university of groningen

faculty of science and engineering

Machine Learning Classification of the Coronary Artery Disease and Clustering of Free-Form Medical Complaints

Boris Marinov

university of groningen

faculty of science and engineering

University of Groningen

Machine Learning Classification of the Coronary Artery Disease and Clustering of Free-Form Medical Complaints

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Human Machine Communication
at University of Groningen under the supervision of

Dr. ir. Peter van Ooijen (UMCG, Groningen)
Prof. Dr. Fokie Cnossen (Artificial Intelligence, University of Groningen)
Moniek Koopman (UMCG, Groningen)

Boris Marinov (s2957531)

August 26, 2021

Contents

Page

Abstract

Cardiovascular diseases are some of the leading causes of death across the world, with the coronary artery disease (CAD) accounting for the highest mortality rates. Effective treatment relies on an early and accurate detection. The signs and factors of the disease are well understood, however often require expensive and invasive methods which are not easily accessible to most practitioners. Machine learning (ML) application to the field carries promising solutions to the problems, and is already a widely researched topic.

The CONCRETE nationwide project was set up with the hopes of investigating whether ML-based prediction and analysis can be applied to easily obtainable, self-reported patient data. The data, despite being scarce, is made up of categorical quality-of-life answers and free-form, unstructured text complaints. Utilizing supervised method, this project shows that multi-label classification is possible to a degree when using the introspective answers only. Feature selection is used to quantify and discuss the gender and age-specific differences which contribute to the risk of the disease, promoting a more tailored future detection. Finally, unsupervised and NLP methods are combined to propose a form of topic modeling which goes beyond previous modeling methods and successfully discovers clusters of similar complaints for each disease group. The presented results pave a promising path and outline the future potential of the CONCRETE project in gathering more data and improving the efficiency of the Dutch health-care system.

1 Introduction

Machine and deep learning research provides attractive and promising solutions to many of the problems and tasks observed in the medical field [1]. These problems and tasks vary greatly in their nature and purpose, often requiring creative and unique approaches to tackling them. Luckily, the field of machine learning (ML) is ever growing, with new methods continuously being researched, improved and refined. This promotes and encourages the continuous application and approaches to different task types. Some of the most common applications of machine learning to medical problems fall into screening [2], risk stratification [3], prediction and assessing decision-making in a medical context [4]. The depth of research and extent of the successful application also depends on the mortality rate of the disease, with severe diseases attracting more attention.

Cardiovascular diseases (CVDs) have been reported as the leading causes of death by the World Health Organization (WHO) [1]. Out of that group of diseases, the coronary artery disease (CAD) is the most common type, and is caused by the narrowing or blockage of the coronary arteries, typically due to atherosclerosis [5]. Further, the disease varies in rates around the world, with more frequent cases in Russia and the Middle East [6], suggesting that there are more susceptible populations or genetic factors which could increase the likelihood of falling ill. To combat the increasing mortality rates, patients and hospitals benefit from an early and accurate detection. Understanding the different factors which predispose someone to the disease can also have great benefits in the long run, accommodating the early detection. CAD is relatively well understood, and there a range of methods for detection. Angiography (a CT scan with an injected contrasting material) is one of the standard procedures, however it is invasive and risky, unlike the potential promises of an ML-based approach.

Another reliable way is to check the calcium levels within the arteries, through the use of a computer tomography (CT) scan. The accumulation of calcium in the arteries has been shown to be a strong predictor of CAD [7]. While this procedure is not invasive, it needs to be performed in a medical center or hospital by trained personnel. Currently in the Netherlands most general practitioners (GPs) do not have access to such advanced diagnostics. Therefore the standard procedure is to refer all patients with non-acute chest pains and complaints to a cardiologist. After referral to the cardiologist, an exercise ECG is often performed. While this diagnostic is painless, it often gives sub-optimal outcomes, with almost 50% of reported test results being false negatives. It would be beneficial to understand which complaints and groups of symptoms could be predictive for the different CAD severity levels, across different patient demographics, aiding GPs in their decision making.

A nationwide project called CONCRETE [2] was undertaken in order to investigate whether a more reliable and early diagnosis and treatment of CAD is possible by giving GPs access to the CT calcium testing. The project consists of gathering a range of data from patients which had visited their GP with atypical angina pectoris (chest paints) and non-specific thoracic complaints. Patients are then referred to a CT scan in order to determine the amount of calcium present in the coronary arteries, which in turn indicates the presence of CAD and its severity. Overall the main objectives of the project are to:

- Evaluate whether GP access to CT calcium scoring leads to earlier CAD diagnosis and treatment
- Assess and optimize gender and age-specific diagnosis stratification based on the calcium score
- Determine which (clusters of) symptoms and risk factors could assist in web-based self-assessment of CAD

[1] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] https://concrete-project.nl

- Translate study findings to initiate a change in the Dutch health care policy by providing data on cost-effectiveness

At the current stage the CONCRETE project is still in its infancy, meaning that the collection of data is ongoing and new instances, information and ideas are continuously being updated. This research project aims to provide an initial attempt at tackling some of the described goals, namely, to investigate whether the collected data can be used to make predictions about the disease severity, as well as assessing the differences in the predictive factors. Additionally, whether specific clusters of similar complaints can be discovered and associated to the different disease severity levels. The collected data mainly falls under two categories, rating scores with regards to quality of life questions, and free-form answers describing particular complaints. Limited demographic information is also included, such as the age and genders of the patients. In particular the project aims to answer the following research questions:

- **RQ1** Can self-reported, introspective quality of life rating scores be used to predict a patient's CAD severity?
- **RQ2** Are there gender and age-specific differences in the predictive factors for CAD classification?
- **RQ3** Can free-form and unstructured text complaints be clustered to reveal groups of symptoms for a particular CAD severity?

As will be described in the related works section, ML application to detecting CAD is no novel task. Typically though, predictions are based on specific and precise instrument measurements, laboratory data and demographic information. These measurements are expensive, invasive and time consuming to gather. Having the ability to predict someone's risk of CAD based on survey questions, as well as understanding the gender and age-specific differences, carries many benefits. It can benefit the early and online detection, while also allowing to tailor the treatment and interventions more effectively to the specific patient.

1.1 Thesis Outline

This thesis will start off by investigating the already existing literature on the topic (Chapter 2), followed by a theoretical background of the techniques used to address the posed research questions (Chapter 3). The methodology (Chapter 4) describes the CONCRETE data set, some of its features and the two main pipelines used to gather the results. The results are presented (Chapter 5), followed by a discussion with regards to the research questions (Chapter 6), some limitations and future work suggestions. The thesis ends with the main conclusions drawn from the presented results and discussions (Chapter 7), and a full description of the data is included in the Appendix.

2 Related Works

This section summarises some of the work done in the past on CAD classification, outlining the success in the field and the promises it carries for ML integration in the Dutch health system. At the same time it highlights the differences and novelty of the CONCRETE task. Previous work on clustering of patient data is also presented, finishing off with some general information about the disease risks and properties of different demographics.

2.1 Heart Failure Classification

Attempting to predict the risk of a disease is, at its core, a classical ML task. The exploration of ML-based solutions to the detection of CAD dates back to the early 90's [8]. Combined with the mortality rate and high-impact of the disease, a vast number of attempts have been published. These attempts cover a wide range of data set types, as well as a variety of different ML-based methods. A recent review provides a comprehensive summary and comparison of 149 research papers related to CAD detection [9]. This overview was created with several goals in mind:

- Investigate the various ML techniques that have been applied, comparing their advantages and disadvantages
- The types of data sets used and how their characteristic impacted the overall performance
- The performance achieved by each ML technique on the specific data set

The review covers 67 different data sets and draws some informative conclusions from the available literature. For starters, most CAD data sets are rather small, with a median sample size of 350. The data sets also contain different features, and results have shown that different ML methods perform better depending on the feature types (Table 3 of the overview [9]). For example, Artificial Neural Networks seem to be the best performing classifiers for demographic, symptom and examination data, while Support Vector Machines and Decision Trees are more suited to laboratory, ECG and other heart-related measurements [9].

Nevertheless, some of the most commonly used categories of features are outlined to be: demographic; symptoms and examinations; laboratory measurements; ECG signals as well as fluoroscopy and echo measurements (slightly more advanced X-ray and heart based tests). Across all data sets, the best performing ML-based methods have been found to be Artificial Neural Networks, Decision Trees and Support Vector Machines, with Naive Bayes and K-Nearest Neighbors classifiers also receiving some attention. As noted by the review, accuracy scores are extremely high, virtually all falling in the 99-100% range [9].

The presented scores, as well as multitude of published papers and research on the topic support the use of ML-based detection in the Dutch health system. However, the aim of the CONCRETE project differs slightly to what has been achieved in the past. Most of the top performing papers focused on a simple, binary detection of the disease. In the CONCRETE project, the presence of CAD is broken down into five different severity levels. This turns the problem into a multi-class prediction.

Further, the data sets used contained a variety of features, most of which relied on a professional and trained personnel to be collected. In some cases these were also invasive and expensive to be collected. The CONCRETE project differs in that regard, and aims to investigate whether multi-label classification can be achieved by using easily obtainable, survey-like data provided by the patients themselves. Since ML models are so task and context specific, a new assessment and investigation into the topic is required.

2.2 Clustering in the Medical Field

While the previous research and the first goal of the CONCRETE project are mainly focused on the classification of the disease, the grouping of experienced complaints and symptoms is a form of clustering task. As such, it is draws on a different sub-field of ML and follows separate research paradigms.

The clustering of text data, whether it may be medical or from some other domain, is closely tied to topic modeling. Topic modeling is the ability to identify themes within a given set of documents and can reveal interesting information about the data. There are no fixed ways to perform topic modeling, however there are general steps one can take to ensure better results. Typically the text is turned into some numerical representation, which is then fed into the topic modeling method. Popular methods fall into two main categories: generative probabilistic models, such as the popular Latent Dirichlet Allocation (LDA) [10], which assume and attempt to model some probability distribution of the words, or non-negative matrix factorisation (NMF) which directly perform operations on some term-document matrices [11]. Each method however is suited to a particular text. LDA is better suited to longer documents as it assumes that the given text contains multiple topics, with methods such as SeaNMF being developed to handle the shorter text cases [11].

Another approach is to directly apply some clustering method on the text representation. Again, clustering methods are numerous and each carry their own strengths and weaknesses (Chapter 2.2). Combined with the difficulty of interpreting the discovered clusters or topics, the task of identifying informative groupings within a text is no-trivial task.

Attempts to utilise this in the medical field have been documented. One of these is applying the aforementioned LDA topic modeling to extract words which can be used for further classification of patient reports [12]. While successful, the research only focused on a binary classification task. A slightly more complex and well developed method, which also uses LDA, was carried out to attempt and predict the risk of depression [13]. Similarly, the research focused on extracting keywords and probability features using LDA and training a range of classifiers on them. Pure clustering on medical data is also documented [14], with clear topics being discovered in regards to five different classes.

While this research is applicable to the CONCRETE task, it needs to be reconsidered for the current state of the data and desired task. A lot of the topic modeling methods rely on forming large matrices of co-occurrence. The clustering research also represented the text with bag-of-words (Chapter 3.5.1). Both these approaches would result in sparse and less informative representations when applied on limited data, as such is the case for the CONCRETE project during the time of analysis of this thesis. Further, topic modeling methods typically output groups of individual words, and the same was done for the clustering paper [14].

The CONCRETE experienced complaints come in a free and unstructured form. Rather than attempting to discover clusters of single words, it would be beneficial to create a method which can group sentences of varying lengths and structure. This would not only boost follow-up interpretation and capture more information, but also continue to accommodate freedom in the way patients describe their complaints.

2.3 Risk Group Demographics

Risk of CAD varies across the world [6], suggesting that there are some geographical differences in the risk of the disease. This finding in a way reduces some of the applicability of the already developed ML detection methods, as most data sets are sourced from specific locations.

Further, there are likely differences between the risk factors across genders, and similarly for the different age groups. It has been reported that cardiovascular diseases (CVDs) has a male predominance in the younger population, especially in the 35 to 44 year old range [15]. This however diminishes with age, reaching a more equal sex-ration at ages 75 to 84. CVDs encompass a range of diseases, with CAD being the most common and lethal one, across both genders [16].

There are also differences between the genders in terms of mortality. Around half of the mortality caused by some CVD for women is due to CAD, and this is observed across all ages [15]. In males however, the mortality rates seem to be higher at a younger age (under 65) compared to later in life. This is an interesting findings, highlighting the need for early detection and research to increase detection and care for the female population.

In terms of risk factors, things which apply to the young can also be transferred to the older population. These factors include hypertension (high blood pressure), dyslipidemia (abnormal levels of cholesterol), impaired glucose tolerance, reduced physical activity and cigarette smoking [15].

It has been suggested that the aforementioned geographical differences in CVDs can be somewhat explained by local variations in these risk factors, socioeconomic positions and health services [17]. Figure 1 illustrates the most up-to date information on CAD death rates across the world, with the information being sourced from the World Health Organisation[3], with the highest rates observed in former Soviet Union States, the Middle East and North Africa [6]. The Netherlands is in-fact ranked much lower, 176th out of 183 countries. This by no means suggests that research should not be prioritised, as findings can easily be transferred to other regions and future research. Not only are there global differences, regional differences have also been widely reported, even in well developing countries such as the US [18], or Canada [19]. To account for this, the CONCRETE project is established across the whole country, aiming to involve as many regional GPs as possible.

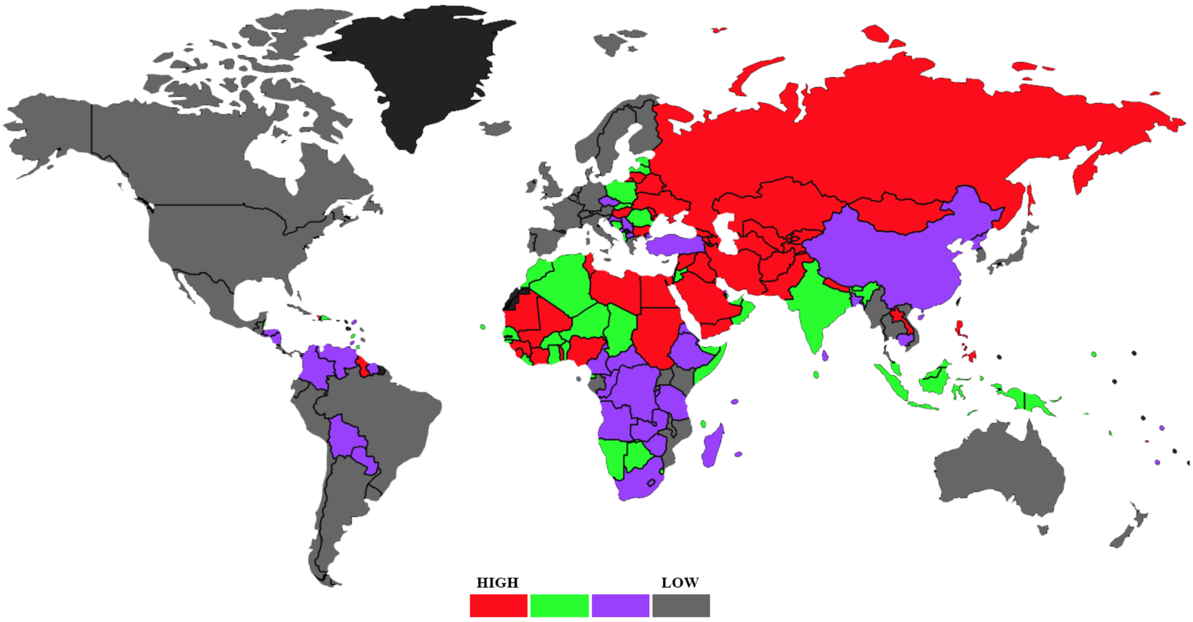


Figure 1: Global CAD death rates per 100,000 citizens, as reported by the WHO

[3] https://www.worldlifeexpectancy.com/cause-of-death/coronary-heart-disease/by-country/

3 Theoretical Background

This section of the thesis provides a theoretical base of the methods applied in answering the posed research questions (RQs). The methods are largely ML-based, with the different RQs being addressed by different types of learning. Natural Language Processing (NLP) is used to transform the free-form textual data into a machine-understandable representations which can then be grouped together to discover clusters of symptoms (**RQ3**).

3.1 Machine Learning

Machine learning (ML) is an active area of research in the field of Artificial Intelligence (AI). It is concerned with the study of algorithms which are able to learn and improve from a given set of data. The algorithms are able to extract patterns from a given *training set* of data, and generalize these patterns and learned behaviour to an unseen set of samples, often referred to as the *test set*. In a real world scenario, data is rarely in a set format, therefore the methods for learning and the overall approach to the task can vary greatly. While ML offers attractive and creative solutions to many real-life problems, this task-specificity is often a limiting factor. Nevertheless, the field of ML is ever-growing and can currently be divided into three major areas, relating to the type of learning involved in the process.

3.2 Types of Learning

The three main types of learning that make up the field of ML are supervised, unsupervised and reinforcement learning. The type of learning used is determined by the format of the available data, the formalisation of the task and the desired outcome. This project makes use of supervised learning approaches to investigate whether the numerical portion of the data (self-reported quality of life scores) can be used to predict the different disease severity levels (**RQ1**), while unsupervised learning is used to attempt and discover clusters within the symptoms and experienced complaints (**RQ3**).

3.3 Supervised Learning

In a supervised learning scenario, the algorithm is trained to learn a function $f : x \rightarrow y$ which maps an input x to some output y. Therefore, the given training data is labeled, structured as a series of paired instances of the input features x and the desired output y, (x, y). Since the model is shown the desired output, the method of learning is supervised. The input features and the output do not necessarily need to be in the same domain, and the input can be an arbitrary length of features. Typically the input features are in a numerical form or representation. Likewise, the output can be binary or of multiple classes. Some of the most common supervised learning tasks fall into classification or regression.

In the context of this project, supervised learning is used to attempt and classify a set of input features into one of the pre-determined CAD severity levels (**RQ1**). The set of input features are the different rating scores given by the patients with regards to the quality of life questions, and the output is one of the five CAD classes (no CAD, minimum CAD, mild CAD, moderate CAD and severe CAD). These class labels are based on the total calcium scores recorded by the GPs.

As both the input features and output are numerical and do not require extensive transformations, a variety of different classification methods can applied to see which performs best on the given task. Further, the process of training the models can benefit greatly from feature selection, i.e selecting the top features from the input which can still capture the patterns to be learned. The following sections describe common feature selection procedures, as well as the theory behind each of the used classifiers in the follow-up experiments.

3.3.1 Feature Selection

The importance of feature selection in the context of ML is well documented [20], to the point where it is a necessary step before training any models. This is also highlighted in the review paper for CAD classification [9], with an extensive comparison between some of the more popular techniques. The benefits of feature selection are numerous, it can reduce the computational costs of training a model in cases of large data sets, while also improving the performance of the model by providing it with only the ”useful” features and variables. Further, having a way to quantify the importance of a feature allows for comparisons to be drawn between different demographics and groups of patients (**RQ2**). Feature selection is often discussed alongside dimensionality reduction. While they aim to achieve similar goals, the methods and principles differ greatly and are not to be confused. Dimensionality reduction is also used in this project and is discussed in a later section (Chapter 3.4.1).

Due to the nature of the data and the supervised scenario, this project makes use of a statistical feature selection method. It is statistical because it evaluates and selects input variables that have the strongest relationship to the target, based on some statistical metric. In this case, the statistic is the ANOVA correlation coefficient (F-score). In short, the ANOVA F-score captures the explained variance by a feature, or a group of features, by computing the following ratio:

$$Variance = \frac{SST}{TotalSS} \tag{1}$$

where SST stands for the Treatment Sum of Squares and the Total SS is the Total Sum of Squares. The higher the the ratio, the higher the proportion of variance that can be explained by these selected features, and the more likely that they will be used in training the final models.

3.3.2 Classifiers

The type of classifier can also have a great effect on the overall performance of the classification task. Some of the factors which can influence this decision include the number of data points, number of selected features or whether the output classes are linearly separable. It is difficult to know prior to running the experiments which classifier will perform best, therefore it is common practice to set up experiments in which performance is compared amongst several candidates. The following section describes the theoretical workings of the classifiers which will be trained the numerical portion of the CONCRETE data set. Despite their different mechanisms and internal assumptions, all of the selected methods can handle the same input format and can be used for both binary and multi-class predictions.

K-Nearest Neighbor The K-nearest-neighbor (kNN) classifier is one of the simplest methods to both apply and understand, and is often considered a standard baseline approach when one is not sure about the exact nature of their data [21]. First introduced in the early 50’s [22], the method can be used for both classification and regression problems and falls under the class of non-parametric approaches. The method requires no training, each instance of the available ”training” data is represented as a vector in some multi-dimensional space, where each dimension corresponds to an attribute of the data. Each instance is also stored with its associated class label.

The *k* in k-nearest neighbors refers to the number of neighbors considered when a new instance is to be classified. In other words, given an unseen sample, the algorithm finds the *k* closest vectors based on some distance metric. The class of the new instance is then based on a majority voting system, with the most common label of the k neighbors being the final output. In a $k = 1$ scenario, the unseen sample is purely assigned to the closest vector in the data set.

The choice of distance metric is also dependent on the type of data at hand, outlined by several studies which investigate the performance of the classifier on different data sets [23, 24, 25]. Nevertheless, the Euclidean distance is typically the go-to choice [26]. Further, its performance on medical data sets, both numerical and categorical, has also been evaluated [27], making the classifier an appropriate candidate for this project.

Decision Trees Decision trees (DT) are a group of classifiers which have a long standing place in the field of ML, statistics and pattern recognition [28], introduced formally to the field in the late 80's [29]. They are hierarchical models which are expressed as a recursive partition of the instance space. The trees consist of nodes, with a single ”root” node at the base of the tree which has no incoming edges. All other nodes in the tree have exactly one incoming/outgoing edge. Nodes with an outgoing edge (so nodes within the tree) are called internal nodes. All other nodes are called leaves, or terminal nodes.

During training, the entire data set is assumed to belong to the root node. Following that, the tree is formed by splitting the instance space at each internal node into two or more sub-spaces (smaller branches), where the splitting decision is based on an input attribute value. The end of the tree is formed when each leaf, or terminal node, is assigned to one of the output classes. In this fashion, each branch relates to one class, and new instances are classified by navigating from the root to the leaf, according to the tests and splits made along the path [28].

The idea behind decision trees is straightforward, resulting in effective applications and intuitive interpretation of the results [30, 31]. Decision trees are also suitable for both categorical and continuous data, and are able to capture non-linear interactions [32]. Decision trees do carry some pitfalls, such as their tendency to overfit or be highly dependent on the training data, leading to great variations in classification accuracy and poorer generalization [30]. Nevertheless, their success is outlined in the CAD classification review [9], making them an attractive candidate for this project.

Random Forests Random forests (RF) fall under the decision forest class of classifiers. The are relatively newer compared to the other presented methods [33], but have already proven their efficacy in the medical field. Applications include predicting drug responses in cancer cells [34], recognising DNA proteins [35] and localizing cancer tissues [36].

A decision forest is based on the previously mentioned decision trees, combining a collection of them into an ensemble. The different decision trees each learn to form some prediction, and the overall forest prediction is based on some kind of majority voting system. As such, errors in predictions caused by an individual tree would have a smaller effect on the overall forest predictability, leading to improvements in the overall accuracy [37]. A key idea behind the improved group performance of the forest is that the trees are uncorrelated to one another. In a random forest scenario, each decision tree is trained on a random subset of the data and the features contained in it. Thus, the entire forest is able to capture and describe the full data set, through the combination of its random, disjointed parts.

As the core idea is the same, decision forests benefit in the same fashion as their individual counterparts, while also providing solutions to some of the described pitfalls, such as the overfitting or high training data dependency [38]. On the other hand, the joint behaviour of all the individual trees reduces the overall interpretation of the model, forming a type of black-box. Nevertheless, due to the documented success of the individual decision trees in CAD detection [9], random forests are also investigated in this project.

Support Vector Machines Support vector machines (SVM) have earned their popularity in the field of ML, largely due to their applicability, relatively easy concept and overall high performance on a variety of tasks [39, 40, 41]. At its essence, an SVM is an algorithm which learns to draw a separating line (often referred to as a hyper-plane) in some n-dimensional space, where n refers to the number of attributes/variables of an input. Figure 2 visualises the concept in two-dimensions, where half of the points belong to one class (red), and the rest form a separate class (blue). Plotting the points based on their attributes (each attribute relates to one of the two dimensions) reveals a clear separation between the classes, and the line which separates these is what the SVM classifier learns. Knowing where the line stands then allows for the classification of unseen and new samples.

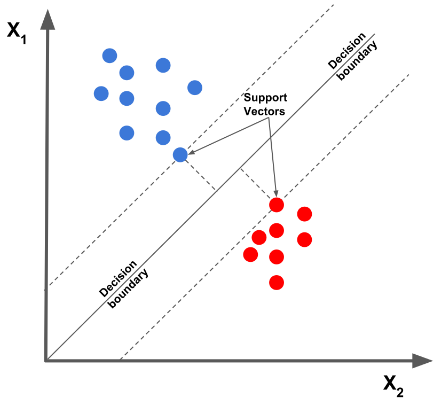


Figure 2: Visual representation of the SVM concept.

Extending to three-dimensions, the line now becomes a plane which can still linearly separate the points into their respective classes. SVMs are not the only classifier which follows the principle of a hyper-plane separation, however the way they decide on where exactly the line should sit is unique. SVMs follow the principle of the maximum-margin hyper-plane, which roughly speaking aims to place the separation in the middle [42].

Further, SVMs are attractive candidates for ML application due to their versatility. While they were designed for binary classification, multi-class application is possible too when used in a one-versus-all fashion. This extensions creates a separate SVM for each class, and has been successfully used for gene selection in cancer classification [43].

3.4 Unsupervised Learning

Unsupervised learning is the second main sub-field of ML. It differs to its supervised counterpart in that the model or algorithm is not provided the target output (class label) y during training. Instead, the model only has access to the input x and its goal is to find and learn patterns or groupings in the unlabeled data set. Typical problems in the field of unsupervised learning are clustering and dimensionality reduction approaches. As mentioned earlier, dimensionality reduction is closely tied to the field of feature selection, and can similarly be divided into both supervised and unsupervised approaches. Both clustering and three different types of unsupervised dimensionality reduction methods are employed in this project to attempt and discover groups of symptoms and complaints in the free-form text data (**RQ3**). The following sections will go over the basic principles of the involved methods, while section (Chapter 3.5) will describe the processes involved in transforming the text into a representation which can be used by the unsupervised techniques.

3.4.1 Dimensionality Reduction

To be able to apply the clustering algorithm on the free-form text data, the text is first converted into some numerical representation (Chapter 3.5). These representations often contain very high number of dimensions, which, as mentioned earlier, can hinder the performance of the applied algorithms. To yield and boost the performance, dimensionality reduction should be applied first, before attempting to cluster the data.

Clear links can be made between feature selection and dimensionality reduction, as they are both typically applied to improve performance and remove parts of the data which can be seen as redundant to the model. Further, both can contain methods which can be applied to supervised and unsupervised scenarios. There is a key difference however, in that feature selection methods do not alter the nature or values of the data. Rather, they select and exclude certain features deemed as unimportant for the task at hand.

While dimensionality reduction also lowers the number of dimensions (which can also be seen as features), it does so by transforming the features onto a lower dimensional space. In other words it can alter and change the values of the data, while still creating a representation which can explain most of the original variance and information. This project makes use of and compares the performance of three commonly seen dimensionality reduction methods in the field of machine learning: Principle Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) and the Uniform Manifold Approximation and Projection (UMAP). It is worth noting that the methods are not mutually exclusive, in fact it is even recommended in certain cases that one is applied before the other. PCA has been shown to be beneficial prior to applying t-SNE [44, 45], and a similar methodology is applied to the project at hand (Chapter 4.4.2).

PCA Principle Component Analysis (PCA) is one of the most well-known and established dimensionality reduction methods, often being the first option in standard ML practice [46, 47]. It reduces the dimensions of the data, while ensuring that most of the information is kept the same. In simple terms, the method achieves this by first standardizing the data points, scaling them to be within the same range. Following this, it determines the main principle components by computing the feature covariance matrices, eigen vectors and eigen values, which, through linear algebra principles can highlight the directions of maximum variance. Higher variance relates to higher amounts of information in the data. Therefore, the principle components are ordered in descending order based on their eigen values. The original data is now represented by the highest ranked principle components and lower ranked ones can safely be removed without fear of losing important information. When using the method, one can select how much of the explained variance they would like to keep, typically setting this to 90-90%.

t-SNE T-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique developed with the purpose of visualising high-dimensional data onto a two or three dimensional map [48]. The need to visualise high-dimensional data becomes apparent in many different domains, from medical visualisations of genes [49] to music analysis [50], cancer research [51] or geological domain interpretation [52, 53]. T-SNE, like most dimensionality reduction methods, carries deep mathematical roots. In simple terms, it achieves the lower dimensional mapping by checking the similarity between data points, repeating that for each pair. These similarities for the high-dimensional points are then translated to probability distributions, where similar points have higher probabilities than dissimilar ones. A lower-dimensional map is then constructed by randomly assigning the high-dimensional points, and a second probability distribution is created for this lower-dimensional space. The method then attempts to minimize the divergence between the two distributions, with respect to the locations of the points in the lower-dimensional map. T-SNE is not a trivial method to understand, and while it is successful in the lower-dimensional mapping, it requires numerous hyper-parameters and is notoriously non-rigid with respect to them. In other words, the choice of parameters can greatly influence the final results, to the point where clusters can be discovered in non-clustered data [54]. In an unsupervised learning scenario this only introduces more difficulty to the problem. Interactive selection of the hyper-parameters is generally recommended [55, 56], however without knowing how the data is supposed to look and whether actual clusters are present in it, it is very easy to introduce false findings. Nevertheless, this project makes use of t-SNE to reduce the numerical representations of the free-form text complaints, and a ”pooling” strategy is proposed to overcome the parameter selection issue.

UMAP The Uniform Manifold Approximation and Projection (UMAP) method is heavily inspired by t-SNE, and there are many similarities between the two. It does however offer advantages over t-SNE, such as speed and better preservation of the global structure of the data. The workings of UMAP are also based on complex and advanced mathematical concepts. In simple terms, UMAP, just as t-SNE, use graph layout algorithms to map the data into the lower-dimensional space. The differences between the two methods are mostly in the initial stages of how the high-dimensional graph is created, and the addition of some optimisation tricks of the mapping to the lower-dimensional graph. The biggest differences between the outputs of the two methods is the balance between local and global structure (Figure 3)[4]. The presented figure, taken from a very intuitive and high-level Google blog post, clearly shows how well clustered each different category is for the UMAP output (local structure), while clusters which belong to a similar categories (”pullover”, ”t-shirt/top”, ”dress”) still tend to colocate (global structure). T-SNE is still able to form meaningful clusters, however UMAP extends this further but also separating the clusters more clearly from each other. This is beneficial in aiding the application of subsequent clustering algorithms. UMAP however also suffers from the careful choice of hyper-parameters, and the proposed ”pooling” strategy is also applied for this technique in attempting to group the free-form text complaints across the CAD severity levels.

There is a notable feature for both the UMAP and t-SNE outputs which needs to be addressed, as it affects the choice of clustering algorithm. That is, the distances between clusters in both outputs are mostly meaningless. This is because both methods use local distances in the initial high-dimensional graph construction. Therefore, clustering algorithms which rely on distance measures between the clusters are likely less suited for this case.

3.4.2 Clustering Algorithm: HDBSCAN

Clustering algorithms can roughly be divided into four main groups: hierarchical, centroid-based, graph-based and density-based clustering. Each of those have their own assumptions and limitations, and most require several hyper-parameters to be optimised and achieve meaningful results[5]. The attached `sklearn` page provides a nice overview of the use-cases and applicability of a collection of clustering methods.

Again, as mentioned earlier, clustering is a classical unsupervised learning task, and as such, the correct output is not known. Therefore the selection of these parameters can be a non-trivial task, greatly affecting the final interpretations and results. A common parameter across many clustering

[4] `https://pair-code.github.io/understanding-umap/`

[5] `https://scikit-learn.org/stable/modules/clustering.html`

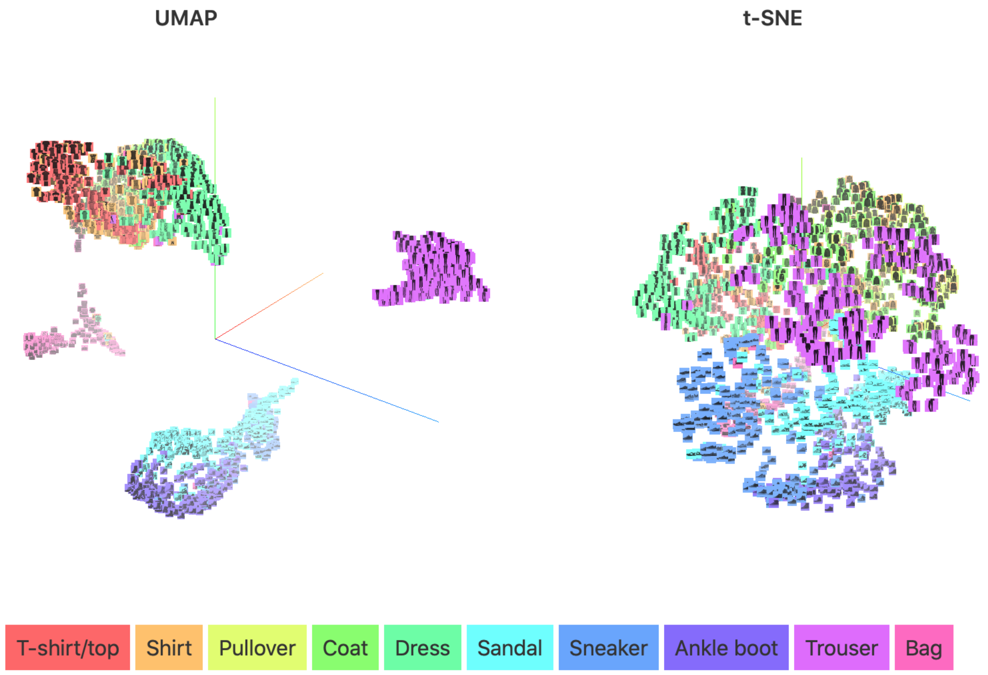


Figure 3: Visualisation of the differences between t-SNE and UMAP outputs. Image sourced from blog in footnote.

techniques is the need for prior knowledge of the number of clusters, something which does not exist in an unsupervised learning scenario.

The need for minimal and trivial hyper-parameters, as well as the reduced meaningfulness of the cluster distances in the UMAP and t-SNE outputs, make density-based clustering methods an attractive option. In particular, this project makes use of of HDBSCAN technique [57], which only requires one parameter to run (the minimum cluster size) and computes the clusters based on distances between the nearest points, rather than the spaced out clusters.

3.5 Natural Language Processing

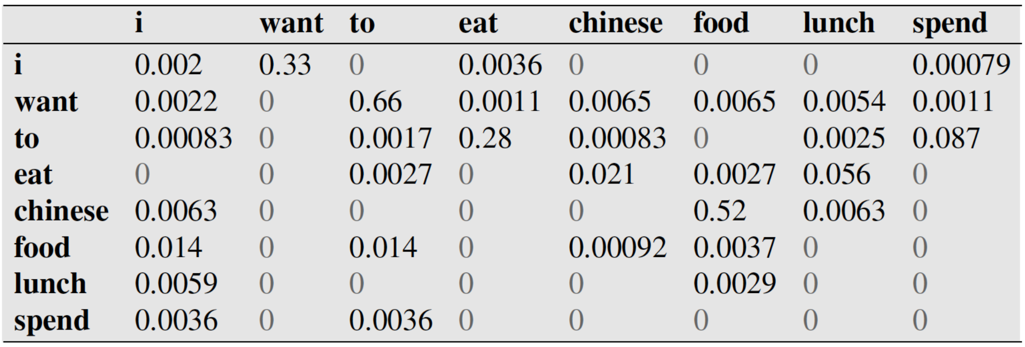
Natural language processing (NLP) is a sub-field of AI concerned with providing computers the ability to understand, process and analyze language. The field, just like machine learning, is a heavily active area of research. This is to no surprise, as language is one of the main forms of communication. It is also famously complex, as it is built up from atomic parts which, when combined with ever changing rules and principles, reach higher levels of meaning and pragmatics. These non-uniform rules also make languages notoriously ambiguous and often challenging to parse, even for humans. Such properties make the understanding of languages by machines a very difficult task. Nevertheless, NLP research is constantly improving and common NLP tasks such as speech recognition, part of speech tagging or language generation are reaching ”human-like performances” on specific data sets and varying benchmark tests [58]. The success in the field is in part due to sophisticated and clever ways of representing languages in a numerical form which can be understood by a computer.

ID	Word1	Word2	Word3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Table 1: One-hot encoding example. The original collection of words would in this case be (Word1, Word2, Word3, Word2).

3.5.1 Language Representations: Vector Space and Embeddings

A computer requires a language to be in a format it can process, which is often some numerical representation. These representations range in complexity and their uses depend on the task at hand. One-hot encodings (Table 1) simply assign a unique value to a word and its categorical identifier, often a binary representation, and are unable to capture the relationship between words or their surrounding context. N-gram language models take this a step further and are able to estimate the probability of a word given the words that come before it (Figure 4). This is an improvement, however languages often rely on long-distance dependencies and would require insufficiently long n-gram models.

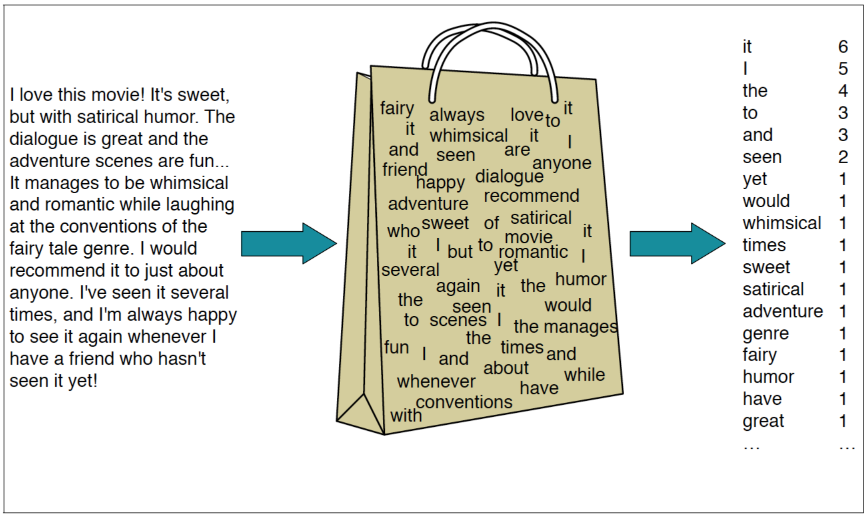


	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Figure 4: Table of bigram probabilities

N-grams are thus able to somewhat represent a text in a sequential manner. Other methods, such as the bag-of-words language models (Figure 5) instead keep track of word frequencies, and allow classifiers to form probabilities of the respective word classes. Thus solving slightly more complex tasks, such as sentiment analysis or text authorship and classification. Nevertheless, none of the described methods are able to capture the meaning (semantics) of a text to a sufficiently high level which would allow the creation of inferences, question-answering systems or human-like speech generation.

This is where word embeddings come into play. These embeddings are vector representations which are able to capture the meaning of a word by taking into account the surrounding context. In a very simplified manner, words are assigned numbers (vectors) which define them in some N-dimensional space. By counting and keeping track of surrounding words in a text, vectors can be assigned to each, such that similar and commonly occurring words are close to each other in that N-dimensional space (Figure 6). This is already an attractive method, however is depended on the document length and can result in sparse and insufficiently long representations.



I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Figure 5: Bag-of-words representation. Position of in text is ignored, representation only makes use of word frequencies.

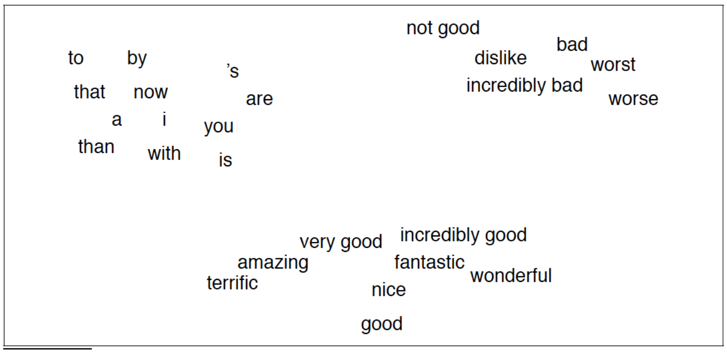


Figure 6: Two-dimensional t-SNE projection of word embeddings in some vector space. Words similar in meaning form clear clusters.

3.5.2 Language Models and BERT

Language models were designed with the purpose of forming denser vector representations, which can subsequently be used by other machine learning models. The idea is to train a model on some task, often the prediction of a target words based on context words, and extract learned model weights to form the word embedding. Another approach is to optimise the document extracted word embeddings directly to reach some desired property. Embeddings are not only limited to the word level. Through the combination of model layers, concatenation of vectors and clever design, embeddings can be formed on the sentence or even higher document levels.

Language models come in many variants, depending on the design of the task they are trained on, or the training corpus itself. Furthermore, they are publicly available, pre-trained and provide easy access to the stored word embeddings. One of the first such language models is **Word2vec** [59], which has already seen wide use cases in NLP tasks [60, 61, 62]. Language models such as **GloVe** [63] take the subsequent approach of optimising the word embeddings themselves, also earning their place in the NLP field [64, 65]. The Bidirectional Encoder Representations from Transformers (BERT) language model is currently one of the most popular models in the NLP field, achieving state-of-the-art performance on a number of language understanding tasks [66]. While the model architecture, training tasks and principles require extensive deep learning knowledge to fully grasp, using it is relatively simple and applications are numerous.

This project makes use of the BERT architecture to extract and form vector representations of the free-form complaints provided by the CONCRETE patients. These embeddings can then be used as input for the aforementioned dimensionality reduction techniques and clustered to hopefully discover meaningful and informative groups for each CAD severity (**RQ3**). In particular, the Dutch BERT variant BERTje [6] is used to fit the nature and language of the data. While BERT models are multilingual, meaning that they have been trained on corpora of various languages, BERTje has been fine-tuned further and consistently scores higher across NLP tasks in the Dutch language [67].

[6] `https://github.com/wietsedv/bertje`

4 Methods

The following section will describe the nature and general outlook of the CONCRETE data as well as the main process of collecting it. Certain preprocessing steps are highlighted, mostly in regards to the manual reconstruction of the text data, before dividing the section into the two main approaches used to answer the posed research questions. This division is necessary, as the two research questions vary in their theoretical background and overall steps taken to investigate them.

4.1 The CONCRETE Data set

Existing data sets with regard to the machine learning application to detect CAD vary, not only in the number of samples, but also in the number and type of features [9]. While there are some data sets which contain samples in the thousands, the typical CAD data set is small in size [68]. This is not ideal, as ML model performance benefits from plentiful training instances. Further, it decreases the generalizability of the models as they may not have been trained on all feature combinations.

In terms of the types of features used, most of the reported data sets are made up demographic, laboratory, symptom and more sophisticated heart and muscle activity-based measurements. While these are not shared uniformly across all investigations, certain common features such as the age and sex of a patient are widely observed.

The CONCRETE data is currently in its infancy stage and shares some of the aforementioned properties, while also being unique in its own regard. Unfortunately the current data set consists of approximately 80 samples that can be used to investigate either of the research questions. This is in part due to the early stages of the project, the difficulty of collecting data from numerous GP offices, as well as the limitations posed by the ongoing COVID-19 pandemic. Nevertheless it is still worth attempting to apply ML-based methods, as this could reveal whether the project is worth pursuing further. The CONCRETE data set does also include laboratory measurements, the calcium scores used to determine the CAD labels. These however are not included in any of the follow-up experiments as they already serve the purpose of the ”ground truth” in the data and calcium is a well documented predictor of CAD [7].

The uniqueness of the data set lies in the type of features it is made up of. Rather than having a range of laboratory and precise machine measurements, the CONCRETE data set can currently be divided into two major halves. The first half is made up of self-reported scores, usually on a five-point scale, in regards to different quality of life questions. These were related to chest pain complaints of the patient and any related difficulties they may have experienced as well as general quality of life complaints. Tables 20 and 19 in the Appendix contain the full set of questions and possible responses that the patient was allowed to give. This half of the data, as well as the age and gender of the patient, contributed to the investigation and training of the models used to attempt and predict the CAD severity levels (**RQ1**). In other words, attempting to predict the disease mostly based on patients' own feelings and introspective answers. A small portion of the participants had also undergone follow-up surveys at 6, 12 and 24 months, during which the same quality of life questions were asked. While these could have provided interesting insights into the development of the disease, they are also not included in the experiments purely due to the limited and incomplete surveying.

The second half of the data are the free-form, unstructured text complaints. Table 18 summarises the variables and the associated questions which the patients had to answer. Again, most of these were in regards to experienced complaints and the patients had the freedom to report anything they thought might be important. This was done with the hopes of gathering data which might be unique and predictive of a specific CAD severity level. Most importantly, this half of the data was used to

Question	Original Answer	Clean Answer
How severe are the by you described complaints on a scale of 1 to 10?	Vraag 15 7 op de schaal van 10 Vraag 16 6 op de schaal van 10	6.5
	Het is geen pijn wat je voelt meer een warme gloed van binnen uit alsof de adrenaline door je lichaam giert.	Het is geen pijn, eerder een gevoel van warme gloed die door het lichaam straalt *(moved to open_uitstralen question)*
	2 of 3	2.5
	Op het ergste punt wel een 9	9
Can you explain with what kind of complaints you contacted your GP?	Altijd koud , geen goede bloed Door stroming.	Altijd koud, slechte doorbloeding
	Moe! 'S morgens opstaan en je nog moe voelen Hijgen , terwijl we al meer dan een half jaar aan het sporten zijn en aan het lijnen zijn Mijn man voelt zich fit en is 20 kg kwijt en ik ben moe en 0 afgevallen Zat op de bank en mijn man zei wat zit je te hijgen terwijl ik niets deed	Moe, zelfs 's morgens bij het wakker worden. Hijgen en kortademigheid
How often do you experience the described complaints?	in augustus tijdens de hitte. ook 's morgens vroeg met krant lezen. even hevige pijn en duizelig .Niet misselijk geweest. Toen ook paracetamol genomen en rustig op bed gelegen.	In augustus tijdens de hitte en ook vroeg in de ochtend *(answer still not exactly what is asked)*

Table 2: Some of the numerous examples of text entries which had to be cleaned. Redundant or irrelevant information removed, often moved to another question. Ratings given in a range are given a number in the middle. Some cases still not perfect, as to not structure the data too much.

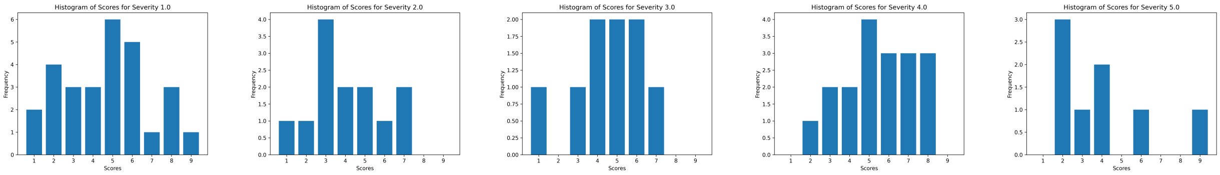
attempt to develop a method which can form and locate clusters of similar complaints with regards to the disease (**RQ3**).

4.2 Manual Cleaning and Exploration of the Unstructured Text Data

While the patients were encouraged to answer the experienced complaints questions as closely as possible, they were also given complete freedom to their answers. This opens up the possibility of gathering redundant, unrelated or repeating information. While these ”artefacts” might have less of an affect in a large-scale data set scenario, the CONCRETE data set is sparse and it was important to investigate the nature of the text data closely. Indeed, an initial look into the given answers revealed numerous cases which motivated a full clean-up and reconstruction of the textual data. Table 2 provides a few picked out examples, as well as their cleaned up entries. These include cases where patients either provided excess information which was irrelevant to the question at hand, repeated or gave an answer which was more suitable to a different question, or simply gave an invalid answer (giving a verbal description when asked to simply give a rating or answering ”See Question X”).

The cleaning of the data was done manually and while it introduces the risk of subjective manipulation of the answers, it was a necessary step in providing some minimal structure to the limited data and increasing the chances of finding meaningful clusters. Cleaning also included removing or changing invalid characters, expanding certain abbreviations provided by the patients and creating a mapping dictionary to change certain phrases uniformly across the data set.

Cleaning the data set manually also allowed for a closer inspection into the general nature and distribution of symptoms that the patients reported. At an initial look, it seemed that there were little differences in the reported symptoms between the CAD severity levels, even when comparing the No vs Severe CAD entries. To investigate this further, Table 3 summarises the percentage occurrences of the top most informative words in the data set across all severity levels. Indeed, the distributions are rather uniform, with only a few obvious cases of differences highlighted in red. For example, it seems like ”rest” is mostly present in less severe CAD levels, suggesting that it does not reduce the symptoms for Severe CAD. Words relating to family matters (mother, father) are also more present on the severe end of the disease, suggesting mentions of genetic predispositions. Similarly, Figure 7 displays the distribution of the reported pain scores by the patients for all severity levels. There are no clear patterns, for example, higher pain scores associated with more serious severity levels, and this could be due to the nature of complaints. The patients in the study have atypical angina pectoris or non-



(a) No CAD (b) Min CAD (c) Mild CAD (d) Moderate CAD (e) Severe CAD

Figure 7: Summary of reported pain scores across all CAD severity levels.

specific thoracic complaints, which are different to expressed complaints from patients experiencing typical angina pectoris issues. There is a possibility that these patients experience less serious pain scores, while still having a severe underlying CAD. This observation, combined with the limited data samples, supported the decision to focus on clustering of the data, rather than attempting to build a classifier which can predict the multiple classes.

4.3 Predicting CAD Severity Levels (RQ1/2)

In total there are 83 samples which can be used to train the classifiers in attempting to predict the different CAD severity levels. These are patients which have answered all relevant questions completely and accurately, and have also undergone the calcium measuring test in order to have an accurate CAD severity label.

4.3.1 Task Definition, Experimental Groups and Evaluation Metrics

There are five unique labels in the data set: no, minimum, mild, moderate and severe CAD. This makes the prediction a multi-class classification task. While all the selected classifiers are capable of handling such tasks, the limited data set and nature of the features (self reported scores) support the need for a simpler task classification. Therefore a binary task was also introduced, for which all labels, apart from no CAD, are set to 1. This was done with the hopes of overcoming the sparsity of the data, as well as getting an initial outlook at whether the features can be used as predictors.

To assess whether there are differences between the predictability and feature importance between the different demographics of the data (RQ2), several experimental groups were predefined. These were the full data, males versus females and young versus old patients. The age split was set at an age equal to or younger than 62 years. While this number might appear arbitrary, it was largely based on data availability, ensuring that each group would contain at least some instances from all classes.

As evaluation metrics, accuracy, recall and F1-scores were used in comparing and determining the best fitting models for each of the tasks. While accuracy is a very common measure, it has been shown that it could be misleading for highly imbalanced data sets [69], therefore numerous metrics are employed. Further, purely relying on it, especially in the bio-medical field is not advised [70].

4.3.2 Addressing Class Imbalance

Table 4 illustrates the distribution of the class labels, for each data subset experimented on. Aside from the lack of data in certain cases, there is also a noticeable class imbalance, something which had to be addressed as it is very likely to hinder the model performance. This occurs due to the models tending to ignore the minority class, leading to decreased performance in predicting it. In this case, the minority class is the Severe CAD, the class one would argue is the more important in predicting

		CAD Severity Score				
		No CAD	**Minimum CAD**	**Mild CAD**	**Moderate CAD**	**Severe CAD**
Word	**Translation**	**Word Distribution %**				
borst	chest	76.7	77	80	65	87.5
arm	arm	26	38	30	35	12.5
links	left	33.3	38.5	30	45	25
pijn	pain	66.7	69.2	60	65	62.5
druk	pressure	46.7	46.2	40	35	50
inspanning	effort	50	53.9	40	50	12.5
rust	rest	73.3	76.9	80	45	25
dag	day	33.3	53.8	60	25	50
hart	heart	36.7	69.2	30	35	12.5
schouderblad	shoulder	23.3	23.1	20	25	37.5
minuten	minute	36.7	30.8	40	35	62.5
stress	stress	26.7	15.4	40	35	12.5
werk	work	13.3	53.9	0	30	0
kortademig	shortness of breath	30	23.1	20	25	0
moe	tired	33.3	30.8	50	25	50
dagelijks	daily	30	15.4	10	25	0
hartslag	heartbeat	3.3	0	10	5	0
moeder	mother	6.7	0	0	10	25
spierpijn	muscle pain	3.3	0	10	10	0
vingers	fingers	3.3	0	10	10	0
sport	sport	10	7.7	10	5	12.5
hartinfarcten	heart attack	10	0	0	10	0
ribben	ribs	3.3	7.7	10	10	12.5
vader	father	10	0	10	20	25
bloeddruk	blood pressure	10	23.1	10	10	0
continu	continuous	13.3	15.4	0	15	12.5
slapen	sleep	16.7	0	10	20	25
hoofd	head	16.7	0	0	10	12.5
duizeligheid	dizzy	10	7.7	10	15	12.5
adem	breath	26.7	15.4	0	15	12.5
tinteling	tingling	3.3	15.4	10	5	0
knie	knee	0	23.1	0	0	0
	Total number of words	2695	1007	1012	2194	699
	Data Instances	30	13	10	20	8

Table 3: Word distributions across all CAD severity levels for the most common and informative words in the data. Total number of words, as well as number of patients for each category included as well.

	CAD Severity				
Experimental Group	**No CAD**	**Minimum CAD**	**Mild CAD**	**Moderate CAD**	**Severe CAD**
All	30	13	11	21	8
Males	11	3	5	9	4
Females	14	6	4	8	3
Young (≤62)	24	9	6	14	2
Old (>62)	5	4	5	7	6

Table 4: Number of samples for each class across all experimental groups.

correctly. Class imbalance is typically addressed by over- or under-sampling of the data, depending on where the imbalance lies. As there was a lack of instances for the severe CAD class, the data is over-sampled using the Synthetic Minority Over-sampling Technique (SMOTE) [71]. Rather than duplicating already existing data points which would add little to no new information to the data set, SMOTE synthesizes new examples. In short, the method achieves this by selecting samples which are neighbors in their respective feature spaces, creating a line between them and synthesizing a new point at that line.

4.3.3 Model Building and Running of Experiments

Once the data is over-sampled to account for the class imbalance, all entries are standardized using the `sklearn` Standard Scaler implementation. Standardization of the data is a common preprocessing step for machine learning experiments as it ensures that entries come from more or less the same distribution, and varying scales will not influence variable selection in subsequent stages. Feature selection is then performed by selecting the variables with the 19 highest ANOVA F-scores. These were also stored for all experimental groups to assess the differences in feature importance between them (**RQ2**).

The classifiers used for the experiments and the prediction of the CAD severity levels were k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Decision Trees (DT) and Random Forests (RF). All model implementations were sourced from the `sklearn` machine learning library. Each of these models has a specific set of hyper-parameters which need to be selected manually and can greatly affect the model performance. Prior to running the experiments, it is near impossible to know which set of hyper-parameters will perform best for the given data, therefore a grid search is performed for each of the models. The grid search implementation used also performed k-fold cross validation, which is another standard practice in machine learning experiments. K-fold Cross validation is the process of splitting the data into k groups. At each run of the validation, one of the groups is treated as a test set, while the remaining data is used as the training set for the model input. The model is then evaluated on the held out test set, and the process is repeated with the remaining groups, retraining a new model on each new test set. Cross validation is an important step as it ensures less bias and less optimistic estimates of the model, thus more accurate results. In a simple training/test split procedure a model could happen to be trained on a particularly ”nice” part of the data, producing overly optimistic results.

The k-fold cross validation grid search is performed for all four classifiers, outputting the best set of hyper-parameters for each. Once these are determined, a final k-fold cross validation is performed with the selected hyper-parameters to get the final evaluation scores for each classifier. For both the grid search and final cross validation the k value was set to 10. Figure 8 shows a schematic of the overall experiment pipeline described above. The entire process is repeated for each classifier, task type and experimentation group, resulting in a total of 40 different experiments. From these, the best classifiers for each demographic and task type were selected and presented in the latter results section (Chapter 5).

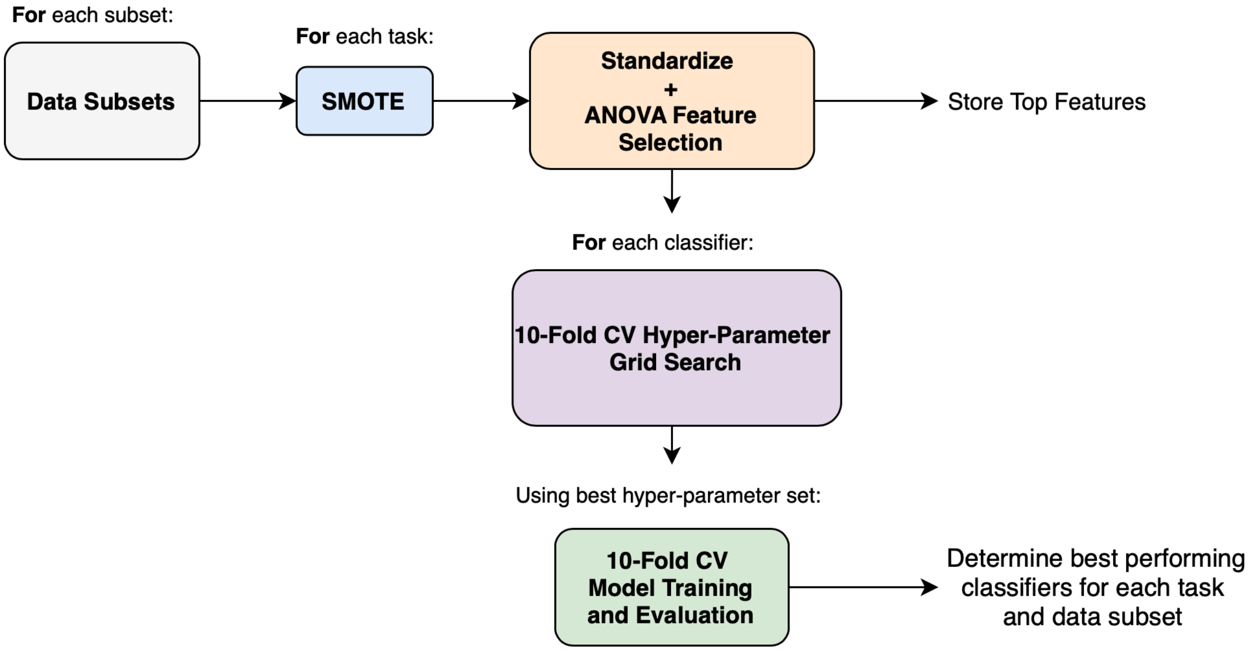

For each subset:
Data Subsets
For each task:
SMOTE
Standardize + ANOVA Feature Selection
Store Top Features
For each classifier:
10-Fold CV Hyper-Parameter Grid Search
Using best hyper-parameter set:
10-Fold CV Model Training and Evaluation
Determine best performing classifiers for each task and data subset

Figure 8: The proposed ML-pipeline for preprocessing the data, feature selection, hyper-parameter optimization and running the experiments to determine the best classifiers.

4.4 Clustering of Experienced Symptoms (RQ3)

For the second main task of this project, there were a total of 81 patients who had sufficient and valid data entries which could be used in clustering the described symptoms.

4.4.1 Determining the Embedding Level

The first step in the process was to decide at which level of the data it would be most effective to create the vector embeddings. That is, whether to create separate embeddings of all individual words, of each separate response provided to a question or to attempt and combine the responses for each patient into longer sentences describing all of their symptoms at once. Creating word embeddings would help overcome the sparsity of the data, however unigrams (single words) contain no contextual information and would likely result in less informative clusters.

An initial attempt was made to group the answers for each participant and form long sentences which would contain in-depth information about their symptoms. This method however quickly proved problematic for several reasons. Firstly, it reduces the data points to 81 samples which would likely not lead to successful clustering, especially considering the overlap in word distributions (Table 3). Secondly, BERT-based sentence embeddings are contextually and grammatically dependent, meaning that sentences which might contain the same semantic message, but are worded differently, would get different embeddings. For the purpose of this project it may have been feasible to manually create grammatically correct sentences, however this is not a scalable solution to larger data sets.

Therefore, it was decided to leave the answers as they were, and separate each answer into its own instance along with the associated CAD severity label. As there were 15 different questions, this resulted in a total of 1215 instances that clustering was performed on.

4.4.2 Clustering Pipeline

The clustering pipeline was broken down into several main stages that were ran for each of the dimensionality reduction or feature selection methods and their associated hyper-parameters. Clustering was also performed on the raw embeddings, with all experiment combinations shown below.

The first step, prior to all experiments, was to convert each of the given answers into a 768-dimensional vector using the Dutch version of BERT, also known as BERTje (Chapter 3.5.2). Following that, the clustering was either performed on the raw data, or on different combinations of feature selection and dimensionality reduction techniques. For some of the experiments the data was reduced using either PCA, ANOVA, t-SNE or UMAP alone, and for the rest, PCA or ANOVA were performed prior to running the t-SNE or UMAP reductions. When using PCA, the explained variance was set to 90%, while ANOVA selected the top 50 features.

The procedure for running the clustering on the raw, PCA or ANOVA reduced data was a bit simpler, since these methods do not have specific hyper-parameters. Once the data was reduced, it was fed into the HDBSCAN clustering algorithm. The benefit of HDBSCAN over other clustering algorithms, as mentioned earlier, is that it only requires a single hyper-parameter: the minimum cluster size. To determine the optimum minimum size, several clustering runs are performed with min_size ranging from 1 to 50. For each of those, a score is computed by looking at the number of points with assigned cluster probability of less than 0.05, divided by the full length of the data. The optimum min_size was then the one associated with the lowest score. Once clustered, samples which were not assigned to a cluster label were removed, as well as samples with cluster probabilities lower than 0.8. Then, within each unique cluster, the most commonly occurring CAD label was extracted, along with the associated sentences. This was repeated across all discovered clusters, keeping track of the sentences in their respective CAD severity levels.

When it came to clustering on the t-SNE and UMAP reduced data, the described process was simply repeated for all possible hyper-parameter combinations of the two methods. As mentioned previously, these dimensionality reduction techniques, especially t-SNE, are highly dependent on the selected parameters. Figure 9 shows the variability in the data and the difference in resulting visualisations when changing some of the parameters of t-SNE. As the combinations are numerous and there is a large overlap in the different CAD classes of the points, it was hard to determine a set of well-performing parameters. Therefore, clustering is ran on each parameter combination, and the resulting grouped sentences (clusters) are stored for each run. Throughout all the runs the frequency of the discovered clusters is also stored, and used later to compute and evaluate a ”goodness” score for the particular experiment. For the raw, PCA and ANOVA reduced experiments, these frequencies are always 1, as there is a single clustering being performed. T-SNE and UMAP on the other hand benefit from the numerous hyper-parameter combinations, forming a king of ”pooling” strategy of discovering as many clusters as possible. Figure 10 illustrates a schematic of the complete clustering pipeline.

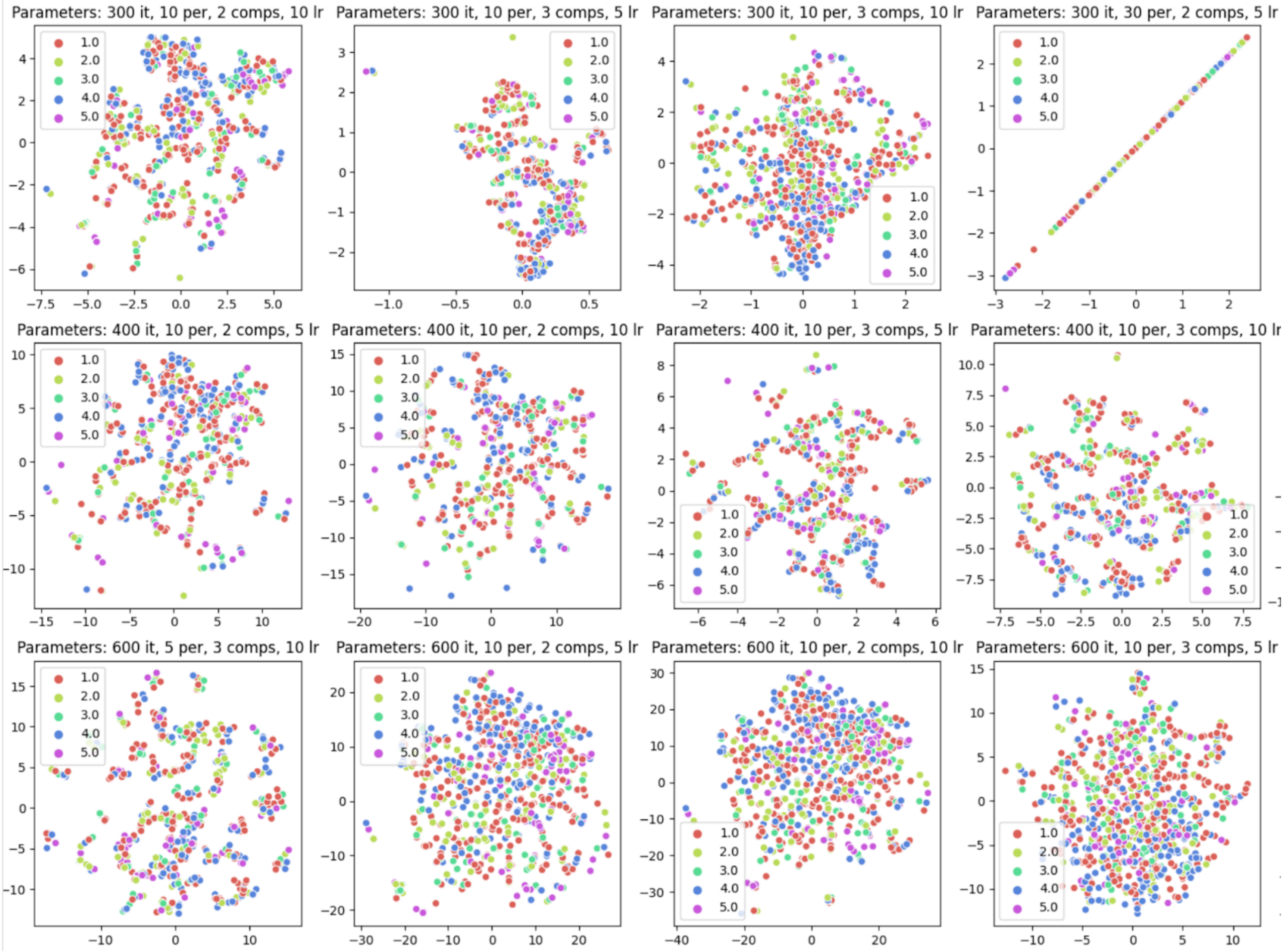


Figure 9: Visualisations of different t-SNE parameter runs on the same data. The results vary greatly, with certain cases (top right) forming clearly wrong mappings to lower dimensions. Classes overlap greatly and it is difficult to pick out a suitable and appropriate reduction.

4.4.3 Evaluating the Clustering Procedure

A schematic of the general output of the clustering procedure can be observed in Table 5. For each class label (CAD severity) the discovered clusters are displayed next to their associated frequency. It was also possible for a cluster to contain only a single sentence. Further, the total number of discovered clusters varied greatly as expected. Both the UMAP and t-SNE experiments iterated over a set of hyper-parameter combinations, resulting in many more runs and thus a higher number of discovered clusters. As many different experiments were performed, it was important to come up with some way of quantifying the discovered clusters in order to assess which method performed best.

This was done in a few steps. Firstly, clusters smaller than three sentences were removed from the output as they bring little information about the symptoms and go against the overall goal of the procedure. Following that, each cluster is assigned a similarity score for the sentences within it. The similarity measure used for this was the cosine similarity score. Cosine similarity is a measure between two non-zero vectors, in this case the sentence embeddings, and is a common measure for assessing semantic similarity in NLP tasks [72]. A cosine score of 1 means that two sentences are identical to one another. Once computed, the cosine score is multiplied by the frequency of the

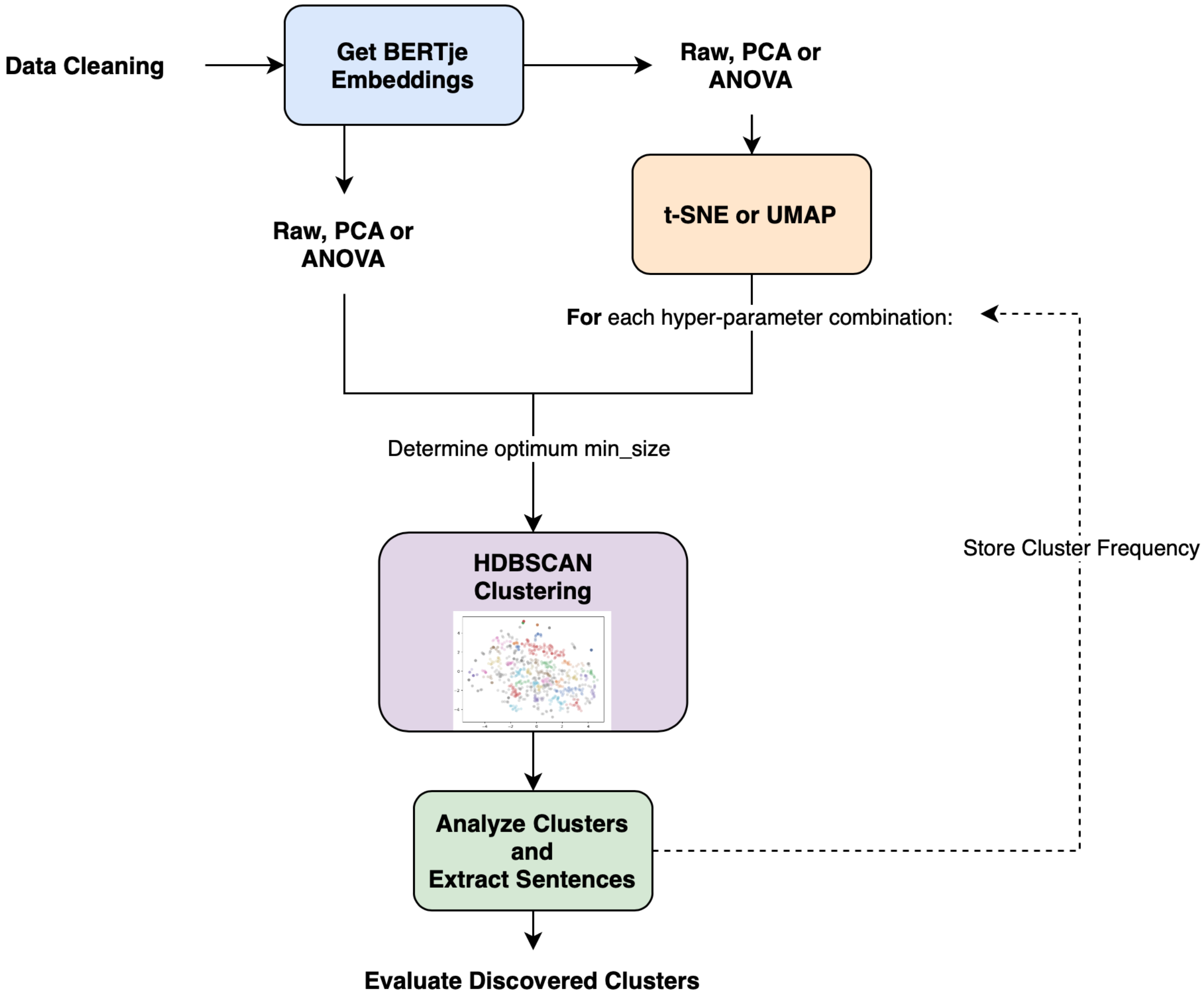

Data Cleaning
Get BERTje Embeddings
Raw, PCA or ANOVA
t-SNE or UMAP
Raw, PCA or ANOVA
For each hyper-parameter combination:
Determine optimum min_size
Store Cluster Frequency
HDBSCAN Clustering
Analyze Clusters and Extract Sentences
Evaluate Discovered Clusters

Figure 10: The proposed clustering pipeline for forming the embeddings, running several different dimensionality reductions, clustering and evaluating the discovered clusters

CAD Severity						
1		**2**		**...**	**...**	**...**
Clusters	**Frequency**	**Clusters**	**Frequency**			
('Sentence1', 'Sentence2', 'Sentence3'....)	X	('Sentence1')	X			
('Sentence1')	Y	('Sentence1', 'Sentence2')	Y			
('Sentence1', 'Sentence2')	Z	('Sentence1', 'Sentence2', 'Sentence3'....)	Z			
...	...	...	...			
...	...	...	...			

Table 5: Schematic of clustering pipeline output. For each CAD severity the discovered clusters are ordered based on their frequency. Clusters can vary in size greatly, and so can their frequencies. For the experiments on the raw, ANOVA and PCA reduced data, all frequencies are 1.

discovered cluster and this product is then summed across all discovered clusters. Finally, to account for the ”pooling” procedure of the UMAP and t-SNE experiments, this final score sum is divided by the total number of discovered clusters. Equation 2 summarises the computation of the proposed score.

$$Score = \frac{\sum_{i=1}^{n} cos\_sim(c_i) * freq(c_i)}{n} \tag{2}$$

Where c stands denotes a single cluster and n is the total number of discovered clusters. The score is designed with the hopes of giving more weight to frequently found clusters which also contain sentences that are close in meaning, while accounting for the number of runs of each experiment.

5 Results

This chapter presents the results from both experimental pipelines. The first half of the results focuses on the classifier performance with regards to predicting the CAD severity levels on the described tasks and demographic groups (**RQ1**). The top features used for classification are also presented for each experimental group and task, and differences between them are assessed briefly (**RQ2**). The second half focuses on the clustering results, by first presenting the different evaluation scores for all clustering experiments. Then, qualitative results are presented with regards to the different CAD severity levels, by summarising and interpreting the topics extracted from the discovered clusters (**RQ3**).

5.1 Classifier Performance and Selected Features (RQ1/2)

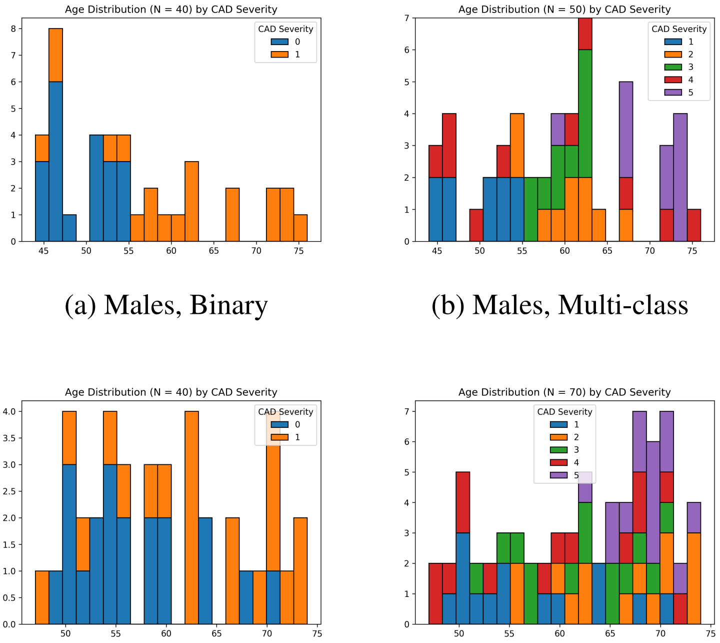
Table 6 presents a summary of the best performing classifiers across all described experiments. That is, the evaluation scores for each task type (binary or multi-label), for each of the presented experimental groups (all data, males, females, young and old). As a reminder, the younger population threshold was set to less than or equal to 62 years. The sample sizes for each group are also included in the table. Note that these sizes are prior to over-sampling, as the final sample size varied depending on the task. Patients for which the gender was unknown/missing at the time (marked as ”Other” in the data) were excluded from the gender split experiments, hence the mismatch between the total sample sizes for males and females.

Several observations can already be drawn from the table. In most cases, performance on the binary task is higher than the multi-class counterpart. This is expected, as the task is easier and only needs to predict two labels, compared to the five of the original task. This behaviour however is not observed for the Female and Young demographics, where the multi-class predictions perform better. This could be due to the way SMOTE is oversampling the data, and will be addressed in the following discussion section (Chapter 6.1). The lowest observed scores are for the multi-class predictions of the male patients, where performance is mostly in the mid 60%. For the rest of the groups, scores vary, however stay above chance level and can get as high as 93%. This suggests that the self-reported introspective quality of life scores do indeed hold some predictive power for a patient’s CAD severity.

Table 7 presents an in-depth view of the selected features for three of the demographics (All, Males and Females), ranked based on their ANOVA F-scores. Note that the table presents the top 10 features for each group, while the top 20 were used during the experiments. This is done mostly for ease of interpretation. Nevertheless, interpreting and comparing between the three groups at a first glance is difficult, and requires a good understanding of what each of the variables mean. Feature names

Data	**Task Type**	**Classifier**	**F1-score**	**Recall**	**Accuracy**
All (83)	Binary	RF	0.77	0.78	0.78
	Multi-class	KNN	0.69	0.71	0.71
Males (32)	Binary	KNN	0.82	0.84	0.84
	Multi-class	KNN	0.58	0.66	0.66
Females (35)	Binary	KNN	0.69	0.72	0.72
	Multi-class	KNN	0.79	0.83	0.83
Young (55)	Binary	SVM	0.7	0.71	0.71
	Multi-class	SVM	0.74	0.76	0.76
Old (28)	Binary	SVM	0.92	0.93	0.93
	Multi-class	KNN	0.6	0.69	0.69

Table 6: Summary of the best performing models across all classification experiments



(a) Males, Binary

(b) Males, Multi-class

(c) Females, Binary

(d) Females, Multi-class

Figure 11: Age distributions across the two tasks for both genders

starting with **eq_** are the general quality of life data, while **hqol_** relates to the specific heart quality of life complaints, both being categorical answers in most cases. The full list of features and what they mean is included in the Appendix (Table 19 and 20). A full and complete verbal comparison between the selected features for each group is provided in the discussion section of this thesis (Chapter 6.2).

There are noticeable differences between the groups. One should however mostly focus on comparing the males versus females, as the ”All” group contains samples from the ”Other” (unknown/missing) gender. There is an initial, obvious difference between the two groups, and that is in the strength of the Age predictor for the males. Across both tasks, the age variable is scored much higher compared to the rest of the features. Such a striking difference between the first and second selected feature is not observed in the females group. In fact, while age is second highest for the females in the binary task, its importance drops quite a bit in the multi-class prediction.

The strength of the age predictor for the males is overwhelmingly high and prompted a further look into the data. It could be the case that the predictions of the classifiers are solely based on the age, ignoring the information provided by the other features. Figure 11 illustrates the age distributions for the two genders, across both tasks. Note that the plots are created after over-sampling with SMOTE to accurately represent the data the classifiers make their predictions on. This is also the reason for the varying sample sizes observed in the plots. Depending on where the class imbalance lay, SMOTE would synthesize the needed instances to balance out the data. Indeed for both genders, but even more so for the males, there is a hint of separation of the classes by the specific ages. Especially for the binary task for the males, all No CAD labels are below 55 years old. For the females multi-class task, this division is not so obvious, explained by the reduced predictability of the age variable in the feature rankings.

Due to these findings, the entire ML-pipeline is repeated again for both genders, this time with the age feature removed from the data. Table 8 illustrates the results from these experiments. While the scores across all experiments have indeed decreased slightly, performance is still above chance

	Task			
	Binary		Multi-class	
Group	Feature	F-Score	Feature	F-score
All	Age	19.26	Age	21.91
	hqol_stress	3	hqol_lopen	5.27
	eq_mobiliteit	1.84	hqol_heuvel	4.92
	hqol_depressief	1.55	hqol_binnen	4.71
	hqol_bezorgd	1.47	hqol_frustratie	3.80
	hqol_bewegen	1.15	eq_mobiliteit	3.49
	hqol_lichbep	0.71	hqol_lichbep	3.41
	hqol_frustratie	0.68	hqol_tillen	2.12
	eq_angst_somb	0.59	hqol_kortademig	2.04
	hqol_heuvel	0.46	hqol_bezorgd	1.99
Males	Age	27.57	Age	13.15
	hqol_moeheid	6.62	hqol_lichbep	3.34
	hqol_tillen	5.78	hqol_lopen	3.29
	hqol_lichbep	4.10	hqol_moeheid	2.95
	hqol_kortademig	3.84	hqol_frustratie	2.48
	hqol_bewegen	3.73	hqol_tillen	2.46
	hqol_lopen	3.24	hqol_kortademig	2.14
	eq_thermometer	2.23	eq_pijn_ongemak	1.61
	hqol_heuvel	1.86	hqol_heuvel	1.57
	eq_activiteiten	1.28	hqol_actief	1.40
Females	hqol_bewegen	16.08	hqol_bewegen	12.47
	Age	8.12	hqol_lopen	11.40
	hqol_lopen	7.79	eq_mobiliteit	11.35
	hqol_lichbep	7.66	hqol_moeheid	9.25
	hqol_moeheid	7.43	hqol_frustratie	9.18
	hqol_tillen	6.97	hqol_heuvel	8.60
	hqol_heuvel	5.96	eq_thermometer	7.47
	eq_mobiliteit	5.12	hqol_lichbep	7.38
	hqol_kortademig	4.89	Age	7.38
	hqol_buiten	4.21	hqol_bezorgd	7.16

Table 7: Top 10 selected features for three of the experimental groups, ranked based on ANOVA F-scores

levels. This is a promising sign that the self-reported quality of life scores are indeed predicative of the disease levels, and additional demographic information boosts the performance of the classifiers.

Data	Task Type	Classifier	F1-Score	Recall	Accuracy
Males (32)	Binary	KNN	0.8	0.82	0.82
	Multi-class	KNN	0.54	0.62	0.62
Females (35)	Binary	KNN	0.63	0.68	0.68
	Multi-class	KNN	0.72	0.77	0.77

Table 8: Summary of best performing classifiers for both genders with feature Age removed

Table 9 presents the top 10 features for the two age groups. This time, large differences between the first and second features are observed only for the young patients in the binary task, where age is deemed as a very important factor. Rather, the F-scores for the young, multi-class and old, binary experiments are all uniformly high. Again, there are observed differences between the feature importance for each group and task. For example, the feelings of frustration (**hqol_frustratie**) is not deemed as important for the younger patients, while it boost classification for the older group. Similarly, physical obstruction (**hqol_lichbep**) seems to be more of an issue for the older group, while issues with climbing the stairs (**hqol_heuvel**) has a stronger presence in the younger population. A complete and in-depth verbal comparison between the features and groups is provided in the discussion section of the thesis (Chapter 6.2).

	Task			
	Binary		Multi-class	
Group	Feature	F-Score	Feature	F-score
Young	Age	8.62	hqol_heuvel	22.98
	hqol_lopen	3.23	Age	18.44
	hqol_stress	2.46	hqol_stress	17.86
	hqol_bezorgd	1.49	hqol_buiten	14.05
	eq_angst_somb	1.32	hqol_kortademig	9.91
	hqol_tillen	1.14	hqol_moeheid	8.38
	hqol_heuvel	1.05	hqol_tillen	7.46
	Gender	0.88	eq_thermometer	6.14
	hqol_bewegen	0.79	hqol_binnen	6.08
	hqol_kortademig	0.46	hqol_actief	5.5
Old	hqol_lichbep	18.90	Age	4.87
	hqol_lopen	17.09	eq_mobiliteit	3.39
	eq_mobiliteit	16.05	hqol_frustratie	2.70
	hqol_tillen	13.40	hqol_lopen	2.56
	hqol_frustratie	12.80	hqol_lichbep	2.21
	hqol_bewegen	9.39	eq_zelfzorg	1.96
	hqol_binnen	8.61	hqol_tillen	1.61
	hqol_depressief	6.72	eq_pijn_ongemak	1.48
	Gender	6.65	Gender	1.37
	Age	5.5	hqol_bezorgd	1.22

Table 9: Top 10 selected features for two age groups, ranked based on ANOVA F-scores

5.2 Clustering Results and Evaluation (RQ3)

Table 10 presents all clustering experiments and the associated ”goodness” scores for the discovered clusters, for each CAD severity level. The numbers within the brackets are the number of discovered clusters used to compute the score. Note that this number does not include discovered clusters of sizes smaller than three.

Without looking at the content of the clusters, one can already see the benefits of applying the proposed ”pooling” strategy and reducing the data with either t-SNE or UMAP. Simply trying to cluster once on the raw, PCA or ANOVA reduced data leads to poor results. This is both in terms of the number of clusters found (in some cases 0) and in the proposed scoring of the discovered cluster quality. It is not surprising that some of the experiments failed to discover any clusters. Figure 9 and Table 3 demonstrated the uniformity of the data, with little noticeable differences between the classes in terms of the described complaints. Running a single clustering procedure on these embeddings will likely fail to capture anything meaningful and only discover a small number of clusters for the classes with higher data points.

Overall, the number of clusters found for all experiments also reflects the size of the data for each CAD severity (Table 3). No CAD and Moderate CAD both have higher number of samples, and therefore the method is able to discover more clusters. The high number of discovered clusters can also be slightly deceiving, as many of these clusters have a frequency of 1 and could result from a poor hyper-parameter configuration. As these clusters will likely be ignored in the qualitative interpretation of the results, and in fact they could be skewing the evaluation scores, the experiments with t-SNE and UMAP are repeated from the beginning. This time, prior to computing the scores, all clusters

	CAD Severity					
Experiment	**No CAD**	**Minimum CAD**	**Mild CAD**	**Moderate CAD**	**Severe CAD**	**Average Score**
Raw data	0.951 (5)	1 (1)	0 (0)	0.946 (1)	0 (0)	0.5794
PCA	0.951 (5)	1 (1)	0 (0)	0.946 (1)	0 (0)	0.5794
ANOVA	0.940 (7)	0.872 (3)	1 (1)	1 (2)	0 (0)	0.7623
t-SNE	1.332 (3492)	1.832 (343)	1.275 (147)	1.386 (1043)	0.955 (127)	1.356
PCA + t-SNE	1.553 (2915)	2.059 (284)	1.166 (124)	1.540 (828)	1.557 (116)	1.575
ANOVA + t-SNE	1.178 (3701)	1.712 (489)	1.348 (232)	1.194 (1139)	1.65 (209)	1.416
UMAP	1.441 (6722)	1.942 (720)	1.625 (298)	1.398 (2306)	1.462 (288)	1.573
PCA + UMAP	1.441 (7032)	1.7185 (863)	1.803 (319)	1.418 (2373)	1.51 (351)	1.578
ANOVA + UMAP	1.239 (7881)	1.7936 (1013)	1.652 (399)	1.185 (2980)	1.477 (483)	1.469

Table 10: Evaluating the quality of discovered clusters across all experiments. Number in bracket refers to number of clusters used in computing the score.

	CAD Severity					
Experiment	**No CAD**	**Minimum CAD**	**Mild CAD**	**Moderate CAD**	**Severe CAD**	**Average Score**
t-SNE	9.07 (237)	13.21 (29)	7.61 (11)	9.24 (73)	3.36 (9)	8.49
PCA + t-SNE	**10.2 (241)**	**11.53 (34)**	**4.43 (12)**	**10.24 (64)**	**6.28 (17)**	**8.54**
ANOVA + t-SNE	7.92 (221)	8.35 (61)	5.96 (24)	7.41 (69)	7.08 (30)	7.35
UMAP	7.95 (473)	12.12 (64)	8.8 (25)	6.88 (169)	5.69 (33)	8.3
PCA + UMAP	**8.03 (481)**	**11.45 (66)**	**10.54 (29)**	**7.17 (169)**	**6.44 (36)**	**8.73**
ANOVA + UMAP	7.98 (385)	7.95 (131)	8.64 (39)	5.64 (173)	5.41 (61)	7.12

Table 11: Rerunning the clustering pipeline and computing evaluation scores with cluster frequencies of 3 or higher. Rows in bold represent the two best performing experiments.

with frequencies smaller than three are also filtered out of subsequent analysis. The results of these can be seen in Table 11.

It is difficult to select the best performing experiment. While PCA+UMAP has the highest proposed average evaluation score, it performs worse on the No and Moderate CAD category compared to the second highest experiment, PCA+t-SNE. On the other hand, it greatly outperforms in the Mild CAD category. PCA+UMAP also seems to discover more clusters. This is actually in line with the previously mentioned differences between the two methods, as UMAP tends to preserve a better global structure and produce more distant and better formed clusters (Chapter 3.4.1). In the end, PCA+UMAP is selected for further analysis, based on the higher average score, as well as greater number of discovered, and likely more separated, clusters to pull information from.

5.2.1 Clustering Results: Most Frequent Sentences

Attempting to manually compare hundreds of discovered clusters and sentences between the five CAD severity levels is unlikely to yield any accurate and complete interpretations. An attempt here is made to capture and highlight some differences between the different CAD levels.

A simple way of doing this is by looking at the most frequent sentences across all discovered clusters. Since the clusters are not mutually exclusive (sentences often repeat across clusters), an assumption can be made that a commonly occurring topic or complaint within a severity will likely be discovered several times by the procedure. Table 12 displays the top 50 (where possible) most frequent sentences from the discovered clusters. While the table contains a lot of information at a first glance, it does make it easier to compare the clusters across the different levels. It is important to note that each one of these sentences does in fact come from a cluster containing other, similar sentences, as the final clustering results were limited to cluster sizes of three or more. The results presented in this table can be built upon in the later discussion section, by sourcing the rest of the sentences within these clusters and ordering the emerging topics based on these frequencies. A few initial observations can be made from the table. Firstly, the pain rating scores are captured for a lot of the CAD levels. These pain scores actually reflect the score distributions quite nicely (Figure 7). For the No CAD category, 5, 2 and 6 come out on top, and indeed they are they are reported most frequently. The same for Min CAD, and partially for Moderate CAD, where the most reported score of 5 is still within the top 10 sentences. Shorter time spans are reported for the No CAD level, and sentences mentioning the genetic factors and past family deaths are only observed in the Moderate and Severe CAD category. The unique presence of the knee location is also seen in the Min CAD category. At the same time, there is a lot of observed overlap between the complaints and symptoms. All of these findings are reflected in the word distributions of the data (Table 3). A more complete list of clusters and topics, based on the ranked sentences here, is presented in the latter discussion (Chapter 6.3.2).

CAD Severity									
No CAD		Min CAD		Mild CAD		Moderate CAD		Severe CAD	
5	88	inspanning	29	stress	9	8	28	het zakte van zelf weer weg	12
2	77	3	15	rust	9	7	27	gaat vanzelf weer over	11
6	69	moe	15	pijn neemt toe tot een gevoel ik probeer kalm te blijven en het gaat na een uur weer weg	6	inspanning en stress	14	ging weer vanzelf over	11
geen idee	61	extra inspanning	12	pijn trekt vanzelf weer weg het opgejaagde gevoel wordt minder op het moment dat ik iets doe maar zodra ik weer relax komt het weer terug	6	het gaat vooral om zorgen mbt het hart een jongere broer 57 is eind vorig jaar plotseling overleden aan een hartinfarct hij had geen klachten laten merken maar zijn hart bleek voor meer dan 50 verkalkt een oudere broer moest het jaar ervoor plotseling gedotterd worden zelf heb ik zonder de simvastatine ook een te hoog cholesterol gehalte	11	het is iedere dag en gaat niet over	9
dagelijks	48	ongerust	11	het komt op een onverwachts moment als ik mij opgejaagd voel probeer ik mijn te concentreren op iets anders het gaat weg als ik afgeleid ben	5	continu	10	het komt en het gaat	9
1	40	weet ik niet	9	het is spontaan ontstaan en is ook weer weg	4	vanwege stress komen ook hartaandoeningen en beroertes voor in de familiegeschiedenis van zowel mijn moeder als vader	10	het straalt niet uit	8
vijf tot tien minuten	39	kort	7	rusten	3	5	9	werd er door wakker	8
zwaar werk	36	knie rug en schouderklachten	7	vermoeidheid	3	druk op de borst met soms uitstraling naar hals en schouderblad	9	ik maak me geen zorgen maar omdat het al eerder is gebeurd wilde ik het laten nakijken	5
3	35	moe heid	6	het begint slecht te voelen maar verdwijnt na een minuut of twee	3	benen en handen zwetend over het hele lichaam	8	soms meerdere keren per dag en soms niet	4
spanning	34	hoge bloeddruk	6	tijdens een etentje met veel mensen kreeg ik inéén het gevoel dat ik niet goed werd na een paar minuten zakte het weer weg	3	erfelijke aanleg mijn vader heeft op de leeftijd die ik nu heb ook een hartinfarct gehad en is op 63 jarige leeftijd ook aan en hartinfarct overleden ook mijn ooms van vaders kant hadden hart problemen misschien een stukje voeding ik eet denk ik niet heel ongezond maar zou beter kunnen	8	soms elke dag dan weken weer niet	4
4	29	dagelijks	4	dat is denk al meer dan een half jaar geleden en is alleen maar erger geworden	3	ik denk met het hart of met een vernauwing in de aderen dit ook omdat mijn moeder opa oma en tante op jonge leeftijd zijn gedotterd	8	erg bezorgd vader had een hartstilstand broer stierf aan hartfalen en moeder had een pacemaker	4
inspanning	28	de pijn is de hele dag aanwezig en ik vind het moeilijk om te ontspannen en me geen zorgen te maken	4	twee keer	3	ik maak me een beetje zorgen omdat mijn vader is overleden aan vernauwde kransslagaders tijdens de wintersport	8	ik maak me zorgen omdat hartfalen erfelijk is van mijn vaders kant vader en zijn twee broers en zussen stierven aan hartfalen	4
op de borst	27	oefeningen in de sportschool	4	het komt plots op en zakt dan langzaam af	2	hele dag door	8	ik heb geen lust meer om wat te doen	4
midden op de borst achter borstbeen	24	bij traplopen en sporten of bij zwaar werk	4	uren	2	huishouden stofzuigen	8	ik probeer er op uit te gaan	4
rondom het borstbeen	22	schouderbladen en rug	4	pijn en druk op de borst tussen de ribben in kortademig pijn tussen de schouderbladen pijn blijft lange tijd aanwezig geen pijnscheuten	2	naar de linkerarm	7	het kwam van zelf door dat ik last kreeg van mijn hoofd	4
vijf tot vijftien minuten	20	regelmatig	3	twee dagen	2	naar de arm	7	2	3
tussen tien en twintig minuten	20	continu	3	wanneer in rust	2	operatie en stress	7	er wordt getwijfeld of er toch nog iets anders is dan de astma klachten	3
tien tot vijftien minuten	19	linkerarm	3			alle dagen	7	geen idee misschien erfelijk bepaald	3
ik maak me er zorgen om je wordt belemmert in je doen en laten	19	maak je geen zorgen omdat de toestand verbetert maar denk nog steeds niet dat het normaal is	3			elke nacht tijdens rust	7	vijf tot tien minuten soms meerdere keren per dag en soms niet	2
aan de linkerkant boven mij linkerborst	19	ik kan overdag niets doen omdat ik me zo moe en niet lekker voel	3			diep van binnen en achter de ribben	7	overlijden moeder broer mijn bruidsjongen	2
rustig aan doen	18	paracetamol	3			kan niet op de linkerkant liggen de beweging van de vingers op de linkerhand is beperkt	7	ik sta te trillen op de benen ik heb geen grip meer op de grond	2
8	18	linker arm	3			sommige dagelijkse activiteiten zoals stofzuigen	7	ja ik ben astma patient	2
elke dag	17	arm heel af en toe been	3			na inspanningen	7	ik maak mij hele grote zorgen hier over want het gaat maar niet over	2
borstkas en linkerarm en schouderblad	17	naar arm of pols en benen bij kortademigheid bij het strottehoofd gebied	3			tuinieren zoals harken en gras knippen bukken en tillen knutselen in huis zoals werken boven mijn hoofd	7	deels maak ik me zorgen wat er nu speelt	2
ik weet niet wat ik heb daar maak ik me zorgen over	17	knieen	3			aanleiding was knutselen en tuinieren tillen bukken en het tillen van zware voorwerpen	7	ik voel mij elke dag ziek en word dan bang om dood te gaan	2
soms een uur soms de gehele dag	17	naar mij linker arm	3			rondom het maagkuiltje bij de rechter aanhechtingsspier aan mijn ribbenkast lopend vanaf het maagkuiltje	6		
twee keer per week	17	soms naar de schouderbladen	3			ietsje hoger dan de borsten naar het middenrif toe	6		
vier tot vijf keer per week	17	zal niet weten	2			iedere dag	6		
arm schouderblad en nek	16	dat verschild per dag	2			laat drie ribben aan zowel de linker als de rijzijde zakken	6		
aan de linkerkant van het borstbeen langs het schouderblad	16	per dag verschild	2			hoofd en iets boven mijn linker tepel op borst	6		
tijdens inspanning	16	heftig	2			bij mijn borstbeen en onder het sleutelbeen en schouderblad	6		
maak me af en toe zorgen	15	inspanning snel een paar honderd meter lopen traplopen	2			na operatie	6		
in het middenrif	15	kan niet werken omdat ik zo misselijk en moe ben	2			ik maak me zorgen omdat ik niet weet waardoor ik deze klachten heb weet ook niet wat ik moet doen op zon pijnaanval moment	6		
links op de borst onder de tepel	15	voeten knieën linkerheup handen borst en keel	2			ik heb net een gastroscopie gehad met een positieve uitslag wel iets buik vet rond mijn maag scheurtje in het middenrif ben aan het krimpen waardoor alle organen kennelijk volgens de mdl arts wat in de knel komen te zitten	6		
twee of drie keer in de week	15	hijgen gevoel of de keel dichtgeknepen wordt	2			duizeligheid ongeveer een minuut pijn op de borst afwisselend van enkele minuten tot een kwartier tot half uur	5		
soms maak ik mij zorgen erom ik ben er veel mee bezig om het te voorkomen	15	lichte pijn in de hartstreek uitstralend naar linker arm	2			van een paar minuten tot paar uur	5		
regelmatig	14	pijn op de borst	2			varieert van binnen enkele minuten tot langere tijd	5		
borst arm en linkerkaak	14	lopen en lichamelijke inspanning gaan minder goed fietsen gaat wel redelijk	2			van mild tot bijna ernstig	5		
soms paar minuten soms enkele seconden	14	knie rug en schouder	2			naar de rechterarm	5		
gedurende de inspanning en korte tijd erna	14	4	2			ongeveer acht maanden	5		
soms maanden niet maar soms ook twee of drie keer per dag	14	kortademigheid bij traplopen en fietsen	2			naar binnenkant hand en vingers	5		
twee tot drie keer	14	druk op de borst straalt door naar hart	2			alleen deze ene keer	5		
bij lichte inspanning	14	borst	2			gehele bovenlichaam voornamelijk aan de linkerzijde	5		
vaak	13	naar linkerarm	2			bij activiteiten zoals huishoudelijk werk of tillen	5		
overgang	13					tussen en net boven de borsten	5		
twee keer in de afgelopen vier weken	13					6	5		
tussen de vijf en tien keer in het afgelopen jaar	13					totale rust fysio	5		
momenteel ongeveer een keer per maand	13					ik denk dat ik bang en gestrest was dat ik mijn werk niet kon doen	5		

Table 12: Most frequent sentences from the discovered clusters across all CAD severity levels

6 Discussion

This part of the thesis provides a discussion for all three parts of the results. That is, describing and discussing the classification results, verbalising the differences between the selected features for each of the demographics, and providing an extensive overview of the clusters and topics found within each CAD severity level. Possible pitfalls and current limitations are presented and lastly suggestions and ideas for future work on how the project can be improved and described.

6.1 CAD Classification

Looking at the classification results across all experiments (Table 6 and Table 8), it is a promising start to see that all scores are above chance-level. This is especially important for the multi-class tasks, where the models had to predict five different classes. This finding suggests that the self-reported quality of life answers are somewhat predictive of CAD. Further, as will be shown in the following section, it is not the case that the classifications are all based on a single predictor, such as the age. Rather, there are differences across the groups and different features carry varying strengths for each demographic.

It seems like there is an added benefit of trying to predict CAD separately across groups, rather than keeping all patients together. This is noted by the higher scores contained in each group compared to the full data (All) experiments. It is worth mentioning that this could also be due to the way the data is split during testing, or due to some skewing created by the over-sampling. Since a 10-fold cross validation procedure was ran, the models had to predict a larger portion of the data for the All experiments, as the splits resulted in larger portions being held out as a test set. For groups such as the Males and Old, where the data was scarce, a binary task prediction would result in a test split of around 3-4 samples. This is comparatively easier than trying to predict 8-9 samples for the full data experiments, thus increasing the chances of a higher score. To counter this, the second largest experimental group is the younger population, where the binary scores are actually slightly lower and the multi-class predictions are higher than the full data experiments.

Another observation is that for two of the groups (Females and Young) the multi-class task scores are higher than the binary ones. A possible reason for this could be due to the class-imbalance (Table 4) and the way the data is over-sampled. For both of those groups, during the multi-class task, more than 50% of the data was actually synthesized by SMOTE. There has been some evidence that in such extreme cases SMOTE actually overgeneralizes the models [73], meaning that the classifiers perform inaccurately ”well” on the fake data. When faced with real world examples, this performance would actually decrease. To overcome this, it is recommended that SMOTE is only performed on the training data, leaving the validation and testing sets untouched. However this was not possible in this case, with the Young population only have 2 Severe CAD patients.

There are cases though, such as the Young and All data binary tasks, where the classes were more balanced, with less synthesized data. Seeing as there the performance of the classifier was still relatively high, this suggests that the quality-of-life scores do indeed hold some predictability for CAD.

It is also difficult to compare the scores of the gender and age-specific experiments, as patients for which the gender information was missing at the time (”Other” in the data) have been excluded from the first split. The gender distribution for the younger population is 21 males, 22 females and 12 ”Other”, which is a reasonable portion of the data. This distribution, and mostly the balance between the genders, resembles the original data the most, hence the higher similarity in scores.

Overall however, as there are such high class imbalances which cause the creation of ”fake” data, as well as the sparsity issue, it is difficult to credit the success/failure of the model classifications in respect to each demographic and hence the difference in scores. Nevertheless, results across the full data set do show some success in predictions and a discussion into the varying feature importance for each group is also provided.

6.2 Demographic Differences

Both Tables 7 and 9 give a good representation of the features deemed important for classification for all experimental groups. Using these, a discussion is presented that describes potential differences between the demographics. Information gathered here can be useful in accommodating earlier detection and tailoring specific treatments for each group. While the results of the classification experiments presented here (Table 6) are not as high as those in the relevant research on the topic [9], the data is unique and much easier to acquire. Further, the results are above chance level across most tasks, therefore drawing and discussing differences with regards to the features is worthwhile and likely valid.

Starting with the gender-specific differences (Table 7). As noted in the results, age seems to be an overwhelmingly strong predictor for males. This is in-line with past research, as it has been shown that men exhibit signs of calcium accumulation at earlier ages than women, and the calcium build up is overall faster [74]. Further, as mentioned in the risk demographics section (Chapter 2.3), there are higher rates of CAD for males at younger ages, with this difference between the genders evening out with increased age [15]. The average male age in the data set is 57.5 years, compared to 60.4 years for the females. Indeed there is a small difference between the two, and this could also be a factor for the high feature F-score.

Nevertheless, it was also shown that classification is not purely based on the patient age (Table 8), supporting the importance of the other features. Comparing the two genders for the binary tasks, difficulties experienced during exercising or physical activity (**hqol_bewegen**) are rated much higher for the female group. Similarly, but to a lesser extent, the same can be said for difficulties with walking a 100m in a single pass (**hqol_lopen**). On the other hand, experiences with a lack of energy (**hqol_moeheid**) are ranked slightly higher for the males, and also difficulties breathing and shortness of breath (**hqol_kortademig**). The scores on a scale of 1-100 with regards to how the patient was feeling on the day about their health (**eq_thermometer**) is only observed for the males, and the same for reported difficulties with problems with daily activities (**eq_activiteiten**). On the other hand, issues with walking (**eq_mobiliteit**) and working in the garden or house (**hqol_buiten**) are only observed for the female patients. In summary, it seems like issues with physical activities, exercising and walking are stronger predictors for females in the binary CAD classification, whereas for males it is more about the lack of breath or energy that they experience.

Some of these differences can be transferred to the multi-class task as well. For starters, difficulties with exercising (**hqol_bewegen**) is still the highest predictor for females, and is in fact now unique to the gender. Males do report more physical obstruction (**hqol_lichbep**) than females this time, and this could potentially overlap with the difficulties in physical tasks and exercising. For both genders feelings of frustration (**hqol_frustratie**) are now a predictive factor, which was not the case in the binary task. The shortness of breath (**hqol_kortademig**) is now even more pronounced for the males as it becomes unique, similarly to difficulties with moving heavy objects (**hqol_tillen**) and reports of pain or discomfort (**eq_pijn_ongemak**). For the females, issues with walking up or climbing staircases is rated higher (**hqol_heuvel**) and the other previously unique issue with walking (**eq_mobiliteit**) is now an even stronger predictor. Females seem to express feelings of worry (**hqol_bezorgd**) and the scores they provide in regards to how they felt about their health on that day (**eq_thermometer**) gain importance too. Both of these were not observed in the binary task features. Finally, it seems like the age is no longer as important as in the binary task, and no where near as important compared to that of the male patients. In summary, the general features mostly follow the patterns in the binary task. Male features are more concerned with shortness of breath, physical obstruction and difficulties performing heavy load work. For the females, difficulties with exercising, walking, climbing stairs and expressions of worry about their symptoms seem to have more importance during classification.

A similar discussion is presented for the feature differences observed between the younger and older populations (Table 9). It is important to look at the gender distributions for each, as it could be the case that an older population contains mostly females for example, in which case the differences might be skewed and repeated from before. The younger population contains 21 males and 22 females, while for the older these are 11 and 13 members respectively. As these are mostly balanced, feature differences discussed forward should capture actual differences between the ages.

Differences across the binary task start with the age of the patient, which seems much more important for the younger population. There are overall a higher number of features which are unique for each group, compared to the previous gender cases. For the younger patients these are feelings of worry (**hqol_bezorgd**), reports of anxiety or depression (**eq_angst_somb**), issues with climbing stairs (**hqol_heuvel**) and shortness of breath (**hqol_kortademig**). Unique features for the older patients start with reports of physical obstruction (**hqol_lichbep**), which is also the strongest predictor for the task. This is followed by issues with walking (**eq_mobiliteit**), expressions of frustration (**hqol_frustratie**), issues with walking inside the house on the same floor (**hqol_binnen**) and expressions of depression (**hqol_depressief**). Difficulties with exercising and physical activities (**hqol_bewegen**) are also ranked slightly higher for the older population. The differences here present an interesting pattern, which is also somewhat expected. The older population reports mostly physical issues, difficulties walking and the resulting mental effect this would have on them, such as frustration and depressive thoughts. The younger population on the other hand, reports issues with worries, anxiety and depression. Further, they report issues with more physically demanding tasks, such as walking up stairs and the resulting shortness of breath.

The feature rankings for both age groups change for the multi-class task, with unique variables being highlighted in some cases. For the young patients, age is still an important predictor, however now comes second to issues with climbing stairs (**hqol_heuvel**). The new unique features are issues with working around the house and gardening (**hqol_buiten**), an experienced lack of energy and tiredness (**hqol_moeheid**), walking around the same house floor (**hqol_binnen**) and a wider difficulty with a range of tasks (**hqol_actief**). The shortness of breath (**hqol_kortademig**) is still a unique complaint for that group, being ranked even higher than in the binary task. It also seems that the general score of well-being they report (**eq_thermometer**) is also contributing to the classification. When looking at the older patients, there is very little overlap in the features, similarly to the binary task. Contrary to the binary task though, age is now the strongest predictor. Following that, and uniquely, problems with walking (**eq_mobiliteit**), expressions of frustration (**hqol_frustratie**), difficulty with walking a 100m in a single pass (**hqol_lopen**), physical obstruction (**hqol_lichbep**), issues with washing and clothing oneself (**eq_zelfzorg**, reports of pain and discomfort (**eq_pijn_ongemak**) and general worries (**hqol_bezorgd**) are all features present for the older population. In summary, it seems like the younger patients now exhibit slightly more issues with physical activities, mostly in regards to walking and completing activities around the house, compared to the binary task, with the shortness of breath being present in both. As for the older patients, a lot of the same features remain as in the binary task, with even more emphasis on difficulties with easy tasks (washing and clothing oneself) and higher reports of pains and discomfort which worry them.

6.3 Cluster Topics

6.3.1 Evaluating the Overall Method

Tables 13, 14, 15, 16 and 17 summarise the topics and cluster examples for all CAD severity levels. These topics were extracted from Table 12, by taking each sentence in the ranked order and searching for the associated clusters in the output file of the clustering pipeline (PCA+UMAP). The ordering of topics here is based on the frequency of the discovered sentence, following the assumption that if a particular topic was more common in the data, the clustering procedure would discover it more frequently at each hyper-parameter run. This however is just indicative and should not be taken concretely, as it seems that the overall results and number of discovered clusters greatly depend on the availability and nature of the data. Another thing to note is that many of the discovered sentences in Table 12 were related to each other their frequencies overlap. ”Spanning” and ”Inspanning” from the No CAD column were usually found in the same clusters, and were therefore grouped for the presented topic in the respective topic summary table (Table 13). Another note to make is that, for sake of interpretation and readability, not all cluster examples are presented. For topics such as the ”Location of symptoms”, this would result in unreadable tables and too many examples to follow and compare against. These tables were formed mainly with the purpose of trying to extract some differences between the CAD levels, as well providing a qualitative assessment of the cluster pipeline.

The first message to draw from the tables is the success of the proposed clustering procedure, especially the ”pooling” strategy behind it. The experiments which did not use either UMAP or t-SNE discovered a maximum of 7 clusters (Table 10), with some of those clusters containing hundreds of sentences. It seems like the method greatly benefits from iterating over numerous parameters. The hyper-parameter sets here were not selected with some special care, both UMAP and t-SNE are notorious for having parameters which are hard to interpret and understand how they would affect the final output. Instead, sufficient ranges were provided for iteration, with the idea that configurations which worked well would also form more meaningful clusters more frequently.

Additionally, the formed clusters are indeed meaningful. The sentences for almost all presented examples are related to each other semantically and make it relatively easy to understand what topic the particular cluster is associated with. Not only is this the case for sentences which are short and relatively similar in wording (”Dagelijks” and ”dagelijks vermoeid”), but also for sentences of varying lengths and completely different syntax (”geen idee” and ”ik zou niet weten”). This is especially impressive for topics describing how the pain of the complaints progresses (Table 15), or the capturing of the genetic predispositions and family issues (Table 16). As highlighted in the results section too, the captured pain scores in these tables also reflect the most frequently reported pains (Figure 7), mostly staying in line with the reported distributions.

It is not to say that the clusters are perfectly formed though. There are cases where a sentence in the clusters is unrelated to the rest of the members (”Varieer, geen specifiek patroon” and ”rust”, Table 15), or actually completely opposite in meaning (”Het is iedere dag en gaat niet over” and ”het zakte van zelf weer weg” in Table 17). While opposite in meaning, they both are still somewhat related to each other, talking about the progression of the pain.

One can also argue that the topics are too broad. This is most evident in cases about the location of symptoms, across all CAD severity levels. In fact, these are the clusters where a lot of examples were omitted from presentation, as a lot of the complaints were deemed as very similar. Locations such as the chest, sternum, left arm were extremely common, and finding subtle and specific differences between those would require extensive and cumbersome manual analysis. It is also hard to say whether this is due to the method not being sensitive enough to highlight the differences between the CAD levels, or it is simply down to the word distribution in the data (Table 3). For example, the clustering was able to capture the knee location complaint in Min CAD, which was also unique for that category.

Nevertheless, this is a successful clustering on a very limited data set. It is successful as it captures and forms meaningful clusters, while also not being limited to outputting single words such as typical topic modeling methods. Being limited to single words can greatly limit the interpretation of results. This success is also largely due to the embeddings and their ability to capture semantic information, supporting the use of language models and their versatility across domains. Embeddings also do not require the preprocessing of the data, which is something commonly required to get the most of the topic modeling.

6.3.2 Differences in CAD Levels

Using the topic tables, a quick discussion is presented to try and highlight some differences and notable features for each severity. Starting with the No CAD (Table 13) category. The presented pain scores vary, however are all in the mid ranges, not suggesting any severe pains. Patients seem to report that they are not aware of what causes their complaints (or what reduces/enhances them), and the experienced problems seem to occur on a daily basis. While the complaints are frequent, they do not seem to last a long time, highlighted by the numerous reported time-spans of several minutes. When reporting a cause of symptoms, this seems to be due to stress or increased effort and high intensity work. The sternum (”borstbeen”) is a commonly mentioned location, amongst the other usual chest and arm reports, where the patients experience pain and pressure. The patients are also worried about their symptoms, however, report that resting and taking it easy seems to subside the patterns. Finally the table also captures some contradictions to the regards of the daily occurrence of symptoms. It seems like symptoms also vary in their duration and in cases occur only several times in the week.

No CAD	
Topic	**Cluster Examples**
Pain Ratings	(’5’, ’5’, ’5’, ’5’, ’5’, ’5’); (’2’, ’2’, ’2’, ’2’); (’6’, ’6’, ’6’, ’6’, ’6’); (’4’, ’5’, ’6’, ’6’, ’5’, ’5’, ’5’, ’6’, ’4’, ’5’, ’4’, ’5’, ’6’, ’6’)
Unknown Causes/Relief	(’Geen idee’, ’geen idee’, ’Geen idee’, ’geen idee’, ’Geen idee’); (’niet lang’, ’Ik zou niet weten’, ’Geen idee’, ’geen idee’, ’Geen idee’, ’geen idee’, ’Geen idee’)

Frequency of symptoms (daily, often)	(’Dagelijks’, ’dagelijks vermoeid’, ’Dagelijks’, ’Dagelijks’, ’dagelijks’);
	(’regelmatig’, ’Dagelijks’, ’Dagelijks’, ’Vaak’, ’Dagelijks’, ’actief zijn’);
	(’regelmatig’, ’dagelijks vermoeid’, ’Vaak’);

Duration of symptoms (in the minutes)	(’vijf tot tien minuten’, ’vijf tot vijftien minuten’, ’tien tot vijftien minuten’, ’vijf tot tien minuten’, ’tussen tien en twintig minuten’); (’Ongeveer twintig minuten’, ’Ongeveer een kwartier’, ’vijf tot vijftien minuten’, ’tien tot vijftien minuten’, ’vijf tot tien minuten’, ’tussen tien en twintig minuten’)

Causes of symptoms	(’pijn alleen bij stres’, ’Zwaar werk’, ’Zwaar werk’);
	(’tijdens inspanning’, ’Zwaar werk’, ’Zwaar werk’, ’bij lichte inspanning’);
	(’Dagelijks op het werk’, ’dagelijks vermoeid’, ’Op mijn werk’, ’Zwaar werk’, ’druk, gestrest en vermoeid zijn’, ’Druk en gestrest’, ’Zwaar werk’);
	(’Overgang’, ’Inspanning’, ’Inspanning’, ’Flinke Inspanning’);

Location of symptoms	(’Midden op de borst achter borstbeen’, ’In het middenrif’, ’Rondom het borstbeen’); (’Druk op mijn borstkas. Niet pijnlijk.’, ’Pijn, drukkend gevoel borstkas met uitstraling naar mijn arm.’, ’Druk op de borst, ongemakkelijk gevoel. En benauwd.’) (’Midden op de borst achter borstbeen’, ’aan de linkerkant boven mij linkerborst’, ’Rondom het borstbeen’)
Worried/Hindered	(’ik maak me er zorgen om. Je wordt belemmert in je doen en laten.’, ’Ik weet niet wat ik heb daar maak ik me zorgen over’, ’Maak me af en toe zorgen’) (’Soms maak ik mij zorgen erom. Ik ben er veel mee bezig om het te voorkomen.’, ’ik maak me er zorgen om. Je wordt belemmert in je doen en laten.’, ’Ik probeer kalm te blijven omdat ik me hier zorgen over maak’)
Rest helps	(’rustig aan doen’, ’rustig aan doen’, ’Rustig aan blijven doen’); (’rustig aan doen’, ’rustig ademen en blazen’, ’rustig aan doen’, ’Rustig aan blijven doen’); (’rustig aan doen’, ’rust nemen’, ’rustig ademen en blazen’, ’rustig aan doen’, ’Rustig aan blijven doen’)
Varied occurrence/ symptoms	(’soms een uur, soms de gehele dag’, ’Soms paar minuten soms enkele seconden’, ’gedurende de inspanning en korte tijd erna’, ’Slechts een moment’, ’soms maanden niet maar soms ook twee of drie keer per dag’, ’Varieert, soms drie keer per week, soms even geen last.’); (’soms een uur, soms de gehele dag’, ’soms maanden niet maar soms ook twee of drie keer per dag’, ’Varieert, soms drie keer per week, soms even geen last.’, ’Meestal lichte druk. Soms wat intenser om vervolgens af te nemen’);
Frequency of symptoms (less frequent)	(’twee keer per week’, ’vier tot vijf keer per week’, ’Momenteel ongeveer een keer per maand’); (’Een paar uur per dag’, ’twee keer per week’, ’Momenteel ongeveer een keer per maand’); (’twee keer per week’, ’vier tot vijf keer per week’, ’twee tot drie keer’);

Table 13: Some of the main topics and frequent cluster examples for the No CAD class. Enclosing brackets in the cluster examples relate to one discovered cluster.

For the Minimum CAD category (Table 14), the most commonly captured topic is the increased effort and fatigue which causes or is caused by the symptoms. The pain ratings are noticeably lower than the No CAD category which is not something one would expect. Again, the patients do not seem to be sure of what is causing the issues, and report locations predominantly on the left side of the body.

Here, a unique complaints about knee pain is observed and captured by the clustering, something not observed elsewhere in the data (Table 3). A topic of increased blood pressure emerges as well, which is also inline with the word distributions. Frequency of symptoms seems to be daily and continuous, causing worries and hindering the patients. Finally, efforts and difficulties with sports and physical activity are also reported, unlike in the No CAD category.

Min CAD	
Topic	**Cluster Examples**

Effort/Tired and worrried	('moe', 'kort', 'extra inspanning', 'inspanning', 'inspanning', 'ongerust');
	('extra inspanning', 'inspanning', 'inspanning');
	('moe', 'Moe heid', 'kort', 'extra inspanning', 'inspanning', 'inspanning', 'ongerust');

Pain Ratings	('3', '3', '3');
	('3', '3', '2', '3', '3');
	('3', '4', '4', '3', '3', '3');

Unsure of causes	('weet ik niet', 'weet ik niet', 'weet ik niet'); ('weet ik niet', 'Zal niet weten', 'weet ik niet', 'weet ik niet', 'Zal het niet weten');
Location of symptoms (knee in particular)	('Linker arm', 'knie, rug en schouderklachten', 'Voeten, knieen, linkerheup, handen, borst en keel'); ('Lichte pijn in de hartstreek uitstralend naar linker arm', 'knie, rug en schouderklachten', 'naar arm of pols en benen. Bij kortademigheid bij het strottehoofd gebied'); ('Linker arm', 'knie, rug en schouderklachten', 'knieen', 'Voeten, knieen, linkerheup, handen, borst en keel', 'knie, rug en schouder', 'midden borst', 'Schouderbladen en rug'); ('knie, rug en schouderklachten', 'knieen', 'knie, rug en schouder', 'Schouderbladen en rug') ('Lichte pijn in de hartstreek uitstralend naar linker arm', 'Pijn op de borst', 'naar arm of pols en benen. Bij kortademigheid bij het strottehoofd gebied'); ('naar linkerarm', 'Naar mij linker arm', 'soms naar de schouderbladen', 'arm heel af en toe been')
Blood Pressure	('Bij opnemen van bloeddruk', 'hoge bloeddruk', 'hoge bloeddruk'); ('hoge bloeddruk', 'hoge bloeddruk', 'door overmatige inspanning')
Frequency of symptoms	('dagelijks', 'Regelmatig, continu', 'dagelijks');
Worried/Hindered	('Maak je geen zorgen omdat de toestand verbetert, maar denk nog steeds niet dat het normaal is', 'De pijn is de hele dag aanwezig en ik vind het moeilijk om te ontspannen en me geen zorgen te maken.', 'Ik kan overdag niets doen omdat ik me zo moe en niet lekker voel', 'Voelt niet fijn en maak mij zeker zorgen', ''Kan niet werken omdat ik zo misselijk en moe ben');
Physical activity (stairs, sports)	('Oefeningen in de sportschool', 'Bij traplopen en sporten of bij zwaar werk', 'Inspanning. Snel een paar honderd meter lopen. Traplopen'); ('kortademigheid bij traplopen en fietsen', 'Oefeningen in de sportschool', 'Bij traplopen en sporten of bij zwaar werk', 'Inspanning. Snel een paar honderd meter lopen. Traplopen')

Table 14: Some of the main topics and frequent cluster examples for Minimum CAD. Enclosing brackets in the cluster examples relate to one discovered cluster.

The number of discovered clusters for Mild CAD (Table 15) were some of the smallest across all categories, despite Severe CAD having less patients and less total number of words (Table 3). Nevertheless, Table 15 still demonstrates some clear clusters. Stress and effort are again the top topic, as was the case in Minimum CAD. Rest is reported quite frequently as a way of easing the symptoms, and there is a clear pattern in how the symptoms progress. It seems like the complaints come on suddenly, last for a few minutes to half an hour, and then the pain subsides on its own. Finally, the chest and left arm are again reported as locations of symptoms.

Mild CAD	
Topic	**Cluster Examples**

Stress and effort	('Stress', 'Stress', 'Stress');
	('vermoeidheid', 'uren', 'stress');
	('vermoeidheid', 'stress', 'inspanning');
	('vermoeidheid', 'uren', 'stress', 'zware inspanning');

Rest	('rusten', 'Rust', 'rust'); ('Varieer, geen specifiek patroon', 'rusten', 'Rust', 'rust'); ('rusten', 'Rust', 'rust', 'wanneer in rust')
Symptom progression (pain increases, then goes away, spontaneous)	('Het begint slecht te voelen, maar verdwijnt na een minuut of twee', 'Pijn neemt toe tot een gevoel, ik probeer kalm te blijven en het gaat na een uur weer weg', 'Pijn trekt vanzelf weer weg. Het opgejaagde gevoel wordt minder op het moment dat ik iets doe. Maar zodra ik weer relax komt het weer terug.', 'Het komt op een onverwachts moment. Als ik mij opgejaagd voel probeer ik mijn te concentreren op iets anders. Het gaat weg als ik afgeleid ben', 'Het ene moment kant dat een half uur aanhouden, het andere moment is het met vijf minuten weg. Het onrustige gevoel heb ik afgelopen weken soms wel elke dag of nacht gehad', 'tijdens een etentje met veel mensen kreeg ik ineen het gevoel dat ik niet goed werd. Na een paar minuten zakte het weer weg.') ('het komt plots op en zakt dan langzaam af', 'Gaat vanzelf weg', 'Het is spontaan ontstaan en is ook weer weg.', 'Het is spontaan ontstaan en is ook weer weg.')
Location of symptoms	('op de borst en linkerarm', 'midden borst en mond', 'Op de borst linker arm')

Table 15: Some of the main topics and frequent cluster examples for Mild CAD. Enclosing brackets in the cluster examples relate to one discovered cluster.

The topics for Moderate CAD are a bit more varied (Table 16). The highest pain score clusters are reported here, and there is a clear topic of genetic predisposition and past family issues. This is not something observed in the less severe categories, and one would expect it to be an issue for increasing levels of the disease. This cluster is as well one of the best examples of how successful the clustering method has been. Symptom duration is again continuous, lasting all day and into the night. In terms of the location of symptoms, aside from the usual, the right hand, jaw and fingers are now mentioned as well. Patients also report problems with lying and sleeping on their left side. The symptoms seem to radiate through the whole body, causing sweating. Patients mention household activities such as gardening, vacuuming, that they are now unable to perform. Stress and increased effort are present again, and a separate topic of more specific locations (slightly above the nipple) can be highlighted. Naturally the patients are worried about their symptoms, which seem to progress from minutes to hours.

Moderate CAD	
Topic	**Cluster Examples**

Pain Ratings	('7', '8', '8', '7', '7', '8'); ('7', '7', '7'); ('5', '5', '5');
	('6', '6', '6');

Genetic, past family cases	('Het gaat vooral om zorgen m.b.t. het hart. Een jongere broer (57) is eind vorig jaar plotseling overleden aan een hartinfarct. Hij had geen klachten (laten merken), maar zijn hart bleek voor meer dan 50% verkalkt. Een oudere broer moest het jaar ervoor plotseling gedotterd worden. Zelf heb ik zonder de Simvastatine ook een te hoog cholesterol gehalte. ', 'Vanwege stress komen ook hartaandoeningen en beroertes voor in de familiegeschiedenis van zowel mijn moeder als vader', 'erfelijke aanleg. Mijn vader heeft op de leeftijd die ik nu heb ook een hartinfarct gehad, en is op 63 jarige leeftijd ook aan en hartinfarct overleden. ook mijn ooms van vaders kant hadden hart problemen. misschien een stukje voeding. ik eet denk ik niet heel ongezond maar zou beter kunnen. ', 'Ik denk met het hart of met een vernauwing in de aderen. Dit ook omdat mijn moeder, opa, oma en tante op jonge leeftijd zijn gedotterd.', 'Ik maak me een beetje zorgen omdat mijn vader is overleden aan vernauwde kransslagaders tijdens de wintersport.')
Symptom duration (mostly continuous/ all day and night)	('Verschilt', 'Continu', 'Continu'); ('Hele dag door', 'Alleen deze ene keer', 'Hetzelfde, continu', 'Het is constante druk'); ('continu met uitschieters', 'Hetzelfde, continu', 'Het is constante druk'); ('continu met uitschieters', 'Varieert, maar meestal dagelijks', 'Het is constante druk'); ('Hele dag door', 'Alle dagen', 'Elke nacht tijdens rust', 'Alleen deze ene keer') ('Hele dag door', 'Alle dagen', 'Elke nacht tijdens rust', 'Iedere dag.', 'Altijd')
Location of symptoms	('pijn op linkerborst, linkerarm en -oksel en achter linker schouderblad en CTO gebied; 's avonds dikke voeten. Kan niet op linker schouder liggen want krijg het dan benauwd.' , 'druk op de borst met soms uitstraling naar hals en schouderblad', 'Kan niet op de linkerkant liggen, de beweging van de vingers op de linkerhand is beperkt', 'druk op de borst en op dezelfde hoogte op de rug, schouder', 'tussen en net boven de borsten', 'Pijn op de borst tot aan de arm'); ('naar de rechterarm.', 'naar de linkerarm', 'naar de arm'); ('in kaak en armen', 'naar de rechterarm.', 'naar binnenkant hand en vingers', 'naar de linkerarm', 'naar de arm'); ('in kaak en armen', 'naar de linkerbovenarm. Soms helemaal door naar de handen en vingers ', 'naar de rechterarm.', 'naar binnenkant hand en vingers', 'nauwelijks. soms naar buikstreek', 'naar de linkerarm', 'naar de arm')
Radiating/whole body	('Klachten rond het kuiltje in de buik en de spieraanhechting aan mijn rechter rib. 'S Nachts een gevoel van een rots rond de maag, die naar boven uitstraalt.' , 'Door het hele lichaam', 'benen en handen, zwetend over het hele lichaam', 'benen en handen, zwetend over het hele lichaam')
Housework/ lifting	('Sommige dagelijkse activiteiten, zoals stofzuigen', 'Huishouden, stofzuigen. ', 'Zware objecten tillen. Krachtsinspanningen. Normaal huishoudelijk werk.', 'Bij activiteiten zoals huishoudelijk werk of tillen.'); ('tuinieren, zoals harken en gras knippen (bukken) en tillen. knutselen in huis, zoals werken boven mijn hoofd.', 'Huishouden, stofzuigen. ', 'aanleiding was knutselen en tuinieren, tillen, bukken en het tillen van zware voorwerpen', 'Als ik aan het stofzuigen ben of iets tillen');

Stress/Effort	('inspanning en stress', 'inspanning en stress', 'Operatie en stress.');
	('inspanning en stress', 'inspanning en stress', 'Overwerk, spanning en werkdruk', 'Operatie en stress.',
	'bij middelmatige inspanning',);

Location (more specific, slightly above nipple/chest)	('Hoofd en iets boven mijn linker tepel op borst', ''links van mn linker korst z'm vijf centimeter boven mijn tepellijn'', 'Ietsje hoger dan de borsten naar het middenrif toe.')
Worried/hindered	('Ik denk dat ik bang en gestrest was dat ik mijn werk niet kon doen.', 'Ik maak me zorgen omdat ik niet weet waardoor ik deze klachten heb weet ook niet wat ik moet doen op zo'n pijnaanval moment.', 'Ik kan niet vol uit het leven leven zoals ik dat graag wil. Ik voel me nooit echt fit. Bij te weinig beweging voel ik me neerslachtig worden. Ik raak gefrustreerd. En ik maak me ernstig zorgen.', 'Ik maak me zorgen omdat ik wil weten wat er aan de hand is. Verder ben ik gezond, met een goede hartslag en bloeddruk', 'Ik maak me zorgen omdat ik niet weet wat het is, ik gelukkig getrouwd ben en vader van 3 kinderen. Voor het dagelijks leven loop ik rond met een gevoel van als ik maar niet zomaar er tussen uit piep',)
Symptom progression (minutes to hours)	('Duizeligheid ongeveer een minuut, pijn op de borst afwisselend van enkele minuten tot een kwartier tot half uur', 'Van een paar minuten tot paar uur', 'Varieert van binnen enkele minuten tot langere tijd', 'Van mild tot bijna ernstig');

Table 16: Some of the main topics and frequent cluster examples for Moderate CAD. Enclosing brackets in the cluster examples relate to one discovered cluster.

Finally, despite Severe CAD having the least amount of data, a few interesting topics can still be observed (Table 17). There are clear reports of the symptoms going away on their own, which could be an indication of a stable angina episode. Such episodes usually happen after overexertion, and symptoms tend to be temporary, lasting typically around 15 minutes. There is however also a report of the symptoms never going away. There are strong expressions of worry and the symptoms seem to vary greatly in their duration and frequency, with a report of the aforementioned 10 to 15 minute window of possible stable angina. There is another example of genetic predisposition and past family issues. Finally, despite reporting to be hindered, patients also seem to give a very low pain score of 2 (in line with Figure 7), and some mention asthma related issues.

Indeed, while there is a noticeable overlap between the topics across all groups, there are also some observable differences. Both the overlaps and differences are nicely reflected in the word distributions, which is indicative of the clustering pipeline working well and revealing the true nature of the data. The method naturally benefits from having more instances and patient reports, so it is a promising tool which can be applied on future, larger iterations of the CONCRETE data.

Severe CAD	
Topic	**Cluster Examples**
Pain goes away on its own (NOTE: "Het is iedere dag en gaat niet over." often in clusters too. Opposite meaning)	('het zakte van zelf weer weg', 'gaat vanzelf weer over.', 'ging weer vanzelf over'); ('Het straalt niet uit.', 'het komt en het gaat.', 'het zakte van zelf weer weg', 'gaat vanzelf weer over.', 'ging weer vanzelf over'); ('Het straalt niet uit.', 'Het is iedere dag en gaat niet over.', 'werd er door wakker', 'het zakte van zelf weer weg', 'gaat vanzelf weer over.', 'ging weer vanzelf over')
Worried	('Ik maak mij hele grote zorgen hier over want het gaat maar niet over.', 'deels maak ik me zorgen wat er nu speelt.', 'Ik maak me geen zorgen maar omdat het al eerder is gebeurd wilde ik het laten nakijken.')
Symptom duration/frequency	('vijf tot tien minuten. soms meerdere keren per dag en soms niet.', 'soms meerdere keren per dag en soms niet.', 'soms elke dag dan weken weer niet'); ('de laatste drie maanden', 'soms meerdere keren per dag en soms niet.', 'soms elke dag dan weken weer niet');
Genetic and family issues	('overlijden moeder, broer, mijn bruidsjongen', 'Erg bezorgd. Vader had een hartstilstand, broer stierf aan hartfalen en moeder had een pacemaker', 'Ik maak me zorgen omdat hartfalen erfelijk is van mijn vaders kant. Vader en zijn twee broers en zussen stierven aan hartfalen.')
Hindered	('Ik sta te trillen op de benen ik heb geen grip meer op de grond.', 'Ik heb geen lust meer om wat te doen.', 'ik probeer er op uit te gaan', 'Ik voel mij elke dag ziek en word dan bang om dood te gaan')
Pain rating	('2', '2', '2')
Asthma patients	('Ja, ik ben astma patient. ', 'weet ik niet ik heb altijd gedacht dat het bij mijn Astma hoorde.', 'Niets, het wordt niet beter')

Table 17: Some of the main topics and frequent cluster examples for Severe CAD. Enclosing brackets in the cluster examples relate to one discovered cluster.

6.4 Future Work

The future directions for the CONCRETE project are numerous. In fact, one of the main take away messages from the presented results is the potential ahead and what can be achieved with more data. As it has been outlined several times throughout the thesis, most methods used would benefit with a wider range of samples, equally distributed across all CAD severity levels. A more balanced data set would result in more reliable results, rather than attempting to draw conclusions and discussion based on experiments where large portions of the data have been synthesised. Further, differences between the CAD severity levels could be subtle in certain cases. For an ML model to pick out on these differences, having more data would only benefit the training process.

CAD is a syndrome which develops over time, seemingly getting worse with age, especially for the female population [15]. The CONCRETE project design has already taken this into account, by requiring the patients to come back at different month intervals (6, 12 and 24) and answering the same quality of life questions. Unfortunately due to the infancy of the project, this has only been achieved for a fraction of the participating patients so far. Future iterations of the project should aim to include these variables into the model training and feature analysis. If the answers to these questions have changed from the initial check-up, they are likely to contain valuable information about the progressing state of the patients. This in turn could reveal useful information about changes in the demographic groups, while also aiding the model predictions.

Another point of improvement could be the addition of varying demographic information. A big motivation of the CONCRETE project is to gather data which is non-invasive and easily obtainable. Looking back at the described risk factors for CAD (Chapter 2.3), information about the patient's eating habits and preferences, exercise routines, past blood pressure problems and whether they smoke and if they do, how often. Other more general demographic information, such as the weight or nationality, could also be added and it might indirectly capture some of the risk factors. Again, these efforts have already been initiated for most patients, but entries are limited. The importance of a single additional demographic factor (Age) has already been demonstrated (Table 7 and 9), so additional ones might only boost the already observable predictability strength of the quality-of-life answers. Further, the nature of this demographic data is in line with the CONCRETE goals, in being easily obtainable and non-invasive.

The clustering results proved to be successful in terms of grouping similar complaints together and reflected the uniformity of the data. Table 11 displayed that more data results in more clusters, as one would expect. Where the currently presented method falls short is in the interpretation and topic extraction from the numerous clusters. The results presented here mostly focused on the most commonly found clusters and thus the most frequently present topics in the data. It is possible that the differences between the CAD groups are more subtle, being captured less frequently by the clustering and thus buried by the very common and broad clusters. Ideally the method sensitivity should be improved. A large improvement can already be made at the final output stage, by finding overlapping clusters and grouping them together. For example, looking at Table 17, the first topic clusters show a clear overlap in the discovered sentences, and all of these can be grouped in a single, larger cluster. For starters, this would greatly reduce the number of clusters which need to be analyzed manually at the end, and could give more weight to the subtle differences. It would also highlight poorly and wrongly clustered sentences, outlining where the method falls short.

More improvements to the clustering can be achieved by keeping track of the question type of each sentence and also from which patient the complaint came from. Keeping track of the question type improves interpretation (it is hard to know without manually checking what ”geen idee” refers to), and could reveal interesting distributions of the question types for each CAD category. A form of feature selection, which outlines which questions seem to be more relevant for the difference disease levels. Keeping track of the patients is useful as it would increase the validity of the topics. In the current state of the data, there are cases where a patient would give a repeating answer to different questions. For example, when asked to report what reduces or increases their symptoms, patients often state that they do not know. As the answer is the same, and the clustering does not take the question type into account, this forms somewhat inaccurate clusters. By keeping track of the question type and patient ID, potential duplicates which skew the clustering could be removed. Again, this could increase the sensitivity of the method.

Finally, patients should be encouraged more strongly to stick to the format of the question and to not provide unnecessary information. A possible word limit for each answer could be introduced, forcing them to be more concise. The clustering was successful in part due to the manual clean up of the data, which introduced some limited structure. In larger scenarios this may not be feasible. A reworking of the questions may be beneficial too. Currently some of the questions carry some ambiguity in how they could be interpreted, and at a first glance some of them may overlap. For example, **open_aanleiding** asks for a clear cause of symptoms, while **open_reden** asks for a reason of the experienced complaints, and patients often provided the same answer to both. Ambiguities are difficult for humans to handle, let alone a machine, and should be eliminated at all possible stages of the process.

A final word on the overall direction of the project. The possibilities and potential of the results form a promising basis for web-based assessment of CAD. A web tool could be created in which patients seeking help are asked to fill out a questionnaire, providing the GP with some risk assessment and overall confidence of the disease. Access to this information, prior to the first check-up, can accommodate better preparation and an earlier diagnosis. Further, it could save unnecessary trips and referrals to expensive and invasive testing, improving the overall efficiency of the system.

7 Conclusion

The concluding messages of these thesis are as follows. The classifier results across all experiments indicated that easily obtainable, introspective and self-reported quality of life questions can indeed be used to assess in the prediction and classification of a patient's CAD risk (**RQ1**). While these scores are not perfect predictors, the used data set is still very much in its infancy. Additional demographic information and the further collection of data is likely to boost the performance and allow for more concrete conclusions. Patterns of feature importance were shown between the experimental groups that somewhat represent and reflect their social roles and stances (**RQ2**). A more solid grasp of these patterns can be used in the future to accommodate the early detection, as well as tailor treatment approaches better to the patient. Finally, a ”pooling”-like clustering method was proposed which achieved success in discovering clusters of similar complaints from a mostly unstructured and free-form text data (**RQ3**). With the method refined in future iterations, and the inclusion of additional instances, greater differences across the CAD groups could be discovered to gain further insight into the disease and how it affects the local population.

Bibliography

[1] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[2] M. L. Giger, “Machine learning in medical imaging,” *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 512–520, 2018.

[3] E. Long, H. Lin, Z. Liu, X. Wu, L. Wang, J. Jiang, Y. An, Z. Lin, X. Li, J. Chen, *et al.*, “An artificial intelligence platform for the multihospital collaborative management of congenital cataracts,” *Nature biomedical engineering*, vol. 1, no. 2, pp. 1–8, 2017.

[4] K. Veropoulos, *Machine learning approaches to medical decision making*. PhD thesis, University of Bristol, 2001.

[5] P. Libby and P. Theroux, “Pathophysiology of coronary artery disease,” *Circulation*, vol. 111, no. 25, pp. 3481–3488, 2005.

[6] R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, *Braunwald’s heart disease e-book: A textbook of cardiovascular medicine*. Elsevier Health Sciences, 2011.

[7] A. Schmermund, S. Möhlenkamp, and R. Erbel, “Coronary artery calcium and its relationship to coronary artery disease,” *Cardiology clinics*, vol. 21, no. 4, pp. 521–534, 2003.

[8] M. Akay, “Noninvasive diagnosis of coronary artery disease using a neural network algorithm,” *Biological cybernetics*, vol. 67, no. 4, pp. 361–367, 1992.

[9] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. M. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, and U. R. Acharya, “Machine learning-based coronary artery disease diagnosis: A comprehensive review,” *Computers in biology and medicine*, vol. 111, p. 103346, 2019.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[11] T. Shi, K. Kang, J. Choo, and C. K. Reddy, “Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations,” in *Proceedings of the 2018 World Wide Web Conference*, pp. 1105–1114, 2018.

[12] E. S. Kayi, K. Yadav, and H.-A. Choi, “Topic modeling based classification of clinical reports,” in *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pp. 67–73, 2013.

[13] Y. Meng, W. Speier, M. Ong, and C. W. Arnold, “Hcet: Hierarchical clinical embedding with topic modeling on electronic health records for predicting future depression,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1265–1272, 2020.

[14] Y. Lu, P. Zhang, J. Liu, J. Li, and S. Deng, “Health-related hot topic detection in online communities using text clustering,” *Plos one*, vol. 8, no. 2, p. e56221, 2013.

[15] W. B. Kannel and P. S. Vokonas, “Demographics of the prevalence, incidence, and management of coronary heart disease in the elderly and in women,” *Annals of epidemiology*, vol. 2, no. 1-2, pp. 5–14, 1992.

[16] W. Kannel, “Lipids, nutrition and cardiovascular disease in the elderly,” 1988.

[17] D. Lawlor, C. Bedford, M. Taylor, and S. Ebrahim, “Geographical variation in cardiovascular disease, risk factors, and their control in older women: British women’s heart and health study,” *Journal of Epidemiology & Community Health*, vol. 57, no. 2, pp. 134–140, 2003.

[18] J. Pu, K. G. Hastings, D. Boothroyd, P. O. Jose, S. Chung, J. B. Shah, M. R. Cullen, L. P. Palaniappan, and D. H. Rehkopf, “Geographic variations in cardiovascular disease mortality among asian american subgroups, 2003–2011,” *Journal of the American Heart Association*, vol. 6, no. 7, p. e005597, 2017.

[19] J. V. Tu, A. Chu, L. Maclagan, P. C. Austin, S. Johnston, D. T. Ko, I. Cheung, C. L. Atzema, G. L. Booth, R. S. Bhatia, *et al.*, “Regional variations in ambulatory care and incidence of cardiovascular events,” *Cmaj*, vol. 189, no. 13, pp. E494–E501, 2017.

[20] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[21] L. E. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[22] E. Fix and J. L. Hodges Jr, “Discriminatory analysis-nonparametric discrimination: Small sample performance,” tech. rep., California Univ Berkeley, 1952.

[23] R. Todeschini, D. Ballabio, V. Consonni, and F. Grisoni, “A new concept of higher-order similarity and the role of distance/similarity measures in local classification methods,” *Chemometrics and Intelligent Laboratory Systems*, vol. 157, pp. 50–57, 2016.

[24] K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, and N. Kerdprasop, “An empirical study of distance metrics for k-nearest neighbor algorithm,” in *Proceedings of the 3rd international conference on industrial application engineering*, pp. 280–285, 2015.

[25] N. Lopes and B. Ribeiro, “On the impact of distance metrics in instance-based learning algorithms,” in *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 48–56, Springer, 2015.

[26] D. T. Larose and C. D. Larose, *Discovering knowledge in data: an introduction to data mining*, vol. 4. John Wiley & Sons, 2014.

[27] L.-Y. Hu, M.-W. Huang, S.-W. Ke, and C.-F. Tsai, “The distance function effect on k-nearest neighbor classification for medical datasets,” *SpringerPlus*, vol. 5, no. 1, pp. 1–9, 2016.

[28] L. Rokach and O. Maimon, “Decision trees,” in *Data mining and knowledge discovery handbook*, pp. 165–192, Springer, 2005.

[29] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[30] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, “Decision trees: an overview and their use in medicine,” *Journal of medical systems*, vol. 26, no. 5, pp. 445–463, 2002.

[31] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[32] T. M. Berhane, C. R. Lane, Q. Wu, B. C. Autrey, O. A. Anenkhonov, V. V. Chepinoga, and H. Liu, "Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory," *Remote sensing*, vol. 10, no. 4, p. 580, 2018.

[33] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] G. Riddick, H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, and H. A. Fine, "Predicting in vitro drug sensitivity using random forests," *Bioinformatics*, vol. 27, no. 2, pp. 220–224, 2011.

[35] G. Nimrod, A. Szilágyi, C. Leslie, and N. Ben-Tal, "Identification of dna-binding proteins using structural, electrostatic and evolutionary features," *Journal of molecular biology*, vol. 387, no. 4, pp. 1040–1053, 2009.

[36] J. D. Cohen, L. Li, Y. Wang, C. Thoburn, B. Afsari, L. Danilova, C. Douville, A. A. Javed, F. Wong, A. Mattox, *et al.*, "Detection and localization of surgically resectable cancers with a multi-analyte blood test," *Science*, vol. 359, no. 6378, pp. 926–930, 2018.

[37] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[38] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*, vol. 52, pp. 456–462, 2019.

[39] L. Wang, *Support vector machines: theory and applications*, vol. 177. Springer Science & Business Media, 2005.

[40] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (svm)," *Informatics in Medicine Unlocked*, vol. 16, p. 100200, 2019.

[41] M. E. Ozer, P. O. Sarica, and K. Y. Arga, "New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines," *Omics: a journal of integrative biology*, vol. 24, no. 5, pp. 241–246, 2020.

[42] W. S. Noble, "What is a support vector machine?," *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[43] K.-B. Duan, J. C. Rajapakse, and M. N. Nguyen, "One-versus-one and one-versus-all multiclass svm-rfe for gene selection in cancer classification," in *European conference on evolutionary computation, machine learning and data mining in bioinformatics*, pp. 47–56, Springer, 2007.

[44] D. Kobak and P. Berens, "The art of using t-sne for single-cell transcriptomics," *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[45] D. Agis and F. Pozo, "A frequency-based approach for the detection and classification of structural changes using t-sne," *Sensors*, vol. 19, no. 23, p. 5097, 2019.

[46] T. Howley, M. G. Madden, M.-L. O'Connell, and A. G. Ryder, "The effect of principal component analysis on machine learning accuracy with high dimensional spectral data," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 209–222, Springer, 2005.

[47] P. Paokanta, N. Harnpornchai, S. Srichairatanakool, and M. Ceccarelli, "The knowledge discovery of [beta]-thalassemia using principal components analysis: Pca and machine learning techniques," *International Journal of e-Education, e-Business, e-Management and e-Learning*, vol. 1, no. 2, p. 169, 2011.

[48] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[49] Y. Wu, P. Tamayo, and K. Zhang, "Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding," *Cell systems*, vol. 7, no. 6, pp. 656–666, 2018.

[50] P. Hamel and D. Eck, "Learning features from music audio with deep belief networks.," in *ISMIR*, vol. 10, pp. 339–344, Citeseer, 2010.

[51] A. R. Jamieson, M. L. Giger, K. Drukker, H. Li, Y. Yuan, and N. Bhooshan, "Exploring nonlinear feature space dimension reduction and data representation in breast cadx with laplacian eigenmaps and-sne," *Medical physics*, vol. 37, no. 1, pp. 339–351, 2010.

[52] M. Balamurali, K. L. Silversides, and A. Melkumyan, "A comparison of t-sne, som and spade for identifying material type domains in geological data," *Computers & Geosciences*, vol. 125, pp. 78–89, 2019.

[53] M. Balamurali and A. Melkumyan, "t-sne based visualisation and clustering of geological domain," in *International Conference on Neural Information Processing*, pp. 565–572, Springer, 2016.

[54] L. Li, *Functional Interrogation of Ventral Tegmental Area (VTA) Gaba Neurons in Anxiety and Novelty Detection, and Classification of Neuron Dendrite Types by Electrophysiological Properties*. PhD thesis, State University of New York at Buffalo, 2020.

[55] N. Pezzotti, B. P. Lelieveldt, L. Van Der Maaten, T. Höllt, E. Eisemann, and A. Vilanova, "Approximated and user steerable tsne for progressive visual analytics," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 7, pp. 1739–1752, 2016.

[56] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-sne effectively," *Distill*, vol. 1, no. 10, p. e2, 2016.

[57] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172, Springer, 2013.

[58] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, *et al.*, "Dynabench: Rethinking benchmarking in nlp," *arXiv preprint arXiv:2104.14337*, 2021.

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

[60] J. Lilleberg, Y. Zhu, and Y. Zhang, “Support vector machines and word2vec for text classification with semantic features,” in *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pp. 136–140, IEEE, 2015.

[61] S. K. Sienčnik, “Adapting word2vec to named entity recognition,” in *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pp. 239–243, 2015.

[62] D. Zhang, H. Xu, Z. Su, and Y. Xu, “Chinese comments sentiment classification based on word2vec and svmperf,” *Expert Systems with Applications*, vol. 42, no. 4, pp. 1857–1863, 2015.

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

[64] Y. Sharma, G. Agrawal, P. Jain, and T. Kumar, “Vector representation of words for sentiment analysis using glove,” in *2017 international conference on intelligent communication and computational techniques (icct)*, pp. 279–284, IEEE, 2017.

[65] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, “Learning gender-neutral word embeddings,” *arXiv preprint arXiv:1809.01496*, 2018.

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

[67] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, “Bertje: A dutch bert model,” *arXiv preprint arXiv:1912.09582*, 2019.

[68] I. Chades and S. Nicol, “Small data call for big ideas,” *Nature*, vol. 539, no. 7627, pp. 31–31, 2016.

[69] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, “Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance,” *Neural networks*, vol. 21, no. 2-3, pp. 427–436, 2008.

[70] K. R. Foster, R. Koprowski, and J. D. Skufca, “Machine learning, medical diagnosis, and biomedical engineering research-commentary,” *Biomedical engineering online*, vol. 13, no. 1, pp. 1–9, 2014.

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

[72] C.-H. Huang, J. Yin, and F. Hou, “A text similarity measurement combining word semantic information with tf-idf method,” *Jisuanji Xuebao(Chinese Journal of Computers)*, vol. 34, no. 5, pp. 856–864, 2011.

[73] H. Guan, Y. Zhang, M. Xian, H.-D. Cheng, and X. Tang, “Smote-wenn: Solving class imbalance and small sample problems by oversampling and distance scaling,” *Applied Intelligence*, vol. 51, no. 3, pp. 1394–1409, 2021.

[74] R. L. McClelland, H. Chung, R. Detrano, W. Post, and R. A. Kronmal, “Distribution of coronary artery calcium by race, gender, and age: results from the multi-ethnic study of atherosclerosis (mesa),” *Circulation*, vol. 113, no. 1, pp. 30–37, 2006.

Appendices

Variable Name	Variable Description	Allowed Answer
open_soort	Kunt u aangeven met wat voor soort pijn- of gevoelsklacht(en) u naar de huisarts bent gegaan? *Can you explain with what kind of complaints (pain/feelings) you consulted your GP?*	open ended answer OR 9. unknown
open_andereklachten	Heeft u naast de door u beschreven pijn-of gevoelsklacht(en) nog andere klachten ervaren? *Did you experience other complaints? In addition to the already described complaints (pain/feelings)*	open ended answer OR 9. unknown
open_duur	Hoe lang duren de door u beschreven pijn-of gevoelsklacht(en)? *How long is the duration of the by you described complaints (pain/feelings)?*	open ended answer OR 9. unknown
open_locatie	Waar zit de door u beschreven pijn- en of gevoelsklacht(en) precies? *What is the location of the by you described complaints?*	open ended answer OR 9. unknown
open_uitstralen	Stralen de door beschreven pijn- of gevoelsklacht(en) uit? (zo ja, waar naar toe) *Do the by you described complaints radiate to other areas in your body? If so, to which body part?*	open ended answer OR 9. unknown
open_ernst	Hoe ernstig zijn de door u beschreven pijn- of gevoelsklacht(en) op een schaal van 1 (mild) tot 10 (zeer ernstig)? *How severe are the by you described complaints on a scale of 1 (mild) to 10 (very severe)?*	Scale 1-10
open_activiteiten	Moet u door de door u beschreven pijn- of gevoelsklacht(en) bepaalde activiteiten laten welke u gewoonlijk wel deed? En welke activiteiten zijn dit? *Do you have to leave certain activitities that you would normally perform due to the by you described complaints? If so, can you describe these activities?*	open ended answer OR 9. unknown
open_hoevaak	Hoe vaak heeft u de door u beschreven pijn-of gevoelsklacht(en)? *How often do you experience your complaints?*	open ended answer OR 9. unknown
open_verloopperkeer	Hoe verlopen de door u beschreven pijn- of gevoelsklacht(en) per keer? *How do your complaints progress each time you experienced them?*	open ended answer OR 9. unknown
open_ontstaanklacht	Kunt u de situatie beschrijven waarbij u voor het eerst de door u beschreven pijn- of gevoelsklacht(en) voelde? *Can you describe the situation in which you experienced your complaints for the first time?*	open ended answer OR 9. unknown
open_aanleiding	Was er naar uw mening een duidelijke aanleiding voor de door u beschreven pijn- of gevoelsklacht(en)? *Was there a clear cause for the your experienced complaints?*	open ended answer OR 9. unknown
open_reden	Waar denkt u dat de door u beschreven pijn- of gevoelsklacht(en) mee te maken hebben? *What do you think is the reason for the experienced complaints?*	open ended answer OR 9. unknown
open_klachtenminder	Waardoor worden de door u beschreven pijn- of gevoelsklacht(en) minder? *What reduces your experienced complaints?*	open ended answer OR 9. unknown
open_klachtenerger	Waardoor worden de door u beschreven pijn- of gevoelsklacht(en) erger? *What enhances your experienced complaints?*	open ended answer OR 9. unknown
open_dagelijksleven	Wat betekenen de door u beschreven pijn- of gevoelsklacht(en) voor dagelijkse leven en/of maakt u zich hier zorgen over? *How do the complaints influence your daily live? Do you feel concerned about it?*	open ended answer OR 9. unknown

Table 18: Summary of the variable names for the open ended, experienced complaints questions. Answers to these were used for the clustering experiments.

Quality of Life Questions		
Variable Name	**Variable Description**	**Possible Answers**
eq_mobiliteit	Welke hokje past het best bij uw gezondheid VANDAAG: *Which category determines the best how you feel about your health TODAY:*	1. no problems with walking (geen problemen met lopen) 2. some problems with walking (beetje problemen met lopen) 3. moderate problems with walking (matihe problemen met lopen) 4. severe problems with walking (ernstige problemen met lopen) 5. not able to walk (niet in staat om te lopen) 9. unknown / missing
eq_zelfzorg	Welke hokje past het best bij uw gezondheid VANDAAG: *Which category determines the best how you feel about your health TODAY:*	1. no problems with washing and clothing myself (geen problemen mij wassen of aankleden) 2. some problems with washing and clothing myself (beetje problemen. . .) 3. moderate problems with washing and clothing myself (matige problemen . . .) 4. severe problems with wahing and clothing myself (ernstige problemen . . .) 5. not able to wash or cloth myself (niet in staat om . . .) 9. unknown / missing
eq_activiteiten	Welke hokje past het best bij uw gezondheid VANDAAG: *Which category determines the best how you feel about your health TODAY:*	1. no problems with daily activities (geen problemen met dagelijkse activiteiten) 2. some problems with daily activities (beetje problemen met dagelijkse activiteiten) 3. moderate problems with daily activities (matige problemen met dagelijkse activiteiten) 4. severe problems with daily activities (ernstige problemen met dagelijkse activiteiten) 5. not able to do daily activities (niet in staat om dagelijkse activiteiten te voeren) 9. unknown / missing
eq_pijn_ongemak	Welke hokje past het best bij uw gezondheid VANDAAG: *Which category determines the best how you feel about your health TODAY:*	1. no pain or discomfort (geen pijn of ongemak) 2. some pain or discomfort (een beetje pijn of ongemak) 3. moderate pain or discomfort (matige pijn of ongemak) 4. severe pain or discomfort (ernstige pijn of ongemak) 5. extreme pain or discomfort (extreme pijn of ongemak) 9. unknown / missing
eq_depressie	Welke hokje past het best bij uw gezondheid VANDAAG: *Which category determines the best how you feel about your health TODAY:*	1. no anxiety or depression (geen angst of depressie) 2. some anxiety or depression (een beetje angst of depressie) 3. moderate anxiety or depression (matige angst of depressie) 4. severe anxiety or depression (enstige angst of depressie) 5. extreme anxiety or depression (extreme angst of depressie) 9. unknown / missing
eq_thermo	*On a scale from 0 to 100, which number determines the best how you feel about your health TODAY:*	number 1 - 100

Table 19: Overview of the quality of life questions. Answers are categorical and form part of the numerical data used to classify the CAD severity levels

Heart Quality of Life Questions		
Variable Name	**Variable Description**	**Possible Answers (same for all questions)**
hqol_binnen	Did the patient experience difficulty, in the past 4 weeks, with walking inside the house (on the same floor), due to their heart complaints?	0. a lot of difficulty 1. much difficulty 2. a little difficulty 3. no difficulty 9. unknown / missing
hqol_actief	Did the patient experience difficulties, in the past 4 weeks, with gardening, vacuuming, carrying groceries, due to their heart complaints?	1-3, 9
hqol_heuvel	Did the patient experience difficulties, in the past 4 weeks, while walking up a hill or climbing up one staircase, due to their heart complaints?	1-3, 9
hqol_lopen	Did the patient experience difficulties, in the past 4 weeks, while walking 100 meters in a firm pass, due to their heart complaints?	1-3, 9
hqol_bewegen	Did the patient experience difficulties, in the past 4 weeks, while exercising or physical activities, due to their heart complaints?	1-3, 9
hqol_tillen	Did the patient experience difficulties, in the past 4 weeks, with moving or lifting heavy objects, due to their heart complaints?	1-3, 9
hqol_kortademig	Did the patient experience shortness of breath, in the past 4 weeks, due to their heart complaints?	1-3, 9
hqol_lichbep	Did the patient experience physical obstructed, in the past 4 weeks, due to their heart complaints?	1-3, 9
hqol_moeheid	Did the patient experience a lack of energy or tiredness, in the past 4 weeks, due to their heart complaints?	1-3, 9
hqol_stress	Did the patient experience feelings of stress or restlessness, in the past 4 weeks, due to their heart complaints?	1-3, 9
hqol_depressief	Did the patient experience feelings of depression, in the past 4 weeks, due to their heart complaints?	1-3, 9
hqol_frustratie	Did the patient experience feelings of frustration, in the past 4 weeks, due to their heart complaints?	1-3, 9
hqol_bezorgd	Did the patient experience feeling of worry, in the past 4 weeks, due to their heart complaints?	1-3, 9
hqol_buiten	Did the patient experience difficulties, in the past 4 weeks, while working in their house or garden, due to their heart complaints?	1-3, 9

Table 20: Overview of the heart quality of life questions. Answers are categorical and form part of the numerical data used to classify the CAD severity levels