			0.73	

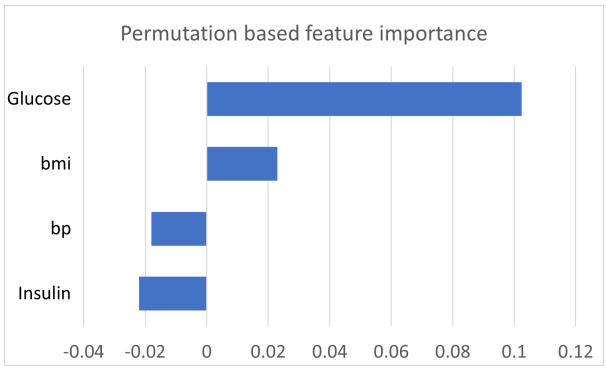
b: Classification report for heart condition classification.

7.3 Feature importance

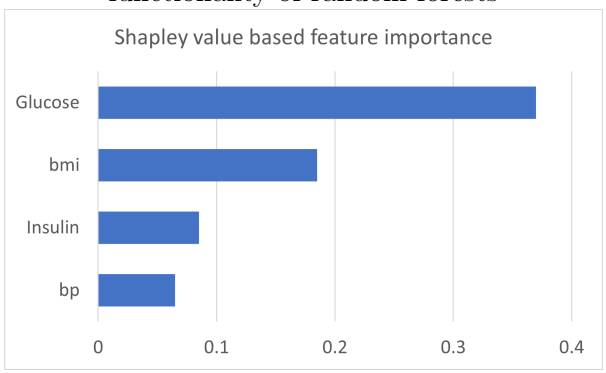
The importance of any arbitrary feature is extracted from the random forest, and by nature this importance is accurate in terms of deciding the classification. The feature importance is calculated using the three methods mentioned in section 2.3.2. The three different feature importance methods have similar results, with minor variations. As this is the case, the Gini importance method was used as it did not require any separate module to be installed for the final prototype. Both the permutation-based and shapley-based feature importance were down-prioritized due to their heavy computational requirements. From existing studies, we know that there may be causation in the data presented; weather it is direct or indirect causation, is discussed further in chapter 8.



a: Feature importance produced by the built in functionality of random forests



b: Results of permutation based feature importance



c: Results of Shapley based feature importance

Figure 7.1: Results of the different feature importance methods

7.4 User test results

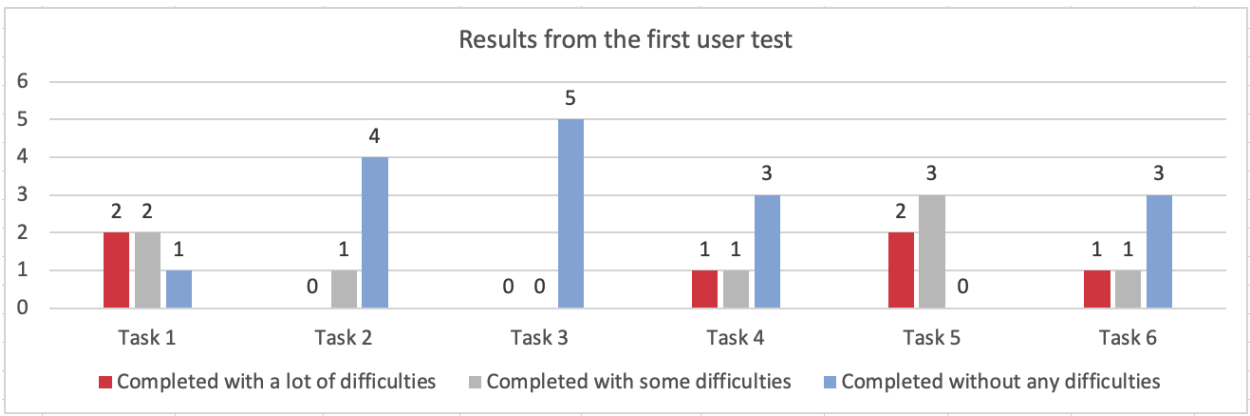


Figure 7.2: Results from the first user test of the high fidelity prototype

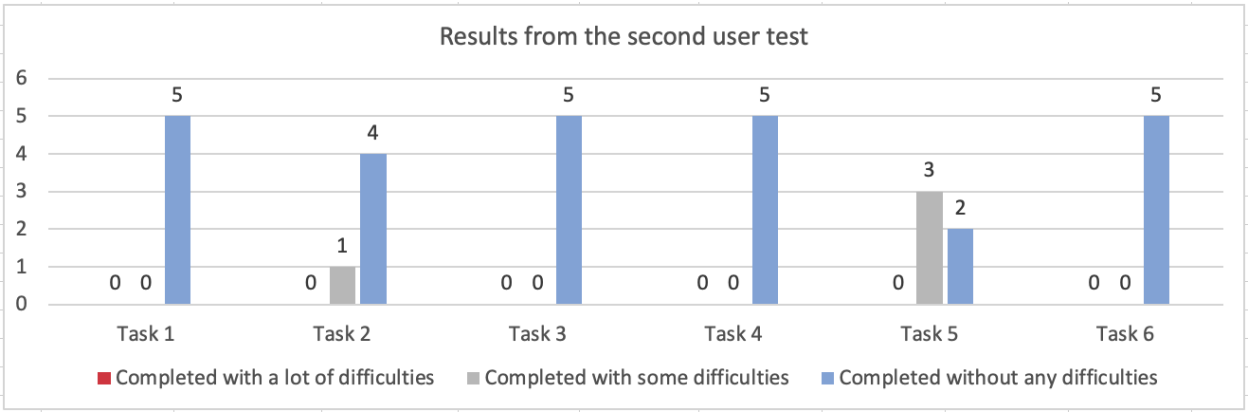


Figure 7.3: Results from the second user test of the high fidelity prototype

The usability of the prototype significantly improved in the second iteration, as seen in fig. 7.3 that shows the test results from the second user tests compared to fig. 7.2 showing the test results from the first user tests. The changes made to the second prototype eliminated previous confusion in the first task, resulting in all users completing the task without any difficulties. The results from the second and third tasks remained the same as non-significant changes were made to the functionality needed to complete these tasks. All five users completed the fourth task without any difficulties in the second prototype compared to the first prototype, where two of the users struggled, and one of them did not manage to complete the task without help from the test leader. The fifth task was the most challenging one. While testing the first prototype, non of the five users managed to solve this task as intended, and two of the users did not manage to complete the task without help from the test leader. The results from the second user test show improvements; two of the users managed to solve the task without any difficulties while three of the users experienced some difficulties. The sixth and last task resulted in all the users managing to complete the task without any difficulties in the second user test, compared to the first user test where two of the users struggled and did not manage to complete the task without help.

Chapter 8

Discussion

The aim of this study was to investigate the viability of an automated system for personalized health advice. The proposed system has two main objectives: to develop and evaluate tools for increased comprehension of medical journals and present personalized health advice based on the automatic analysis of medical journals and the integration of information provided by the user. We have investigated whether these two objectives could be met using contemporary machine learning methods. We have employed various design principles to implement user needs and system requirements and create and evaluate a user interface for an application prototype integrating these tools with usability, another primary design goal.

Our overall findings, based on the results presented in chapter 7, are that the individual parts of the proposed solution produce positive results, though the feature importance could not be proven in a clinical setting.

With the subject of the project being health data and personal medical journals, it was not easy to find data that could be used. This was relevant for both the NLP for interpretation and the data needed for the ML to provide feedback. The data needed to be as accurate as possible, though, without any way to check it, the current results of the feedback module are improbable at best.

An example can be seen in fig. 5.3 with the blood pressure, noted as ap_hi and ap_low. There are datasets available on Kaggle that contain a blood pressure feature. However, the problem comes back when those datasets only contain two features that can be directly tied to lifestyle choices, namely BMI and smoking. This would likely result in the action of reducing BMI having abnormally high importance and many other smaller features being ignored due to low percentages and no clear action tied to them.

The rest of this chapter will seek to answer the research questions presented in the chapter (1), outline the limitations of our study, and suggest future research.

8.1 Research questions

RQ1.1: What issues do patients have when trying to understand medical documents, and are there NLP methods that can alleviate these issues?

Using a combination of interviews and user tests, we found that the users we spoke to faced significant problems when trying to understand medical documents. Interview subjects with different backgrounds faced similar problems, which indicates that their problems are not unique and that a solution would be helpful for the greater medical community.

Some recurring complaints among participants were issues with large amounts of dense text, and difficulty understanding technical terminology. Several participants also stated that they

felt they did not receive the information they wanted from their primary healthcare provider. From the feedback we received when presenting our prototypes and performing user tests, we believe using a combination of automated summarization, intuitive text searching, and explanation of medical terms could alleviate these issues.

RQ1.1:

Can current state-of-the-art Natural Language Processing (NLP) models be used effectively in combination to realize the functional requirements of our system?

From our research, we learned that the field of biomedical and clinical NLP has progressed significantly in the last few years, to the point where we believe useful automated solutions are a possibility. Additionally, the effective use of pre-trained models that are fine tuned on downstream tasks shows that these methods can be utilized in field with limited amounts of data, such as the Norwegian healthcare system.

The semantic similarity results on the BIOSSES dataset show that the model predictions correlate highly with human evaluation. This means that searches performed using these models are likely to select results that are relevant to the user. The results on the MayoSRS, although not as good as the BIOSSES results, are still quite good. The improved performance by the model trained on clinical data indicates that models with additional training for a specific domain can perform significantly better.

The results for summarization on the PubMed dataset were good considering the simplicity of the implementation and that it produces extractive summaries with full sentences. The performance on the MTSamples dataset was subpar, though we believe that to be a misfit between the descriptions from the dataset and our method of summarization. Sentence extraction for such short texts and summaries are bound to perform poorly, and a word based summarization method or keyword extraction should be considered.

Results for assertions were promising. The sentences are skewed towards a positive prediction as only sentences containing the word diabetes were used. This means that a model with only positive predictions would still do relatively well. In the case of this model, it was able to correctly identify negative assertions about the presence of a diagnosis, showing that it is able to distinguish between positive and negative assertions. We believe the next step in the development of the model should focus on distinguishing between assertions about the patient and assertions about someone else.

We were unable to formally evaluate the NER solution because of time constraints, but found the subjective analysis of our solution to be promising.

RQ2

Can specific and significant correlations between health features and medical diagnosis be found using Machine Learning techniques?

This research question was answered early in the project period because we are using data-driven machine learning, and again proved by the results of the random forest, indicating that correlation was prevalent.

The use of a random forest for diagnosis was not necessarily the ideal choice of technology, as improved results would require further fine-tuning, and a more accurate diagnosis could be achieved with different machine learning algorithms, such as the counterfactual algorithm described by Richens et al. [53]. The diagnosis and causation features should be separated for future developments, as the workflows are vastly different. This is further expanded on in *RQ2.1*.

Diving deeper, we can again show a correlation between lifestyle measurements, such as FAVC, NCP, and FAF (see *Abbreviations*). However, part of the focus of the developed application is weighing the importance of a given lifestyle choice. This again will be expanded further in *RQ2.1*.

RQ2.1

Can we extrapolate causation from correlation provided from a machine learning algorithm?

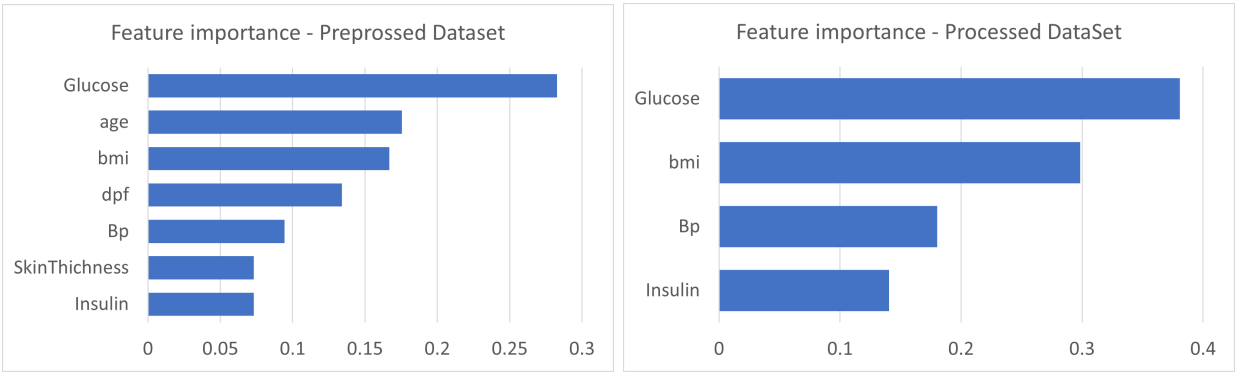
It is fairly proven that correlation does not equate to causation, with websites (such as https://www.tylervigen.com/spurious-correlations) being created to show some of the more humorous examples of correlation. A relevant example is BMI and smoking. Smoking has been shown to cause weight loss[15], reducing BMI. As a result, BMI could be used to classify if a person smokes, but increasing the BMI would not necessarily change the predicted outcome of smoking. This example would be a direct correlation but reverse causation.

Further expanding the example of smoking by mentioning the effects of smoking on cholesterol. There is a causation effect of smoking on BMI, and smoking on cholesterol has been shown. This leads to an indirect correlation between BMI and cholesterol, though there is no causation.

We assumed that finding causation from correlation could be solved by splitting the classification from the factor importance. Training two different models, one for classification and the other for extracting the importance of a feature in terms of causality. This required the splitting of datasets manually where causality was already known, again leading to the point that correlation does not equal causation.

It can be seen when comparing the feature importance of the diabetes dataset before and after the removal of measures that were not a direct cause of the condition that the DPF can be used as a classification measure as shown in fig. 8.1a, though it has neither a direct nor an indirect causal relation to the condition. As a side note, the works of Richens et al. [53] show that the use of causality can lead to increased accuracy in the diagnosis of a condition, though in their case, the importance of each feature was known and inserted beforehand.

To summarize the answer, we could not extract causality from causation, though it is to extract the importance of known causal relationships in relation to how important a feature is in classifying a condition. Such as whether it is more important to work out or have healthy nutrition to reduce BMI or maintain a healthy BMI. If the feature importance in classifying translates to clinical importance to reduce the severity of the condition would require a more extensive study.



a: Feature importance of a full diabetes dataset with all factors present

b: Feature importance of a limited diabetes dataset with all mutable features present

Figure 8.1: Feature importance extracted from a random forest with the same diabetes dataset

RQ3

How can Human-Centered Design increase usability of a Personalized Decision Support System?

User tests were used to evaluate the design solution of a high-fidelity prototype. The first high-fidelity prototype was designed based on the feedback from users received from both the interviews and testing of the paper prototype. The results from the user test of the first high fidelity prototype provided improvements to the second prototype. Five users tested each prototype. The user test consisted of an opening interview, observation, and a closing interview. The opening interview showed the demographical distribution between the users. The demographic distribution differed between the first and second user test. As seen in fig and fig. 5.14, fig. 5.13 and fig. 5.20, fig. 5.19, the first user test was better distributed regarding age and knowledge of medical terms. However, during the user tests' observations, the users who claimed that they had lower computer skills or lower knowledge of medical terms did not struggle with solving the tasks any more than those who said they were highly skilled. Furthermore, the observations and the closing interviews helped determine what parts of the applications were intuitive to the users and what parts were challenging. The feedbacks obtained from the test were used to improve the second prototype. The results from testing the first prototype can be seen in fig. 7.2. The modified version of the prototype obtained the following results fig. 7.3. Testing the second prototype clearly shows that the users had a higher success rate while testing the modified version. The feedbacks from the second user test seen in appendix F combined with test results for the second prototype shows that the changes made thru the iterative process of the HCD increased the usability of a Personalized Decision Support System.

RQ3.1

How can Human-Centered Design help increase understanding of medical terms and documents?

The iterative process of HCD was influenced by the user feedback throughout the whole design phase. This process increased the usability of the designed product as the users tailored the design to their needs. In the beginning stage of the design, users, both health care professionals and laypeople who participated in the interviews, helped by identifying the challenges and needs associated with medical terms and documents. The interview subjects struggled with understanding their medical documents due to difficult medical terms and had to use additional tools to understand them, which tends to be frustrating and inefficient. Testing the paper prototype helped to guide the application's design in the right direction. Users that studied the paper prototype suggested that the application could show a summary of the journal and that it would be helpful to have the ability to see the medical journal and search through it using keywords. Users advised making the design more understandable by adding more descriptive buttons and elements. High fidelity prototype was designed based on feedback from both the interviews and the paper prototype. In the medical journal section of the prototype, the users were able to get an explanation of difficult medical terms found in their journal by hovering over medical terms and clicking on the words. This function contributed to understanding the content of the medical journals without having to use additional resources. Furthermore, the presented summary of the medical journal provided the users with a short description of what the medical journal contained. The searchable functions in the prototype made the searching process for specific information more accessible. Feedback received after the testing was positive; the users especially liked the explanation of difficult medical terms.

RQ3.2

How to illustrate personalized advice and recommendations in a user friendly matter?

In the beginning stage of the design process, the users who participated in the interviews were asked what kind of feedback they preferred regarding health measures needed to improve a specific condition. The majority of the users answered that they preferred to receive both specific feedback and explain why they should do the suggested measures. The final design of the prototype presents health measures sorted in an order which shows the measure with the highest efficiency percentage at the top and the one with the lowest efficiency percentage at the bottom of the list. The percentages are accompanied by a bar chart, providing a visual representation for users who were inclined to a lower numerical literacy. The suggested health measures meet both requirements suggested by the users in the interviews; the users are presented with specific health measures and can see more descriptive information about each measure by pressing the button placed beside each measure.

8.2 Decisions

Trough out the project multiple different avenues were explored to determine what technology would be used to gain the desired results. This section will outline the other options, and the reason they were discarded.

8.2.1 Causal inference

The initial idea of the project was the use of causal inference to determine feature importance for patient feedback. While gathering information on the subject, it seemed that causal inference were set in a Boolean setting. This meant that it would not fit the input, as for example blood pressure has more states then high or low. After attempting to create a model of the causality for the diabetes dataset (see fig. 8.2), the map proved difficult to understand, not really improving on the input data we had already received. This led to the dismissal of the causal inference in search of a model with easier to understand outputs.

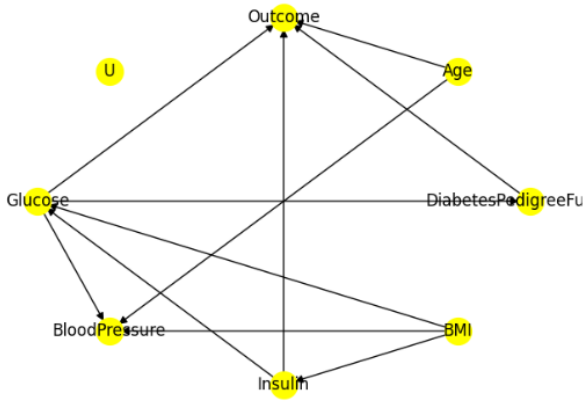


Figure 8.2: Causal model of the Diabetes dataset.

8.2.2 Factor analysis

Factor analysis is the analysis of observed variables to determine the causation of an unknown variable in relation to the observed variables. This seemed relevant as we are able to reduce the number of factors, therefore increasing the relevance of each feature. This was, though,

irrelevant in this case as we can not include external unobserved variables due to the problem with correlation not equalling causality presented in *RQ2.1* in section 8.1.

The factor analysis could still be used as a technique to reduce the number of factors. But with the goal being so provide support in decision making, we still want to provide all the available information, but in a more understandable format.

8.3 Challenges

The challenges section will present the non-technical challenges that were experienced throughout the project period. This is not an extensive list, but rather a summary of the most impact-full challenges.

8.3.1 Covid-19

With the pandemic coming to what seemed to be an ending, it was still a challenge navigating through social interactions without the risk of spreading the virus. This was especially evident when two group members caught the virus and were in quarantine for a week each. This was not the final problem with Covid-19 as we had test subjects cancel their participation in last minute due to Covid-19, and meetings had to be rescheduled.

Chapter 9

Conclusion and Future Work

This thesis explored the possibility of using natural language processing, and the use of machine learning (particularly random forests), in a human centered design of a personal decision support system. The evaluation is based on the Key performance indicators mentioned in section 3.5. The diagnosis showed an accuracy of 73%, though with focus on recall, the diagnosis did not show an improved recall over other diagnosis methods. The feature importance extraction only had the evaluation criteria of comparing the different methods of determination. The results of which showed a low variance in the importance of a set of features. We did see Gini importance having a major advantage in computational speed.

The design of the application had the principles of universal design in focus. The usability testing in the iterative HCD process showed an increased success in task completion with less problems while solving the tasks. The feedback from users often fit directly into the different principles of design. The current prototype did no receive any changes after the final evaluation phase of HCD, as such at least one more iteration would be desired.

The Natural Language Processing part of the project showed promising results, with a high correlation between model predictions and human-evaluated sentence similarity, a relatively high ROUGE score for summarization, and a high level of accuracy in assertion prediction. We were able to produce these results using fine tuned pre-trained models, which indicates that these methods can be used in situations with limited data.

9.1 Future work

Further iterations of the design prototype and feature importance analysis should be produced, with testing of the iterations using the known evaluation criteria, as well as in a clinical setting where the impact of the software can be proven or dis-proven. Further work on the prototypes should include:

- Changing the health measure part so while selected, the users will find themselves directly in the health measure section without answering any questions about more precise health measures. Furthermore, add a button that allows the users to get information from their journals and another button to answer additional questions for more precise health measures.
- Restructuring the health measure section to clarify when the user can find information for themselves and others
- Clarifying that the application uses AI to get health information from the medical journal

- Restructuring the presentation of health measures; potentially present the effectiveness of the health measures in boxes instead of showing them as a percentage
- Perform a new user test and evaluation of the improved prototype. After resolving the previous issues with the prototype, a front-end application would be developed based on the design of the newest prototype.
- Add datasets more noncommunicable lifestyle diseases.
- Add datasets connecting lifestyle choices with variable in the datasets for noncommunicable lifestyle diseases.

Appendix A

Paperprototype feedback

Landing page	General
- Explain what the summary is doing - I don't think that the summary should be the most significant element - Why is the summary there? As it is not the most important element it should not take up the most space	- Have everything in Norwegian - Not very intuitive. I wish that it was more explanations - Miss a good front page - More descriptive buttons

Journal Extension	Decision System
- Log in to Helsenorge to retrieve the medical record from there. (A user had over 100 pages in the medical journal) - Not intuitive, I would like to see more visible buttons. - What are we supposed to upload? Describe more - What happens after the keywords are found? - Why are the keywords displayed? - Possibility to log in? How do I access the journal? - Wants a visible button to upload the journal - What is meant by "Find"? What can you find? - I did not like that there were only things in the middle of the page; they should have been better distributed. - It makes more sense to upload the journal and then see the journal itself and filter things there, and then possibly get keywords. - Have the opportunity to filter the journal on words/date/diseases - Why should you get keywords when you see them on the front page anyway?	- Should have questions that one has to answer to get more specific health measures - Give the user choice if the user wants to fill in a lot of information or not. - Give the user an option to skip questions the user doesn't want to answer. - The options that are selected by synchronizing the journal should appear at the top so the user can easily see them and remove them as they want - Should be able to filter the health measures based on the disease - Specify what is meant by physical activity and other measures - Not very descriptive name for decision system

Appendix B

Consent form

Vil du delta i forskningsprosjektet HIPPO: et PDSS prosjekt

Dette er et spørsmål til deg om å delta i et forskningsprosjekt hvor formålet er å bygge opp data for utvikling og evaluering av prosjektet "PDSS(personalized decision support system) using NLP(Natural language processing) and CIM(Cause inference modelling)". I dette skrivet gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formål
Vi ønsker å gjøre det lettere å få en forståelse for ens egen medisinske status gjennom automatisert analyse og oppsummering av ens egen medisinske journal. Videre vil vi automatisk foreslå steg som kan bli tatt for å forbedre ens egen helse og velvære som er egnet til brukerens egen helsesituasjon. Vi vil oppnå disse målene ved å lage et PDSS (personalized decision support system). Data som blir brukt i prosjektet vil bestå av både offentlig tilgjengelig anonymisert data, og noe kunstig generert data dersom det skulle være nødvendig. Første del, analyse og oppsummering, vil bli gjort ved bruk av NLP (Natural Language Processing). Andre del, foreslå steg som kan bli tatt for å utbedre ens egen helse, vil bli gjennomført ved å inferere kausale sammenhenger. Brukerdata som er planlagt å samle inn er rettet mot behovsvurdering, for å finne ut om sluttproduktet vil være nyttig i dagens helsefag. Det vil også bli samlet data via brukertest, for utbedring av front-end design.

Hvem er ansvarlig for forskningsprosjektet?
Martin Wulf Gerdes ved Universitetet i Agder er ansvarlig for prosjektet.

Hvorfor får du spørsmål om å delta?
Målet vårt med deltakere er å ha et variert utvalg av deltakere med vekt på ulik aldersgruppe, kjønn. Samt en eller flere som jobber innen helse. Det ville være gunstig med noen deltakere som har noe tidligere erfaring ved lesning av medisinske journaler, samt noen deltakere med kroniske sykdommer. Du har blitt spurt om å delta fordi du oppfyller et eller flere av kravene nevnt ovenfor.

Hva innebærer det for deg å delta?

Datainnsamling i prosjektet er delt i to deler: et intervju for å kartlegge potensielle problemer produktet vårt kan bidra til å løse og en brukertest for å teste produktet vi lager, basert på opplysningene vi samler i intervjuet. Det er valgfritt om du ønsker å delta i en eller begge delene.

Dersom du velger å delta i intervjuet, innebærer dette et intervju på ca. 20 minutter, hvor lydopptak og notater blir brukt som referat av intervjuet. Lydopptak og notater blir lagret

elektronisk. Under intervjuet vil deltagere bli spurt om tidligere erfaring med lesing av medisinske journaler, e-mail, navn, alder og tidligere erfaring med helsetjenesten. Det vil ikke være nødvendig å oppgi personlig helseopplysningere. Alle spørsmål er valgfrie å svare på. Dersom du ønsker å delta på brukertesten innebærer dette et kort intervju før testen, en brukertest hvor deltakerne blir bedt om å utføre ulike oppgaver på en datamaskin, samt et kort intervju etter brukertesten. Dette vil foregå på Universitetets i Agder usability-lab. Under intervjuene blir det tatt videoopptak samt elektroniske notater. Under brukertesten foretas det lydopptak, skjermopptak og videoopptak samt elektroniske notater. Alle spørsmål vil være valgfrie å delta på.

Det er frivillig å delta
Det er frivillig å delta i prosjektet. Hvis du velger å delta, kan du når som helst trekke samtykket tilbake uten å oppgi noen grunn. Alle dine personopplysninger vil da bli slettet. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Du kan til enhver tid avslutte intervju og/eller brukertest uten å oppgi begrunnelse for dette.

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger
Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket.

- All data vil bli lagret kryptert.
- Data vil kun være tilgjengelig for prosjektgruppen, samt prosjektveileder.
- Navnet og kontaktopplysningene dine vil jeg erstatte med en kode som lagres på egen navneliste adskilt fra øvrige data

Deltakere vil ikke kunne gjenkjennes ved publikasjon, da data vil bli brukt som anonymisert, bearbeidet data og statistikk.

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?
Opplysningene anonymiseres når prosjektet avsluttes/oppgaven er godkjent, noe som etter planen er 20 juni 2022. All innsamlet data vil da slettes.

Hva gir oss rett til å behandle personopplysninger om deg?
Vi behandler opplysninger om deg basert på ditt samtykke.

På oppdrag fra Universitetet i Agder har Personverntjenester vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Dine rettigheter
Så lenge du kan identifiseres i datamaterialet, har du rett til:

- innsyn i hvilke opplysninger vi behandler om deg, og å få utlevert en kopi av opplysningene
- å få rettet opplysninger om deg som er feil eller misvisende
- å få slettet personopplysninger om deg
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger

Hvis du har spørsmål til studien, eller ønsker å vite mer om eller benytte deg av dine rettigheter, ta kontakt med:

- Universitetet i Agder ved Martin Wulf Gerdes (martin.gerdes@uia.no), eller Vetle Ørstavik Hollund (vetlehoo@gmail.com)
- Vårt personvernombud: Personvernombud@uia.no

Hvis du har spørsmål knyttet til Personverntjenester sin vurdering av prosjektet, kan du ta kontakt med:

- Personverntjenester på epost (personverntjenester@sikt.no) eller på telefon: 532 11 500.

Med vennlig hilsen

Martin Wulf Gerdes (Forsker/veileder)	Vetle Ørstavik Hollund, Marta Hanczarek, Ferdinand Løberg (Prosjektgruppe)

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet Hippo, og har fått anledning til å stille spørsmål. Jeg samtykker til:

☐ å delta i intervju

☐ å delta i brukertest

Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet

.....

(Signert av prosjektdeltaker, dato)

Appendix C

Usability test 1

Questions	User 1	User 2	User 3
How familiar are you with health concepts, what they are, and what they represent?	A little, not educated, personal experience	Very familiar, I have an education in health	Above average, I have an interest in health
How often do you use a computer, what system, and how skilled would you say you are?	Daily, Windows and iOS highly skilled	Daily, Windows, above average	Daily, Windows, above average
How skilled would you say you are on a scale from 1 to 10?	10	6-7	5
Age range (18-25, 26-35, 36-45, 46-55, 56-65, 66+)	18-25	46-55	26-35
Do you work, or have you worked in the health care system?	No	No	Yes, at a nursing home
Do you have reduced vision or any form of color blindness, mobility difficulties, dyslexia, ADHD, or other atypical neurology?	No	Some visual impariment	No
Gender	Male	Female	Female

Questions	User 4	User 5
How familiar are you with health concepts, what they are, and what they represent?	Above average, learned as needed	Above average, I have an interest in health
How often do you use a computer, what system, and how skilled would you say you are?	Daily, Window and iOS highly skilled	Daily, Windows, above average
How skilled would you say you are on a scale from 1 to 10?	7-8	8
Age range (18-25, 26-35, 36-45, 46-55, 56-65, 66+)	56-65	18-25
Do you work, or have you worked in the health care system?	No	Indirectly, worked with health professionals as a secretary/receptionist
Do you have reduced vision or any form of color blindness, mobility difficulties, dyslexia, ADHD, or other atypical neurology?	Some visual impairment	No
Gender	Male	Female

Appendix D

Usability test 2

Questions	User 6	User 7	User 8
How familiar are you with health concepts, what they are, and what they represent?	Average	Average	Above average, I have an interest in health
How often do you use a computer, what system, and how skilled would you say you are?	Daily, Windows , highly skilled	Daily, Windows and iOS, above average	Daily, iOS, highly skilled
How skilled would you say you are on a scale from 1 to 10?	10	9	10
Age range (18-25, 26-35, 36-45, 46-55, 56-65, 66+)	18-25	18-25	26-35
Do you work, or have you worked in the health care system?	No	No	No
Do you have reduced vision or any form of color blindness, mobility difficulties, dyslexia, ADHD, or other atypical neurology?	Light version of dyslexia	No	Some visual impairment
Gender	Male	Female	Female

Questions	User 9	User 10
How familiar are you with health concepts, what they are, and what they represent?	Average	Above average
How often do you use a computer, what system, and how skilled would you say you are?	Daily, Windows, Linux and iOS highly skilled	Daily, Windows, above average
How skilled would you say you are on a scale from 1 to 10?	10	9
Age range (18-25, 26-35, 36-45, 46-55, 56-65, 66+)	45-55	26-35
Do you work, or have you worked in the health care system?	No	Yes
Do you have reduced vision or any form of color blindness, mobility difficulties, dyslexia, ADHD, or other atypical neurology?	Some visual impairment	No
Gender	Male	Male

Appendix E

Closing interviews usability test 1

Below are the results of the closing interviews of the first usability test.

Questions	User 1
What do you think the percentage means?	How much the suggested measures influence the diseases or factors I have chosen.
Was there anything that stood out as more challenging during the test?	- It was not intuitive to upload the journal in the Journal part. - The health measure part was also difficult to understand. I did not understand that you had to import your journal to get more specific health measures.
Was it intuitive to navigate the application?	-Most places it was. - I did not understand that by skipping to answer question you could go directly to selecting your own factors. I would call it something else then just "skip". - When it comes to the journal, if no journal was imported before. Then the import could be an isolated step.
What feelings are you left with after using the application?	- It's very square, not sure if it is positive or negative. - Text is a bit messy when you import the text, the centered text is harder to read. Maybe add more margins as well. - When deleting the journal, you might have a separate button at the bottom.
Would you use the application again?	- If I had needed it then yes. - I was told that my journal is run thru an AI, it would have been fun to see. - It is a medicine tool, so I would like to use it in collaboration with a doctor and look at what it would say compared to a doctor.
Was the placing of the various buttons understandable?	Yes, I think so.
Was the naming of the various buttons and titles understandable?	Could have more explanations throughout the application.
Do you have any general feedback about the system?	- The colors were a bit dark. - Smaller arrows. - Have Home and Back as button not only text. - Good flow throughout the app.

Questions	User 2
What do you think the percentage means?	It is one hundred percent in total, I get more effect from starting with physical activity than quitting smoking and intake of soluble fiber is what gives the least effect.
Was there anything that stood out as more challenging during the test?	I would like that the elements were more consist throughout the application. The question mark as an explanation was not intuitive.
Was it intuitive to navigate the application?	- Yes and no. - I would like that the elements were more consist throughout the application. - But it was easy to navigate, I did not get stuck anywhere.
What feelings are you left with after using the application?	- I liked the yellow explanation in the journal part. - I did not like the question marks. - Left with a feeling that it was a bit chaotic and inconsistent.
Would you use the application again?	- It does not work a 100%. - But the idea of making the journal easier to understand is very important, so you can see on your own and take responsibility for your own actions given some risk factors. That is something important for society. And such a pedagogic solution to get people more aware to see where they are on a scale like this and what they can do on their own is very important.
Was the placing of the various buttons understandable?	- Some buttons for selection are placed in front while the buttons with question marks are placed in the back. I would have them in the same position. - Would have a the import journal button more centered.
Was the naming of the various buttons and titles understandable?	- Just call it "journal" not "journalen". - I get drawn to "skip", possible remove the arrow. - Use a different word than synchronize as it is hard to understand what you mean by it.
Do you have any general feedback about the system?	- From a pedagogic perspective it was not consistent. - Arrow was not good. - Text in front and buttons behind.

Questions	User 3
What do you think the percentage means?	That is the thing that confused me. Why did the physical activity go down when I drank more alcohol?
Was there anything that stood out as more challenging during the test?	- I struggled to understand the presentation of suggested measures. What the percentage ment. It was not clear to me. - I struggled to import the journal.
Was it intuitive to navigate the application?	Yes it was intuitive to navigate.
What feelings are you left with after using the application?	- I liked it a lot. - I liked that one could hover over the words and get explanation. - I think that the application was constructive. - I don't know what the app is meant for but if it is supposed to help people get healthier or help people with health problems to map what they can change/improve then it is spot on.
Would you use the application again?	Yes, i would easily download this application.
Was the placing of the various buttons understandable?	Easy and naturally placed buttons.
Was the naming of the various buttons and titles understandable?	I found it quite intuitive, the only thing i struggled with was to understand percentages and efficiency degree.
Do you have any general feedback about the system?	- I liked the colors. - Very good. - Very healthcare vibes.

Questions	User 4
What do you think the percentage means?	The efficiency, if you want to do something to prevent high cholesterol you can start exercising and following the other measures suggested. Don't know why exactly this percentage is there, and what it means, why not 65%? Not entirely clear.
Was there anything that stood out as more challenging during the test?	- Could have a described what journal and health measures means. - Describe more what synchronizing journal does.
Was it intuitive to navigate the application?	- I like that there is not much hidden stuff. - There are only two options in the main menu. - If you have used this app before, it would be very easy to use it next time. - The one unclear part was the "efficiency degree " it is too cryptic. It confused me.
What feelings are you left with after using the application?	- It was good layout - Good colors - Not stressful
Would you use the application again?	- Yes, I think so
Was the placing of the various buttons understandable?	- Yes, maybe have one button for the "about application".
Was the naming of the various buttons and titles understandable?	- Maybe have an explanation to what those various buttons do. - The most stressful part was when i had to go to the health measures and the app asked to synchronize my journal, I did not know what that ment and what the consequences might be. If i will put my info there i can change some journal. Misunderstood the word. Thought that it would change the journal
Do you have any general feedback about the system?	- Cute hippo - If you add more diseases and factors the UI can be more challenging - There were some technical terms that not all will understand ex. what KOLS is, not everybody knows that. But the ones that have this will understand

Questions	User 5
What do you think the percentage means?	- Probability of someone starting with the measures?
Was there anything that stood out as more challenging during the test?	- I didn't catch at once what the task was, but once I did that it was ok. - I didn't quite understand what the efficiency degree was.
Was it intuitive to navigate the application?	- Easy to click around.
What feelings are you left with after using the application?	- Easy to click around. - Much info on the journal page, wasn't sure at once what i should do.
Would you use the application again?	- It was good to get a nice overview over the journal, but I don't know if i would use the health measure part.
Was the placing of the various buttons understandable?	- Don't know if I would have the search bar there in the journal part, it was a bit unnatural.
Was the naming of the various buttons and titles understandable?	- Yes, I will say that. Nothing that confused me. - I did not understand what the efficiency degree ment, wished it was more explanation to it.
Do you have any general feedback about the system?	- Easy to understand. - Good user interface. - Work a bit more with the design.

Appendix F

Closing interviews usability test 2

Below are the results for the closing interviews for the second usability test.

Questions	User 6
What do you think the percentage means?	I think the graph says something about how active I should be.
Was there anything that stood out as more challenging during the test?	The pop up message if I wanted to use my own journal, because the main thing I got out of it was "is it okay that the application used artificial intelligence to show the results I'll see " but it asks if it's okay to use artificial intelligence to retrieve data from my journal.
Was it intuitive to navigate the application?	Yes, I think so. There was nothing that made me unsure how to get to places or what the buttons did.So that was good.
What feelings are you left with after using the application?	I am left with the fact that the application was very easy to read, I was positively surprised by the health measures. It was fun to see. And useful. I get a feeling that a lot of work has been done to make sure that the users understand what happens in the application and easily get answers to what they need to know. Such as an explanation of difficult words and an explanation of health measures. It was very informative, but it was tucked away in a good way so you could only see if you needed it. Neat app.
Would you use the application again?	I imagine it would have been very useful if I were a person who has some long-term health problems and is a lot at the doctor's. Since I'm rarely at the doctor's, I do not feel that I need to use it a lot myself, but I can see that it is helpful for others.
Was the placing of the various buttons understandable?	Mostly, yes.
Was the naming of the various buttons and titles understandable?	Mostly, yes.
Do you have any general feedback about the system?	Really liked the start page with the hippopotamus, it made me very happy.

Questions	User 7
What do you think the percentage means?	There are four measures also it is how important they are in relation to each other.
Was there anything that stood out as more challenging during the test?	No, I think you have created a user-friendly application.
Was it intuitive to navigate the application?	It wast; I think you have nice boxes without too much text. They are easy to see. When I clicked on buttons, what I thought would happen, happened. And I went where I thought I would.
What feelings are you left with after using the application?	I think it was a good experience; it was not too health professional so that it becomes difficult for ordinary people to use it. Also, it was very explanatory and nice. Nice colors, and I think that the design is good, I do not get stressed by it.
Would you use the application again?	It could have been nice to get a summary of the journal If there are any things to grab you, I do not know how often I had used it, but in the course of life, there will be more things so I had used it several times but certainly not so often.If I get new writing, then I would like to use it.
Was the placing of the various buttons understandable?	Yes, I think so.
Was the naming of the various buttons and titles understandable?	Yes, I do not think anything was strange.
Do you have any general feedback about the system?	I think the hippopotamus was cute.

Questions	User 8
What do you think the percentage means?	How important they are to improve the selected diseases and factors.
Was there anything that stood out as more challenging during the test?	Finding general health measures, because when it was like this then I thought it was all about me. Clarify when you can see measures for myself and others.
Was it intuitive to navigate the application?	Yes, I think so.
What feelings are you left with after using the application?	I think the hippopotamus helps a lot, it makes things much sweeter. It's a little less dangerous. It helps a lot that everything is explained in simple words, what the different medical words mean. Nice colors, right design choices with theme in mind. I think round edges are good, it makes it more welcome than hard squares as one sees otherwise. I liked it. Also, it's cool with AI.
Would you use the application again?	Yes, it would have been nice to know what is written about me in the system.Otherwise, there is not much that is useful for me. But it is because I do not look at health-related things that often. But when I had needed it, I would use it.
Was the placing of the various buttons understandable?	Yes, I think so. Clear and understandable.
Was the naming of the various buttons and titles understandable?	Yes, I think so.
Do you have any general feedback about the system?	Do not think it was anything more. The hippopotamus was the best of it all.

Questions	User 9
What do you think the percentage means?	How important the health measures are regarding the selected factors.
Was there anything that stood out as more challenging during the test?	Make the health measures more clear. I struggled a bit with the yes and no questions. I managed to solve the task but had to think a bit there.
Was it intuitive to navigate the application?	Yes
What feelings are you left with after using the application?	Easy to navigate. It is clear what you can do. Very informative application.
Would you use the application again?	If I needed it, then yes. If I get new medical journals in, it would be nice to see it and get an overview and summary of the journal as well as the health measures.
Was the placing of the various buttons understandable?	Yes, everything felt natural.
Was the naming of the various buttons and titles understandable?	Yes, and there was provided information beside as well.
Do you have any general feedback about the system?	I would use boxes or something else to present the importance of the health measures instead of percentage. Nice design Easy to use

Questions	User 10
What do you think the percentage means?	How much you have to do the suggested measures to reduce the conditions.
Was there anything that stood out as more challenging during the test?	I think that the find health measures part could be done a bit different. I did not catch at once that it was my data from the journal that would be imported, because I read the information a bit too fast.
Was it intuitive to navigate the application?	Yes, I did not struggle with the navigation.
What feelings are you left with after using the application?	Friendly application, I liked the design. I liked that it was a lot of explanations, but that the information was hidden away.
Would you use the application again?	Yes, it would be nice to get a summary of the journal, and explanation of medical terms at once.
Was the placing of the various buttons understandable?	Yes, I think so. Clear and understandable.
Was the naming of the various buttons and titles understandable?	Yes, I think so.
Do you have any general feedback about the system?	Very friendly app, not as scary as some health care apps, liked the hippo.

Appendix G

Needs assessment

The purpose of this report is to document the development of our needs assessment for the HIPPO project. The purpose of the assessment is to get a better understanding of our problem space, and find solutions that can remedy these problems in a way that is useful to our users.

The assessment consisted of a series of steps. The first step was to figure out which problem we wanted to solve, identify stakeholders, and come up with initial ideas for solutions. Then, we gathered data from stakeholders and used it to identify needs and criteria for our solution. We then revisited our solution ideas with this in mind.

Problem introduction

The Covid-19 pandemic has placed a lot of strain on the Norwegian healthcare system. Because of this, some patients may have to wait for a long time to speak to a doctor or other healthcare providers, especially for non-urgent issues such as chronic illness management and prevention. Without access to a doctor, patients might also need to sift through difficult to understand documents to get information about their condition. Even if they are able to understand what is written in the documents, they might have trouble understanding which information applies to their situation.

This is the problem we want to tackle. Our idea is to develop software to allow users to more easily get an understanding of their medical documents, and that can automatically give personalized health advice to patients based on their medical history.

Before we start developing this product, we want to get a better understanding of the problem space. We need to know if this is an actual problem for people, and if our solution addresses the problem in a way that people find helpful. It would also be good to know more details about the problem from people with domain knowledge and any other factors we might not have considered.

Data Gathering

Numerous groups of stakeholders in this project have been identified. Primarily anyone who has had an interaction with the specialized health care service in Norway and further read the feedback given in a written format. On the side of design and use, there are other stakeholders such as developers and designers, but in the interest of the project as a whole the target group is limited to the group mentioned above. Data was gathered through interviews with individuals matching the criteria of the target group mentioned. Given the subjective

nature of the interviews, the information is seen more as a suggestion rather then direct needs, providing some leeway in the identified needs and criteria.

Identified needs and criteria

Some subjects reported that they received a lot of information that was too dense to easily parse. We also found that most of our interview subjects, especially on the patient side, found medical terms and abbreviations difficult to understand. They had to either ask their doctor or look them up and try to figure it out on their own.

From this, we figured that we need a way to get an overview and find specific information in large unstructured text documents. We would also like to make difficult words and phrases easier to understand.

We asked whether our interview subjects felt they got enough information about their situation from their doctor. Several reported that they either did not get enough information in general, or they wanted more information about something specific. When asked about what kind of advice or recommendation they would like to be given, most subjects responded that they wanted specific and applicable advice with some explanation as to why. Medical professionals generally wanted to know more of the underlying reasoning, whereas laypeople wanted a greater focus on what to do and what not to do.

Using our application, our users should be able to specify what they want information and advice about. Advice from our application should be specific enough to users' situation that they do not feel like it is generic advice. Recommendations should mainly be specific and applicable, and some of the underlying reasons behind them should be explained.

We also want our application to be used by people with varying levels of computer literacy and medical knowledge. It should have an simple and intuitive interface. Information should be presented in a way that is understandable by laypeople, but still retains enough of the original meaning to be useful.

Proposed solution

For the initial proof of concept prototype, the focus will be on diabetes and obesity. Creating a PDSS(personalized decision support system) includes the design of software that will use clinical data as input, but the testing of the software in a clinical setting is out of scope of this project. Our project is divided into three main parts: The first part addresses the automatic summarization of health journals with the use of NLP. The second part addresses a data driven way to give individual health recommendations. In the third part, these tools will be combined in an application for end-users, which integrates the tools from the first two parts.

Neil - Finding relevant data points in unstructured text can be difficult. Medical journals contain a lot of unstructured data, so we want to create a tool to turn unstructured medical text data into structured data. Our primary goal is therefore to automatically provide a summary of a medical journal using NLP techniques. Our secondary goal is to allow the users to specify what information they want from their journal.

CIMon - Further, we wish to provide the possibility to infer causality. This is done by analyzing different datasets, comparing data points to find if there are any overlapping points of interest, finally inserting the journal data, to find all actions relevant to the individual user. These actions will be somewhat generalized such as "if you were to lose X weight, Y would be more likely to happen".

Front-end - Finally, we will use human-centered design to create a frontend application that is intuitive and efficient for users with variable backgrounds within healthcare. The frontend will function with both the NLP and causal inference models and work as a connector between the two.

Conclusion

The current medical system in Norway has a gap between the language used by medical professionals. Test results can be difficult to interpret, with words being shortened does not help. In data gather in interviews we can see that the majority of the interview subjects had difficulties understanding the feedback from medical professionals in written form. Often due to the lack of time for each patient, there were not enough time for words and concepts to be explained in detail.

Our project proposes the use of natural language processing (NLP) together with multi variable regression analysis to improve the translation between theoretical words to describe a personalized decision to a layman.

From the data collected, there is a need for an abstraction level between the medical terminology and the everyday language used by patients. We can not determine if NLP can provide this level without further development and testing, though we believe it can.

Appendix H

Interview Questions

The following tables are the results from the initial interviews to determine needs and requirements for the project. The Results are split in two tables, each line marked with an ID pertaining to the given interview.

ID	Age:	Education/Profession:	How satisfied are you with the dissemination of information from your health contacts? Whats good? What could be better?	Have you read your own medical journal or other written communication from your health contacts?	Do you read the medical journals of others?
1	18-25	computer electrician	Yes, but I don't get the results from blood tests I do	I usually read the side effects of medication I take	
2	18-25	Technical IT-support	Usually use HelseNorge, and I've gotten information I needed from there.	Yes	
3	18-25	Sales	Happy with primary physisian. Information from specific departments such as MR was very slow	Yes, I read everything	
4	26-35	Economist and Computer engineer	In public care, no, information was enough, but very rushed. In private care I did not feel rushed, and the doctor had more time to listen and solve the problem.	Read what I got from the private sector. Not gotten much from the public sector.	
5	26-35	Environment worker	not from previous older doctor, I was not taken seriously. My new, youger doctor, is better at explaining.	read notes and some test results	
6	26-35	Police investigator	Very satisfied.	I've read the most important documents.	
7	26-35	Admin & IT consultant	No. It is difficult to find the information, even with right of access, it is difficult.	yes	
8	26-35	Nurse	Its ok. Often only answers 1 question out of the 5 I asked. Written text was to verbose, and often used long latin words	yes.	Yes, a couple of times a week.

9	56-65	specialist nurse	Often not enough time with doctor to go over all the relevant points.Information dissemination should not be left to doctors as the language becomes verbose and heavily influenced by "large words".	yes	on a weekly basis
10	26-35	specialist nurse	yes	yes	multiple times a day
11	26-35	Nurse	Yes	yes	Daily, but the information written by other nurses are often easier to understand.
12	36-45	Primary physician	yes	yes	Multiple times a day

ID	Do you find it easy to understand what is written in you medical journal?	Have you gotten enough information regarding concrete actions you can do to improve your long term health?	What type of feedback do you prefer? concrete actions or the theory behind them?
1	Often feels like a infodump.	In cases where my doctor knew the answer, yes	concrete, but with enough theory to know that im doing it correctly
2	I feel like I understand around 50%, but the are many unknown words. Easier for smaller problems	Yes, either from my doctor, or I get passed on to a specialist	A mix, I lean towards the practical part, but without an explanation I feel kind of blind
3	In no way! My primary physician is good, he explains it like I am 5 years old. In reports from the hospital they use all kinds of fancy latin words	Well, no. In physical therapy follow up took such a long time that I decided to go to a private institution as I figured they had forgotten me. I found out that an injury I have could have been avoided if I had gotten feedback sooner.	The private institution gave me an A4 paper with concrete instructions, this was better then anything I have gotten in the public sector
4	No, many "large words".	No, I have not!	Both
5	In test results, yes and no, there were some expressions that were explained, but others were not.	Was told not to do any physical activities.	Depends on the situation, but maybe both.
6	Yes, most of the language is understandable.	Yes, both written and verbally.	Concrete feedback, with a short explenation.
7	Easy to understand with the use of google.	No, it is fine verbally, but I get almost nothing in written format.	The theory behind would be better, or it would become a follow up question anyways.
8	Depends, I often need to google abbriviations or latin words.	Not relevant.	A mix of both, but with focus on easy to understand language
9	Test results are difficult to understand considering abbriviations.	Usually only get condition explained, not what I can do myself.	Combination, but anything past diagnosis should not be handled by doctor.
10	Usually short, but filled with latin words and abbriviations. Almost need a dictionary to read.	more or less. I wish i got the general feedback verbally with further information in written format.	Combination. Focus on the concrete actions, but with theory available to back it up.
11	No, a lot of large medical words, and a lot of abbriviations.	After recently switching primary physician, it is a lot easier.	Theory, but a mix would be the best.

12	Yes. When it comes to my patients there are major variations when explaining. Youger patients have an easier time using google.	No, but I haven needed it.	With my background I would prefer the backing theory, but I think that without my education I would prefer a focus on the concrete actions.

Bibliography

[1] Oludare Isaac Abiodun et al. "State-of-the-art in artificial neural network applications: A survey." In: *Heliyon* 4.11 (2018), e00938.

[2] Adobe. *Adobe XD*. Version 44.0.12. Apr. 13, 2021. URL: https://www.adobe.com/products/xd.html.

[3] Adobe. *Photoshop CC*. Version 20.0.0. Oct. 15, 2018. URL: https://www.adobe.com/products/photoshop.html.

[4] Betty van Aken et al. "Assertion Detection in Clinical Notes: Medical Language Models to the Rescue?" In: *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*. Online: Association for Computational Linguistics, June 2021, pp. 35–40. DOI: 10.18653/v1/2021.nlpmc-1.5. URL: https://aclanthology.org/2021.nlpmc-1.5.

[5] Emily Alsentzer et al. "Publicly Available Clinical BERT Embeddings." In: *CoRR* abs/1904.03323 (2019). arXiv: 1904.03323. URL: http://arxiv.org/abs/1904.03323.

[6] André Altmann et al. "Permutation importance: a corrected feature importance measure." In: *Bioinformatics* 26.10 (Apr. 2010), pp. 1340–1347. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq134. eprint: https://academic.oup.com/bioinformatics/article-pdf/26/10/1340/16892402/btq134.pdf. URL: https://doi.org/10.1093/bioinformatics/btq134.

[7] David Benyon. *Designing Interactive Systems*. Pearson Education Limited, 2013.

[8] Eta Berner. *Clinical Decision Support Systems: Theory and Practice*. Jan. 2007. ISBN: 978-0-387-33914-6. DOI: 10.1007/978-0-387-38319-4.

[9] Leo Breiman. "Random Forests." In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[10] Mette Brekke, Jeanette Solheimslid Bjørke, and Torben Wisborg. "where did all the doctors go?" In: *Tidsskr Nor Legeforen* (Nov. 2021). DOI: 10.4045/tidsskr.21.0701. URL: https://tidsskriftet.no/en/2021/11/editorial/where-did-all-doctors-go#article.

[11] Anna-Karin Carstensen and Jonte Bernhard. "Design science research – a powerful tool for improving methods in engineering education research." In: *European Journal of Engineering Education* 44.1-2 (2019), pp. 85–102. DOI: 10.1080/03043797.2018.1498459. eprint: https://doi.org/10.1080/03043797.2018.1498459. URL: https://doi.org/10.1080/03043797.2018.1498459.

[12] Álvaro Alonso Casero. "Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature." July 2021. URL: https://oa.upm.es/67933/.

[13] Ilias Chalkidis et al. "LEGAL-BERT: The Muppets straight out of Law School." In: *CoRR* abs/2010.02559 (2020). arXiv: 2010.02559. URL: https://arxiv.org/abs/2010.02559.

[14] Dhivya Chandrasekaran and Vijay Mago. "Evolution of Semantic Similarity—A Survey." In: *ACM Comput. Surv.* 54.2 (Feb. 2021). ISSN: 0360-0300. DOI: 10.1145/3440755. URL: https://doi.org/10.1145/3440755.

[15] Arnaud Chiolero et al. "Consequences of smoking for body weight, body fat distribution, and insulin resistance." In: *The American Journal of Clinical Nutrition* 87.4 (Apr. 2008), pp. 801–809. ISSN: 0002-9165. DOI: 10.1093/ajcn/87.4.801. eprint: https://academic.oup.com/ajcn/article-pdf/87/4/801/23918826/znu00408000801.pdf. URL: https://doi.org/10.1093/ajcn/87.4.801.

[16] Arman Cohan et al. "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents." In: *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 615–621. DOI: 10.18653/v1/N18-2097. URL: https://aclanthology.org/N18-2097.

[17] Pritam Deka, Anna Jurek-Loughrey, and Deepak. "Unsupervised Keyword Combination Query Generation from Online Health Related Content for Evidence-Based Fact Checking." In: *The 23rd International Conference on Information Integration and Web Intelligence*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 267–277. ISBN: 9781450395564. URL: https://doi.org/10.1145/3487664.3487701.

[18] Pritam Deka, Anna Jurek-Loughrey, and Deepak. "Unsupervised Keyword Combination Query Generation from Online Health Related Content for Evidence-Based Fact Checking." In: *The 23rd International Conference on Information Integration and Web Intelligence*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 267–277. ISBN: 9781450395564. URL: https://doi.org/10.1145/3487664.3487701.

[19] National Institute of Diabetes, Digestive, and Kidney Diseases. *Diabetes dataset*. 1990. URL: https://www.kaggle.com/datasets/mathchi/diabetes-data-set.

[20] National Institute of Diabetes, Digestive, and Kidney Diseases. *Obesity dataset*. 2020. URL: https://www.kaggle.com/code/mpwolke/obesity-levels-life-style/notebook (visited on 05/31/2022).

[21] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. "NCBI disease corpus: A resource for disease name recognition and concept normalization." In: *Journal of Biomedical Informatics* 47 (2014), pp. 1–10. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2013.12.006. URL: https://www.sciencedirect.com/science/article/pii/S1532046413001974.

[22] Günes Erkan and Dragomir R. Radev. "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization." In: *CoRR* abs/1109.2128 (2011). arXiv: 1109.2128. URL: http://arxiv.org/abs/1109.2128.

[23] Stanley S. Franklin et al. "Single versus combined blood pressure components and risk for cardiovascular disease: the Framingham Heart Study." eng. In: *Circulation* 119.2 (Jan. 2009). CIRCULATIONAHA.108.797936[PII], pp. 243–250. ISSN: 1524-4539. DOI: 10.1161/CIRCULATIONAHA.108.797936. URL: https://doi.org/10.1161/CIRCULATIONAHA.108.797936.

[24] Kelley Gordon. "5 Principles of Visual Design in UX." In: (Mar. 2020). URL: https://www.nngroup.com/articles/principles-visual-design/ (visited on 05/31/2022).

[25] HR. Han et al. "Using Patient Portals to Improve Patient Outcomes: Systematic Review." In: *JMIR Hum Factors* (2019). DOI: 10.2196/15038.

[26] Tin Kam Ho. "Random decision forests." In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. 1995, 278–282 vol.1. DOI: 10.1109/ICDAR.1995.598994.

[27] *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems*. Standard. Geneva, CH: International Organization for Standardization, July 2019.

[28] Anjali Ganesh Jivani et al. "A comparative study of stemming algorithms." In: *Int. J. Comp. Tech. Appl* 2.6 (2011), pp. 1930–1938.

[29] Alistair E.W. Johnson et al. "MIMIC-III, a freely accessible critical care database." In: *Scientific Data* 3.1 (May 2016), p. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35. URL: https://doi.org/10.1038/sdata.2016.35.

[30] Christian Johnson, Chelsea Richwine, and Vaishali Patel. "Individuals' Access and Use of Patient Portals and Smartphone Health Apps, 2020." In: *ONC Data Brief* 57 (2021). URL: https://www.healthit.gov/data/data-briefs/individuals-access-and-use-patient-portals-and-smartphone-health-apps-2020 (visited on 06/02/2022).

[31] Dan Jurafsky and James H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2009. URL: https://web.stanford.edu/~jurafsky/slp3/.

[32] Wafaa S. El-Kassas et al. "Automatic text summarization: A comprehensive survey." In: *Expert Systems with Applications* 165 (2021), p. 113679. ISSN: 0957-4174. DOI: https://doi.org/

10.1016/j.eswa.2020.113679. URL: https://www.sciencedirect.com/science/article/pii/S0957417420305030.

[33] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." In: *Bioinformatics* 36.4 (Sept. 2019), pp. 1234–1240. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf. URL: https://doi.org/10.1093/bioinformatics/btz682.

[34] Stephen J Leslie et al. "Clinical decision support software for management of chronic heart failure: development and evaluation." en. In: *Comput Biol Med* 36.5 (May 2005), pp. 495–506.

[35] Jiao Li et al. "BioCreative V CDR task corpus: a resource for chemical disease relation extraction." In: *Database* 2016 (2016).

[36] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries." In: *Text summarization branches out*. 2004, pp. 74–81.

[37] Andreas Messalas, Yiannis Kanellopoulos, and Christos Makris. "Model-Agnostic Interpretability with Shapley Values." In: *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. 2019, pp. 1–7. DOI: 10.1109/IISA.2019.8900669.

[38] Justin Morales. "The 7 Best Modern Fonts for Websites." In: (Sept. 2021). URL: https://xd.adobe.com/ideas/principles/web-design/best-modern-fonts-for-websites/ (visited on 06/02/2022).

[39] David Nadeau and Satoshi Sekine. "A survey of named entity recognition and classification." In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[40] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. "Natural language processing: an introduction." In: *Journal of the American Medical Informatics Association* 18.5 (Sept. 2011), pp. 544–551. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2011-000464. eprint: https://academic.oup.com/jamia/article-pdf/18/5/544/5962687/18-5-544.pdf. URL: https://doi.org/10.1136/amiajnl-2011-000464.

[41] *NSD - Norwegian Centre for Research Data*. URL: https://nsd.no/nsd/english/index.html (visited on 05/31/2022).

[42] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. "A Survey of the Usages of Deep Learning for Natural Language Processing." In: *IEEE Transactions on Neural Networks and Learning Systems* 32.2 (2021), pp. 604–624. DOI: 10.1109/TNNLS.2020.2979670.

[43] Fabio Mendoza Palechor and Alexis de la Hoz Manotas. "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico." In: *Data in Brief* 25 (2019), p. 104344. ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2019.104344. URL: https://www.sciencedirect.com/science/article/pii/S2352340919306985.

[44] Ted Pedersen et al. "Measures of semantic similarity and relatedness in the biomedical domain." In: *Journal of Biomedical Informatics* 40.3 (2007), pp. 288–299. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2006.06.004. URL: https://www.sciencedirect.com/science/article/pii/S1532046406000645.

[45] Ken Peffers et al. "A Design Science Research Methodology for Information Systems Research." In: *Journal of Management Information Systems* 24.3 (2007), pp. 45–77. DOI: 10.2753/MIS0742-1222240302. eprint: https://doi.org/10.2753/MIS0742-1222240302. URL: https://doi.org/10.2753/MIS0742-1222240302.

[46] Yifan Peng, Shankai Yan, and Zhiyong Lu. "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets." In: *CoRR* abs/1906.05474 (2019). arXiv: 1906.05474. URL: http://arxiv.org/abs/1906.05474.

[47] *pritamdeka/S-BioBert-snli-multinli-stsb · hugging face*. URL: https://huggingface.co/pritamdeka/S-BioBert-snli-multinli-stsb.

[48] *pritamdeka/S-Bluebert-snli-multinli-stsb · hugging face*. URL: https://huggingface.co/pritamdeka/S-Bluebert-snli-multinli-stsb.

[49] PRNewswire. "WebMD Launches Redesigned, State-of-the-Art Symptom Checker." In: (). URL: https://www.prnewswire.com/news-releases/webmd-launches-redesigned-state-of-the-art-symptom-checker-300624752.html (visited on 06/01/2022).

[50] *pubmed*. URL: https://pubmed.ncbi.nlm.nih.gov/ (visited on 06/03/2022).
[51] Nils Reimers, Philip Beyer, and Iryna Gurevych. "Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 87–96. URL: https://aclanthology.org/C16-1009.
[52] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: https://arxiv.org/abs/1908.10084.
[53] Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri. "Improving the accuracy of medical diagnosis with causal machine learning." In: *Nature Communications* 11.1 (Aug. 2020), p. 3923. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17419-7. URL: https://doi.org/10.1038/s41467-020-17419-7.
[54] James Robertson and Suzanne Robertson. "Volere Requirements Specification Template." In: (Jan. 2000).
[55] *scikit-learn*. URL: https://scikit-learn.org/ (visited on 05/23/2022).
[56] *Sentence-transformers/all-minilm-L6-V2 · hugging face*. URL: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.
[57] Terence Shin. "Understanding Feature Importance and How to Implement it in Python." In: *Towards Data Science* (Feb. 2021). URL: https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285 (visited on 06/01/2022).
[58] Jack Smith et al. "Using the ADAP Learning Algorithm to Forcast the Onset of Diabetes Mellitus." In: *Proceedings - Annual Symposium on Computer Applications in Medical Care* 10 (Nov. 1988).
[59] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. "BIOSSES: a semantic sentence similarity estimation system for the biomedical domain." In: *Bioinformatics* 33.14 (July 2017), pp. i49–i58. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx238. eprint: https://academic.oup.com/bioinformatics/article-pdf/33/14/i49/25157316/btx238.pdf. URL: https://doi.org/10.1093/bioinformatics/btx238.
[60] *Spacy industrial-strength natural language processing in python*. URL: https://spacy.io/.
[61] Reed T. Sutton et al. "An overview of clinical decision support systems: benefits, risks, and strategies for success." In: *npj Digital Medicine* 3.1 (Feb. 2020), p. 17. ISSN: 2398-6352. DOI: 10.1038/s41746-020-0221-y. URL: https://doi.org/10.1038/s41746-020-0221-y.
[62] Daniel Svozil, Vladimír Kvasnicka, and Jiri Pospichal. "Introduction to multi-layer feed-forward neural networks." In: *Chemometrics and Intelligent Laboratory Systems* 39.1 (1997), pp. 43–62. ISSN: 0169-7439. DOI: https://doi.org/10.1016/S0169-7439(97)00061-0. URL: https://www.sciencedirect.com/science/article/pii/S0169743997000610.
[63] *The 7 Principles*. URL: https://universaldesign.ie/what-is-universal-design/the-7-principles/the-7-principles.html (visited on 06/02/2022).
[64] Ken Thompson. "Programming Techniques: Regular Expression Search Algorithm." In: *Commun. ACM* 11.6 (June 1968), pp. 419–422. ISSN: 0001-0782. DOI: 10.1145/363347.363387. URL: https://doi.org/10.1145/363347.363387.
[65] Svetlana Ulianova, Vitalii Mokin, and Ayushi Kaushik. *Cardiovascular Disease dataset*. URL: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset (visited on 03/14/2022).
[66] Özlem Uzuner et al. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text." In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 552–556.
[67] Ashish Vaswani et al. "Attention Is All You Need." In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.
[68] *Vectors/norlm*. URL: http://wiki.nlpl.eu/Vectors/norlm.
[69] Wenhui Wang et al. "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers." In: *CoRR* abs/2002.10957 (2020). arXiv: 2002.10957. URL: https://arxiv.org/abs/2002.10957.

[70] *Web Content Accessibility Guidelines 2.0*. Dec. 2008. URL: https://www.w3.org/TR/WCAG20/ (visited on 06/02/2022).
[71] *WebMD*. URL: https://www.webmd.com/ (visited on 06/02/2022).
[72] *WebMD Symptom Checker*. URL: https://symptoms.webmd.com/ (visited on 06/02/2022).
[73] *world wide web consortium*. URL: w3.org (visited on 06/02/2022).