Diagnostic prediction on anamnesis in digital primary health care

MARIE KINDBLOM
Diagnostic prediction on anamnesis in digital primary health care

MARIE KINDBLOM
Abstract

Primary health care is facing extensive changes due to digitalization, while the field of application for machine learning is expanding. The merging of these two fields could result in a range of outcomes, one of them being an improved and more rigorous adoption of clinical decision support systems. Clinical decision support systems have been around for a long time but are still not fully adopted in primary health care due to insufficient performance and interpretation. Clinical decision support systems have a range of supportive functions to assist the clinician during decision making, where one of the most researched topics is diagnostic support.

This thesis investigates how the use of self-described anamnesis in the form of free text and multiple-choice questions performs in prediction of diagnostic outcome. The chosen approach is to compare text to different subsets of multiple-choice questions for diagnostic prediction on a range of classification methods.

The results indicate that text data holds a substantial amount of information, and that the multiple-choice questions used in this study are of varying quality, yet suboptimal compared to text data. The overall tendency is that Support Vector Machines perform well on text classification and that Random Forests and Naive Bayes have equal performance to Support Vector Machines on multiple-choice questions.
Primärvården förväntas genomgå en utbredt digitalisering under de kommande åren, samtidigt som maskininlärning får utökade tillämpningsområden. Sammanslagningen av dessa två fält möjliggör en mängd förbättrade tekniker, varav en vore ett förbättrat och mer rigoröst anammande av kliniska beslutsstödssystem. Det har länge funnits varianter av kliniska beslutstödssystem, men de har ännu inte lyckats blivit fullständigt inkorporerade i primärvården, framför allt på grund av bristfällig prestanda och förmåga till tolkning. Kliniskt beslutsstöd erbjuder en mängd funktioner för läkare vid beslutsfattning, där ett av de mest uppmärksamte fälten inom forskningen är support vid diagnosticering.

Denna uppsats ämnar att undersöka hur självbeskriven anamnes i form av fritext och flervalsfrågor presterar för förutsägning av diagnos. Det valda tillvägagångssättet har varit att jämföra text med olika delmängder av flervalsfrågor med hjälp av en mängd metoder för klassificering.

## Contents

1 **Introduction** 1
   1.1 Aim, scope & objectives 3
   1.2 Thesis outline 3

2 **Background** 5
   2.1 Clinical decision support systems 5
      2.1.1 Knowledge-based CDSS 6
      2.1.2 Non-knowledge based CDSS 7
   2.2 Motivation behind CDSSs 8
      2.2.1 Human reasoning and error 8
      2.2.2 Benefits and motivation 9
      2.2.3 Acceptance of CDSS in health care 10
   2.3 Theoretical background 10
      2.3.1 ICD code structure 10
      2.3.2 Methods for classification 11
   2.4 Related work 15
      2.4.1 Early approaches: briefing 1950-2000 15
      2.4.2 Drawbacks with early approaches 17
      2.4.3 Current research: state-of-the-art 17
      2.4.4 Utilizing free text data for diagnosis 19
      2.4.5 Conventions in medical research 20
      2.4.6 Summary 20

3 **Methods** 21
   3.1 Dataset 21
      3.1.1 General structure of dataset 21
   3.2 Approach 22
      3.2.1 Single all-forms classification approach 22
      3.2.2 Multiple form-specific classification approach 22
# CONTENTS

3.2.3 Feature groups .............................................. 23
3.3 Preprocessing ................................................. 23
   3.3.1 General preprocessing .................................. 23
   3.3.2 Multiple-choice feature preprocessing .............. 25
   3.3.3 Text preprocessing ..................................... 25
3.4 Models and parameters ..................................... 25
   3.4.1 Testing and training ................................... 26
   3.4.2 Random parameter search ............................ 26
3.5 Validation and evaluation .................................. 28

4 Results ......................................................... 29
   4.1 Data exploration ........................................ 29
      4.1.1 Feature distribution ................................ 29
      4.1.2 ICD code distribution ............................... 31
   4.2 Single all-form classification approach .............. 33
      4.2.1 F1 score & accuracy ................................ 33
   4.3 Form-specific classification approach ............... 35
      4.3.1 F1 score & accuracy ................................ 35
      4.3.2 Collective statistics ............................... 41

5 Discussion ................................................... 43
   5.1 Single all-form classification vs. multiple form-specific classification .................. 43
   5.2 Main findings in form-specific diagnosis ........... 43
   5.3 Text data and multiple-choice questions .......... 44
   5.4 Model comparison ....................................... 45
   5.5 Possible use ............................................. 46
   5.6 Limitations .............................................. 46
   5.7 Ethics & sustainability ................................ 47

6 Conclusions .................................................. 50
   6.1 Future work ............................................. 50

Bibliography .................................................. 52

A Results from form-specific classification ............. 56
Chapter 1

Introduction

Clinical decision support systems (CDSS) are technological systems in the health sector that provide physicians with support during decision-making tasks. CDSSs have been around for decades, yet there is still difficulties present in their incorporation in primary health care. Despite considerable progress in CDSS research, the performance, interpretability and design of such systems in real-world applications is still inadequate [1].

Health care digitalization

Health care is currently undergoing heavy digitalization, which has the potential to reform several fields, one of them being decision support. Digital health care is assumed to contribute on several ends, not only by an increased medical quality but also in terms of accessibility and improved work environments. Furthermore, it allows for the development of equipment that can lead to an increased interest in personal health care [2]. Several actors have emerged in the public sector to enable access to health care via digital means. KRY is a digital health care provider among them, which currently connects patients and doctors through digital video meetings. The patients fill out symptom forms before the meeting as a self-evaluation of their health status, that is part of the foundation in the diagnostic evaluation by the physician. Self-evaluation forms are oftentimes similarly utilized in primary health care for patients to fill out prior to meeting a clinician. The information obtained from these forms could serve as a basis for decision support, yet the usage of this data for diagnostic prediction is rarely encountered.
Clinical decision support in research

The digitalization of health care enables an extended use of digital tools for the benefit of both the clinician and patient. A solid technical ground could allow for a more successful incorporation of CDSSs in the daily workflow of the clinician. Research within the CDSS field has primarily been concerned with the performance of classifiers, for natural reasons. While a few CDSSs have had moderate commercial success, they have failed wide-spread adoption and acceptance in primary health care [1]. As the incorporation of CDSSs in primary health care field presents clear benefits, the importance of continued research is indisputable. However, research for clinical decision support tends to be very limited in its scope. Prediction tasks are often carried out on narrow diagnostic domains, which implies an outcome that is limited to a certain medical diagnostic field. The nature of the classification is often binary classification for a given disease or a specific problem domain and, as a result, it rarely acquaints itself with a larger scope of multiple diseases. Moreover, the data used in research is often limited to e.g. biometric data, such as blood count, sedimentation rate, etc. While it is likely that this type of data is credible, a large set of diagnoses can be deduced without this type of data. Several diseases may be inferred solely with the information from self-assessment, e.g. cough, fever or skin conditions. The application of this data has, however, to a large extent been inadequately utilized in the health care sector.

Novel aspects

The aim of the thesis is to investigate how well various sources of self-assessed information can be used for diagnostic prediction in a digital health care setting. The information sources that are considered are multiple-choice questions (or their subsets) and text data. The research reported here is concerned with several novel aspects in relation to classification. The data used in this report stems from a large diagnostic ancestry, and as previously mentioned, contemporary research is often concerned with a fine set of diagnostic outcomes and therefore rarely evaluates a wider spectrum of diagnoses [1]. Moreover, the data used in research is often biometric data and rarely includes the element of self-assessment which this data solely depends upon. It is also possible, due to the different range of outcomes and atypical
source of data, that this study deals with diagnostic domains that are usually not considered for prediction.

1.1 Aim, scope & objectives

The data used in this study is a mixture of several sources of self-assessed information. The primary objective is to investigate how the various sources of information compare for prediction of diagnoses. The data provided by the user varies between multiple choice questions and text data. The structure of the data is heavily reliant on the initial symptom form that is chosen by the user, which is similar to a categorization of symptom domains. As a secondary objective, two classification approaches will be compared to study the importance of the form selection. This entails a comparison of diagnostic performance on a form-based level (multiple datasets) and with all the forms collected into a single dataset.

The three information sources (text, all multiple-choice questions, a selected subset of multiple-choice questions) are evaluated using various classification approaches to investigate both the importance of information sources (feature sets) and how it depends on the classifier choice. The employed models for classification are Multi-Layer Perception (MLP), Random Forest (RF), Naive Bayes (NB), k-Nearest Neighbors (kNN) and Support Vector Machines (SVM).

One limitation in this study is that only one data provider (KRY) will be considered, which makes the conclusions reliant on their specific data. In relation to this, the data is confidential and may not be disclosed. The data has, however, been visualized in several ways for explanatory purposes.

1.2 Thesis outline

The report begins with the Introduction, which is followed by the Background, where the various concepts related to CDSSs are introduced and the motivation and benefits with CDSSs are highlighted. Following this is on the topic of acceptance of CDSSs in health care, which discusses the shortcomings of integration. The theoretical aspects of decision support are then mentioned, where the relevant models for this thesis are presented. The section is finished off with an
overview of the related work in relation to decision support in health care and research. Methods, gives a brief overview of the dataset, the chosen approach, pre-processing, models and parameters as well as the methods used for validation and evaluation. In Results, the main results are presented and inferred for a selection of forms. This section is followed by the Discussion where the major insights are presented and problematized in contrast to recent work. Finally, Conclusions, summarizes the conclusions that could be drawn from the results and presents suggestions for future work.
Chapter 2

Background

2.1 Clinical decision support systems

CDSSs are a part of the wider domain of health informatics that combines resources from several disciplines to innovate health care services with the use of IT [3]. A CDSS is defined as a computer system with the aim to influence decision making made by clinicians at the point in time when a decision is made. CDSSs span over a large area of decision support, ranging from the suggestion of questions to ranking of diagnoses. There are several ways of categorizing CDSSs, and usual dividers are the point in time in which the system assists the clinician (e.g. pre-diagnosis, during diagnosis or post-diagnosis), their level of activeness and passiveness, their accessibility and whether they are knowledge- or non-knowledge based [1]. As the research and adoption of decision support have primarily been concerned with the diagnostic accuracy, the most fitting categorization to explore further is the division of knowledge and non-knowledge based systems.

Diagnostic Decision Support Systems (DDSS), sometimes referred to as Medical Decision Support Systems (MDSS), have emerged as a specialized kind of support systems to specifically tackle diagnostic support. A DDSS can be defined as a computer-based algorithm that assists a clinician with one or more component steps of the diagnostic process [1]. As the CDSS incorporates the diagnostic functionality of a DDSS/MDSS, the terms are regularly interchanged and CDSS is often loosely used to address the intended works of the more specialized terms that concern diagnostic assistance.
2.1.1 Knowledge-based CDSS

Early research was centered around simulating human thinking while practical uses often focused on specific diagnostic support. While the researchers had different aims, many of the adopted CDSS were focused on assisting the clinician in the decision making rather than providing an answer. The knowledge-based support systems have traditionally been thought of a system divided into three parts, namely the knowledge base, the inference/reasoning engine and the interface for the end-user. Figure 2.1 illustrates the division between the reasoning engine and knowledge base, where the output is fed to the interface.

The reasoning engine is at the heart of the system, where boolean logic and probability have historically played a central role in the machinery. The addition of informal logic is, however, not uncommon for these types of systems. The usual work-flow is to combine the input (which usually is in the form of controlled vocabulary with specific components, such as age, gender, symptom, duration and findings) and additional data to a logical scheme that produces an output. One frequently used scheme is the Bayesian network that relies on probability, which has been especially attractive in the sense of representing the reasoning engine well [1].

The knowledge base should match the inference engine design to the best extent. In the example of a Bayesian network one understands that the prior, conditional and posterior probabilities are based on the knowledge base, as seen in Figure 2.2. The knowledge base is often populated by the use of literature, which possesses an inevitable weakness due to the variety of parameters (as diseases vary within populations, time of year, etc). Logical systems that use discriminating
questions to solve mutually exclusive cases are often applied to narrow problem domains where the patient almost certainly only suffers from one disorder. If-then rules often appear in these systems, and the limitations become more evident when patients display multiple independent diseases as the code base turns immense [1].

Early systems were primarily knowledge-based and mainly Bayesian. The systems were later extended and variations appeared that employed fuzzy set theory, rule-based and heuristic systematicity. The output of knowledge-based systems have, however, remained to generally be in the form of a ranked list of probabilities of diseases. Regarding the interface, CDSSs have in various degrees been incorporated into electronic medical records (EMR) systems [1][4].

2.1.2 Non-knowledge based CDSS

Machine learning techniques have opened up the possibility of non-knowledge-based CDSSs, where the systems are more flexible and can learn from examples. The nature of the inference engine and knowledge base are in some sense static compared to a non-knowledge based CDSS. Neural networks have been the most recognized model for CDSS systems, and are regularly used in a vast group of areas in medicine along with highly orthogonal fields and industries. The system is obviously bound to the quality of its data, and users should therefore not have high expectations. The goal might not always be to deliver the correct diagnosis, but to discriminate between causes or prognoses and function post-diagnosis. The loss of the knowledge base which is
heavily used for inference in the knowledge-based CDSS is the obvious trade-off when selecting a non-knowledge based system [1].

Three popular methods for non-knowledge based systems are the previously mentioned Artificial Neural Networks (ANN), along with Bayesian Networks (BN) and Support Vector Machines (SVM). It is worth noting that non-knowledge based CDSS more often than not tend to have a more narrow scope than knowledge-based CDSSs. Knowledge-based systems often cover diagnosis of multiple disease areas whereas non-knowledge base typically focus on a particular disease and thus a limited subset of the possible symptom scope [1][4].

2.2 Motivation behind CDSSs

2.2.1 Human reasoning and error

A working theory regarding human diagnostic reasoning is that it can be categorized in three strategies, namely probabilistic, causal and deterministic. Observational and experimental studies have shown that humans are not intuitively good statisticians, and human problem solving tends to rely heavily on judgmental heuristics. Humans often evaluate subjective probabilities improperly, incorrectly use prior probabilities and often overlook important information. Humans also have difficulty when it comes to reasoning with probabilities in general and also display reasoning errors such as failing to revise new evidence leading to multiple new options that could conflict with one’s working hypothesis. When reviewing causal reasoning, studies have proposed that humans tend to assign greater impact to causality rather than diagnostic data that are of equal informative weight. Humans also tend to display over-confidence in their predictions when interacting with models that display high uncertainty. It is, however, difficult to estimate the extent to which causal reasoning that accounts for the early stages of hypothesis formulation. Causal reasoning has however been shown to be primarily used when faced with problems that lie outside of one’s domain of expertise, atypical problems and when asked to explain one’s reasoning to others [1][4].
2.2.2 Benefits and motivation

While it is evident that the diagnostic process is exposed to the inevitable fate of human error, medical errors due to diagnostic mistakes are considered to be quite underestimated. A cognitive psychologist named Arthur Elstein specialized in clinical decision making assessed that clinicians on average reach an incorrect diagnosis 10-15% of the time. Another study has estimated that 30% of morbidity and mortality in health care is due to medical error [5][6][7].

Previous studies support the idea that clinicians often arrive at a preliminary diagnosis that can be explained with initial key findings, subsequently leading to an interrupted thinking process – which at times can lead to premature diagnoses. Furthermore, diagnostic error may lead to delayed diagnostic closure, which risks putting patients at harm. Decision support systems enable clinicians to explore related possible diagnoses that can challenge early assumptions made by the clinician and also guide them to appropriate questions due to the spectrum of alternative diagnoses [8]. Many outcomes are critically dependent on the accurate and early detection of the diagnosis, and some research has actually shown that the diagnostic performance of physicians is directly linked to the information given by the MDDS [9]. Clinical decision support systems have also been shown to improve both patient outcomes as well as the cost of care [10]. While it is reasonable to believe that such systems might contribute to redundant questioning and hypothesis switching for the clinician, the gain from reaching a correct diagnosis has a substantive and predominant worth [8].

Health care is predicted to face rapid development and go through heavy digitalization in the foreseeable future [2]. Modelling techniques in CDSSs have been applied to a range of decision processes in medicine (prognosis, therapy planning, diagnosis etc), yet diagnosis remains the most researched for decision making. CDSS and MDDS systems are core areas for artificial intelligence in medicine that are central for research and digitalization. Medical diagnostic decision support systems allow for an increased diagnostic decision accuracy and studies have found that physicians’ diagnostic performance is strongly influenced by the quality of information produced by a diagnostic decision support system. The cost associated with predicting an incorrect diagnosis has pushed the research field and model search forward, now
including multiple methods e.g. ANN, SVM, BN, kNN, decision trees.

2.2.3 Acceptance of CDSS in health care

While a few MDSSs have been commercialized (HELP, DXPLAIN, QMR, ILIAD, etc) the overall application have gained neither widespread acceptance nor usage in health care despite half a century of research. This is thought to be due to a couple of reasons. Issues related to modelling have been concerned with accuracy and pre-processing. Early systems have not displayed high enough accuracy, and only few systems have had diagnostic performance comparable to experts. Natural Language Processing (NLP) methods have historically not been sophisticated enough to handle free text data (record and patient data is usually in this form), which has been a prerequisite for analysis. Moreover, many applications have not been incorporated properly and as a stand-alone tool that only provides diagnostic support (lacking treatment plans, etc), thereby distracting the workflow of the clinician. Another aspect has been regarding the usability, where many systems have lacked user-friendliness and thus required extra allotted time, which is often infeasible in health care. When clinicians repeatedly encounter usability issues, limited use is a natural consequence. Finally, physicians have a higher probability of believing and using the MDSS’s recommendation when provided with an explanation. The lack of an explanatory feature has lead to lesser use. It is evident that although the accuracy has improved a lot over recent years, integrational and usability issues still remain.

2.3 Theoretical background

2.3.1 ICD code structure

ICD stands for the *International statistical classification of diseases and related health problems*, and is produced by the World Health Organization (WHO). The ICD comes in several revision, and the current version is the 10th revision, ICD-10. The purpose of ICD is to have a way for categorical classification of diseases, to enable further analysis, interpretation and comparison across countries. It has in practice become the international standard for diagnostic classification in health
care, and is thus widely used. The ICD code structure is hierarchal, on a top level roughly representing body systems (e.g. diseases of the nervous system or circulatory system or digestive system). This level is of three-character length, which can be further divided into ten different four-character length subcategories. [14] Considering one of the top categories, \textit{X Diseases of the respiratory system}, we see that it further divides into:

- J00-J06 Acute upper respiratory infections
  - J00 Acute nasopharyngitis [common cold]
  - J01 Acute sinusitis
    - J01.0 Acute maxillary sinusitis
    - J01.1 Acute frontal sinusitis
    - J01.2 Acute ethmoidal sinusitis
    - ...
  - ...
  - J05 Acute obstructive laryngitis [croup] and epiglottitis
  - J06 Acute upper respiratory infections of multiple and unspecified sites

- J90-J94 Other diseases of pleura

- J95-J99 Other diseases of the respiratory system

The most specific diagnosis is of four character length, and depending on the category there are specific conditions that qualify or do not qualify for a given diagnostic code, which is considered by the physician [14][15].

2.3.2 Methods for classification

A brief overview of the statistical methods that will be used for classification in this thesis is presented here. Due to the number of methods, only a light overview is provided and mathematical details are omitted for some methods.
Naive Bayes

The NB classifiers belong to the probabilistic classifiers that rely on Bayes’ theorem (given a class features are assumed to be independent). The prediction has a deterministic nature and is determined by choosing the most likely class as such:

\[ P(Y = C | x_1, ..., x_n) = P(C) \prod_{i} P(x_i | C) \]  

where \( C \) is the unknown class that we wish to predict and \( x_1, ..., x_n \) are the observed values for the features of an instance \( X \). \( Y \) is determined as the class \( C \) with the highest probability. Although the simple and slightly naive assumption of independence, NB performs surprisingly well in various classification applications and is one of the most effective learning algorithms used [16].

k-Nearest Neighbors

The kNN algorithm seeks to estimate the conditional distribution of \( Y \) given \( X \), by classifying samples to the class that is estimated to have the highest probability. The kNN algorithm does so by taking a parameter \( K \) and identifying the \( K \) nearest points from the training data (referred to as \( N \)) to a test sample \( X \). The conditional probability for a class \( C \) is the relative fraction of points in \( N \) that are labelled as class \( C \):

\[ P(Y = C | X) = \frac{1}{K} \sum_{i \in N} (y_i = C) \]  

The Bayes rule is then applied by classifying the test observation \( X \) according to the class \( C \) with the highest probability. For the kNN algorithm, the choice of \( K \) has a large impact on the classification performance [17].

Random Forests

Decision trees are popular due to their interpretability and high classification execution speed. They do however suffer from generalization loss on unseen data, i.e. they often overfit to the training data [18]. A Random Forest is an ensemble method where a combination of tree predictors are used for classification or regression. Each tree is dependent on an independent and random selection of features (the distribution
is the same for all trees in the forests) \cite{19}. We may then combine the tree predictions in the forest prediction by for example averaging. If we seek to determine the probability for class $C$ given an instance $X$ over $T$ trees we can estimate as follows

$$p(C|X) = \frac{1}{T} \sum_{t=1}^{T} p_t(C|X)$$  \hspace{1cm} (2.3)

Decision forests have a key component of randomly different trees, and one parameters to consider is the number of random trees used. Another common choice is to stop growing the trees after a depth $D$ \cite{20}.

**Multi-Layered Perceptron**

The MLP is a feedforward ANN that takes inspiration from the architecture of the brain. MLPs consists of at least one hidden layer, as well as the input and output layer, where the layers consists of a number of neurons. All neurons have an activation function that takes the weighted input and produces the output. Common activation functions are the identity, logistic, tanh and ReLu. Training is actually the learning of the weights, which is done by error backpropogation where the error in the final layer is compared to the expected results \cite{21}. Figure 2.3 is an illustration of a neural network with a hidden layer and the input and output layer.
Figure 2.3: Example network with one input layer ($x_1...x_4$, colored mustard), one hidden layer with three hidden nodes ($h_1...h_3$, colored green) and one output layer ($y_1...y_3$, colored blue). Links are accompanied by a specific weight, e.g. $w_{x_1h_1}$ is the weight for node $x_1$ to node $h_1$. Inputs are combined in the hidden nodes and passed through an activation function.

It is notable that several aspects are variable with neural networks, some of which are the architecture (layer and number of nodes), activation function and batch size during training [21].

Support Vector Machines

The SVM is a generalization of a simple and intuitive classifier known as the maximal margin classifier, where classification is done by separating data points by a hyperplane that has the maximal margin to both classes. This means a separating hyperplane that is as far as possible from the observations. There are instances where a hyperplane that separates the classes does not exist or when the boundary between two classes are non-linear, which is where the use of SVMs are essential. The SVM is an extension of the support vector classifier that results from enlarging the feature space in a specific way, using kernels. The way to address non-linearity in SVMs is by enlarging the feature space by adding features that are created as functions (quadratic or cubic, for example) from the original predictors. Due to this, the solution will be non-linear. To enlarge the feature space kernels are used, which depending on the function (e.g. polynomial kernel of a certain degree, radial kernel) attempts to fit a support vector classifier in a higher dimensional space, as opposed to the original feature space where the
case was not linearly separable \cite{17}. In the higher feature space, however, the decision boundary is in fact linear, resulting in a non-linear decision boundary in the original feature space, as seen in Figure \ref{fig:svm examples}. The choice of kernel is important in the SVM, where different parameters apply for different kernels, such as the degree in the polynomial kernel.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{svm_examples.png}
\caption{Example of two SVMs with polynomial kernel of degree 3 (a) and with a radial kernel (b). Both SVMs capture the decision boundary which could not been found linearly in the original space. Reprinted from \textit{An introduction to Statistical learning} (p. 353) by G. James et. al. (2013) Springer-Verlag New York. \cite{17}}
\end{figure}

\section{2.4 Related work}

\subsection{2.4.1 Early approaches: briefing 1950-2000}

Ledley and Lusted \cite{22} revealed in 1959 that physicians do not have clear knowledge of how they actually solve diagnostic problems. Today we know that several parts are involved in the decision making that physicians struggle with, due to the nature of human error. Ledley and Lusted introduced the concepts thought to be central components to medical reasoning, which was logic (boolean algebra and set theory) and probabilistic reasoning (Bayes rule). This was the basis of theory that subsequent years would build upon. The first practical system based on Bayesian logic was developed in 1968 and used for
diagnosis of acute abdominal pain [1]. Probabilistic Bayesian systems have since then continued to play an important role in diagnostic support systems. In the late 50s, additional systems employing alternative strategies emerged, such as the use of heuristics with symbolic reasoning [1]. The trend onwards was the use of mathematical models, where heuristics mainly took place in the 70s and 80s, often due to lack of data. To combat heuristics and simple Bayesian approaches, fuzzy set theory, rule-based systems and Bayesian belief networks were continuously advanced. Since then, artificial neural networks have become the rising star, being the most popular model choice in DDSS implementations [1][4]. The trend in research is well illustrated in Figure 2.5, which is the aggregated result when querying “Computer assisted Diagnosis” on Pubmed and omitting image-, waveform and -omic data (as those domains are hard to apply to the diagnostic research area). Note that the trend in research articles by no means reflect the in-use systems in health care.

Figure 2.5: Trend in over the last decade in research related to Computer aided Diagnosis. Adapted from Modeling Paradigms for Medical Diagnostic Decision Support: A Survey and Future Directions, by K. B. Wahogilkar et al. (2012) [4] Abbreviations: BN = Bayesian Networks, NB = Naive Bayes, FST = Fuzzy Set Theory, SVM = Support Vector Machines, ANN = Artificial Neural Networks.
2.4.2 Drawbacks with early approaches

The early models presented difficulties in various ways. While Bayesian systems have remained the most widely implemented in hospitals, they were considered too simplistic due to the assumptions of symptom independence and mutual exclusive diseases which forced researchers to investigate alternative approaches. Rule-based systems possess the crude quality of growing unmanageable as the number of rules grow. Heuristic systems were often considered less glamorous as they lack a formal proof of correctness. The vague aspect of heuristic systems yet again pushed research in search of something more optimal. Similarly, fuzzy models were quickly dismissed due to their high computational cost. Models that have become increasingly popular like ANNs and BNs still face problems. ANNs have shown promising results but are perceived as “black boxes” by clinicians as they do not provide explanatory reasoning, which is crucial to assure usage by clinicians (see Acceptance of CDSS in health care). Bayesian networks have been problematic to implement for some diseases and often grow very complex [4][23].

2.4.3 Current research: state-of-the-art

The gap between research and CDSSs in health care has been evident for a long time. Although computers were early prime tools in medicine decades ago, research operated on a stand-alone basis for many years. However, the technological advances have over the last two decades transformed many fields, and the medical field has not been an exception to the rule. Many visionaries in the research field have intended to improve and revolutionize medicine by pushing the technologies used forward. Due to this engagement and with the adaptive vision by developers, many systems now use neural networks, genetic algorithms, fuzzy logic, etc. for reasoning purposes [24][25]. However, due to the nature of medical data (imprecise measurements, varying baselines for results, multiple conditions) the decision-making process is often uncertain. This leads to a less straightforward way of determining the most appropriate model, and there is generally not a “one fits all”-solution in machine learning. With this said, medical diagnostic support tools are considered to be reasonably established in medicine and medical technology today, with varying degree of quality, usability and accuracy [23][24][25]. However, the develop-
mentation of diagnostic tools remains to be of great importance, especially
to cater to diseases like cancer, diabetes and cardiovascular diseases,
which are among the most serious and diverse diseases [23]. It is also
important to stress that the status of current research does not by any
means reflect the level of incorporation of these systems in hospitals.

ANNs have many applications in the medical field, and have been
applied to diagnosis to several conditions like appendicitis, back pain,
dementia, myocardial infarction, psychiatric emergencies, sexually trans-
mitted diseases, skin disorders, various types of cancer (oral, colorec-
tal and colon to mention a few), early diabetes and temporal arteritis
[1][23]. As of year 2000 over 500 papers were published yearly in med-
cal journals featuring neural networks with applications in medicine
[10]. Studies have also shown that neural networks may for some con-
ditions outperform the predictions made by physicians [1][23][26]. A
study on the diagnostic prediction on ANN applied for pulmonary
embolisms were as good as or better than the predictions made by
physicians. Another study implemented a neural network for detec-
tion of acute myocardial infarction that managed to outperform two
cardiologists with great experience [1].

Decision trees, random forests and SVMs have in various constel-
lations also proven to be efficient for diagnosis and have in some cases
outperformed other methods [27][28][29]. Decision trees have been ap-
plied to many diagnostic areas, such as cardiology and oncology. De-
cision trees have outperformed Bayesian classification and had higher
but similar accuracy to neural networks and kNN in a study for heart
disease prediction. An extensive study was conducted in 2014 where
classifiers from 17 families were tested on 121 datasets, which featured
many common classifiers (random forests, SVMs, ANNs, boosting,
bagging, kNN, Bayesian, etc.). The datasets were from various do-
mains, but included several datasets related to medical diagnosis, like
breast cancer, inflammation, etc. In this study random forests man-
ged to outperform all other classifiers and was therefore considered
the classifier that was most likely to be the best model pick. Random
forests and decision trees also have a clear advantage over most meth-
ods for their interpretability, which has been an issue for clinicians
historically. SVMs also serve as excellent classifiers and have shown
high accuracy in several cases. In the previously mentioned study on
121 datasets, an SVM with a Gaussian kernel performed second best
where the difference from random forest was not statistically signifi-
cant. It is therefore reasonable to argue that SVMs also serve as safe bets in modelling choices. Overall, state-of-the-art machine learning is often considered to be well suited for medical data analysis and intense work is focused on research within small specialized domains of diagnostic problems [30][31][32][28].

2.4.4 Utilizing free text data for diagnosis

The majority of the presented research have utilized numeric, biometric data. However, it has been shown that a vast amount of information relating to diagnosis can be found in free text, and that free text can be utilized for accurate dating of diagnosis and for identifying misclassified diagnosis [33]. An anamnesis refers to a patient’s recollection of their medical history. Using CDSSs on patients’ anamneses have not been extensively researched. There are, however, instances where the physicians text notes have been subject to classification.

A study conducted in 2013 successfully used Electronic Health Records (EHR) progress notes to identify a relation to diabetes [34]. They found that a bag-of-words SVM-model was an effective approach to filtering out notes related to diabetes patients. One suggestion for extension was to apply the method to other diseases. Another recent study successfully attempted to utilize free text from intensive care unit (ICU) clinical notes to classify procedures and diagnoses for risk adjustment [35]. They used n-gram feature extraction with an SVM classifier. Their choice of model was due to scalability needs. One of the mentioned limitations in this study was that they only considered two procedures and two diagnoses, which lowers the possibility of generalization.

A thorough literature review by D. Demner-Fushman et al (2009) did, however, shed light on the fact that the NLP-approaches are not necessarily concerned with diagnosis but other relevant aspects in health care, such as monitoring, detecting clinical events, extracting findings, etc [36]. This indicates that free text can be utilized in several ways. Their strong belief is that text analysis tools, as NLP, can be of use for all involved actors in clinical processes. While it is apparent that a lot of research is conducted on free text data produced by the clinician, a similar focus on the usage of patients’ self-assessment has not been identified. It is also clear that there is a larger field of application for NLP-approaches, beyond diagnostic motives. Furthermore, several studies have a narrow scope of only a handful of diagnoses,
which is often mentioned as a limitation.

### 2.4.5 Conventions in medical research

While there are certain models that have been implemented with great success, the complex nature of the diagnostic task rules out the possibility of an intelligent model selection. The conventional strategy is to select a model that performs well on the given data and when possible, benchmark it with cross-validation [25].

Because MDSS applications often are driven by commercial objectives, the code, knowledge bases and medical datasets are rarely open-source. The withholding of tooling and lack of large, openly available medical databases make comparisons, evaluation, benchmarking and reuse infeasible. Related fields as bioinformatics have adopted a data-sharing praxis and, as a result, many genetic datasets, tools and libraries are open-source nowadays. Open-source allows for continuous improvement and qualitative additions from a far greater group of contributors than a single research group. The release of datasets and programs is therefore vital [4]. The project will as a result be limited by this fact, as comparisons are infeasible.

### 2.4.6 Summary

Several models have been applied in diagnostic contexts with success. ANNs have been heavily adopted, followed by SVMs, RFs, NB and kNN that have performed equally well in various contexts. This gives us reason believe that this collection of models ought to perform well on diagnostic classification. The approach in this study is clearly non-knowledge based, which would make possible fields of applications those where clinicians use non-knowledge based CDSS. However, as the spectrum of diagnoses is extremely wide-spread, the approach is the complete opposite of traditional non-knowledge based CDSS. This study therefore presents a novel aspect from previous work by studying an extended diagnostic domain. The related work also indicates that the use of available data (patient record data) is usually in free text format, which have not been extensively used for diagnostic prediction. Instead, the focus has predominantly been to wield biometric data. This research aims to utilize free text data, which provides an important differentiation and novelty from previous work.
Chapter 3

Methods

3.1 Dataset

The dataset used in this research has been provided by KRY (Webbhälsa AB), and is data related to meetings conducted between the period of 20th September 2017 to 15th March 2018. A meeting is initiated by the patient in the KRY application, where the patient is requested to fill in information related to the meeting before it takes place. The patient is presented a set of symptom forms and selects the most fitting one, and is then guided through a questionnaire based on the selected symptom form. This submitted form constitutes part of the material that the physician considers before the meeting takes place. The meeting is then carried out over video call, where the physician forms their opinion of the condition and reports the diagnosis and possibly prescribes medication for the patient. The dataset used in this study consists of the answers to the questions in the questionnaires filled in prior to the meeting by the patients, paired with the diagnosis given by the physician. The diagnoses (labels/output) are thus a representation of the narrative composed by the patient (input), which will be used to train and test the classifiers.

3.1.1 General structure of dataset

A sample in the dataset is considered a form filled in by a patient that consists of a set of questions, which is determined by the symptom form chosen by the user. A few questions appear in all forms, such as age and gender. The forms vary in number of questions and the type...
of questions asked. Questions vary between multiple-choice questions and user input such as numbers, text and in some cases photos. As one might realize, some questions do not apply to all types of users, which is often determined by the form, age and gender. As an example pregnancy only applies to women and only children are asked about their weight. Some forms also include a longer, open-ended question formulated as "Why are you here?" or "Describe your issues", which is a text input field that the patient is free to fill in however they want.

The forms are rather specific and chosen by the user. The assumption is that the user will choose the most suitable form, as the outcome is very dependent on the questions asked. The considered forms are urinary tract infection, sleeping problems, eczema, fever, general health questions, eye infection, headache, cold sore, allergies, insect bite, asthma, constipation, ingrown nail and wind.

3.2 Approach

The classification task will be evaluated from two different perspectives, from here on referred to as the single all-forms classification approach and the multiple form-specific classification approach. To measure the importance of the different groups of features, there are three various types of data that is considered, namely all features, selected features and text data.

3.2.1 Single all-forms classification approach

In the single all-forms classifier approach, only a single classifier will be trained and tested. This implies that the feature space will be equal to all unique questions in all symptom forms, as all forms are grouped together. This approach opens up for cross-form predictions, as data from all the different forms is considered. This approach also means that the classifier will have access to the whole dataset, which might prove beneficial during training.

3.2.2 Multiple form-specific classification approach

In the multiple form-specific classifiers approach, classification is done individually on each form. As a result, each form will have a designated classifier that has been trained and tested only with form-specific data.
points. Due to the variations between forms, the feature space will vary with each form. This investigation is interesting as the form-based approach allows for smaller feature spaces yet fewer data points for training and testing.

### 3.2.3 Feature groups

The **feature groups** are the various types of data sources that have been considered in the dataset. The group **all features** refers to the group where all multiple-choice questions have been considered. Note that this corresponds to all multiple-choice questions in all forms for the *single all-forms classifier* approach, and all form-specific multiple-choice questions in the *multiple form-specific classifiers* approach. The second group, **selected features**, are the selected multiple-choice features that were selected based on the findings during the data exploration on the feature distribution (Section 4.1.1). The third group is the **text data**, which is the text answer filled in by the patient to the open-ended question that seeks to answer the general reason behind the visit. This data differs heavily from the multiple-choice data, and therefore examines how the groupings of multiple-choice questions compare to text data. The text question is, however, not included in all forms. As a result, this specific comparison between multiple-choice and text data is evaluated where it is applicable. The specifics of the data is furthered described in the next section, Section 3.3.

### 3.3 Preprocessing

#### 3.3.1 General preprocessing

A number of pre-processing steps have been taken in order to obtain a sound dataset in terms of integrity and compatibility:

- Personal data such as forenames, surnames and birth data have been removed in order to preserve integrity.

- Multiple-choice and numerical questions are kept which is believed to be the most credible approach. These types of questions constitute a vast majority content-wise for all the forms. A few questions that are text-based have been removed due to time limitations, an example is a text-form to explain duration of time.
This can be interpreted in too many ways for it to be reasonable to parse (as it may be in the format of minutes, hours, months, years, etc).

- Many forms include an open and general text-question that aims to comprise a more over-all health status of the patient or reason behind the visit, and is usually composed as "Describe your problem" or "How are you feeling?". This is the only free text data that has been kept as it usually contains a great deal of information, is one of the earliest questions asked and also gives the user the most open-ended chance of communication.

- The data points are accompanied with several statistics such as a tag for whether a patient was helped or not. A negative instance can be due to a number of reasons, such as technical error or that the patient’s need was not compatible with the services that KRY provides (such as physical examination). All data points where a patient was not helped have been removed (as they also do not include an ICD code).

- Other instances that cause a data point to seem unreliable are when there is no ICD code, there is no start- or end-time for the meeting or when certain doctor IDs come from testing during app development. All of these instances have been removed.

- Since ICD codes can be of varying length (depending of specificity), the decision was made to trim all ICD codes to length four by removing the last letter/number. As the ICD code structure is hierarchal, this is not expected to have any unforeseen consequences. ICD codes of shorter length have been discarded.

- As a way to combat skewed ICD-frequencies in the dataset, the dataset has been subject to cutoffs under circumstances when there are too few instances of rare ICD codes. The numerical choice is obviously arbitrary in some sense, yet motivated by Section 4.1.1, ICD code distribution, that illustrates how the dataset is clearly dominated by a subset of the ICD codes. The chosen approach has been to keep ICD codes that have at least 100 occurrences in a given dataset as separate classes, and group diagnoses with less frequency into a group called "X". Note that when predicting on all forms, we may have more ICD codes in total as...
they can appear across forms. However, they might not necessarily appear in any of the form-based predictions where the same rule applies (where the dataset is comprised of data belonging to that specific form only).

After the general filtering has been applied the remaining number of points are 85304, where only 47483 points have text data. To do a fair comparison between models, all classifiers have been trained with the lower number of points when a form contains text, and all points when a form does not have any text data. Note that the number points per form is significantly smaller than the two numbers presented, as they are all points across forms. For the single all-forms classifier approach only points with text have been considered.

3.3.2 Multiple-choice feature preprocessing

The approach has been to treat unique questions as features for the classification. In some rare cases questions between forms have showed enough resemblance to be treated as the same feature. Note that this is only interesting for the single all-forms classifier approach. Multiple-choice questions have been transformed to numerical values and then scaled within a range for each feature using MinMaxScaler in ScikitLearn. The estimator individually scales each feature such that it is between 0 and 1. The data is later split into train and test set, meaning that the scaling has been done while the dataset still is intact.

3.3.3 Text preprocessing

The text has been prepared by conversion to lowercase, removing delimiters (‘,’ ‘,’ ‘‘, etc) and then stemmed by a Swedish stemming library in Python. The text has then been converted to a token counts matrix with standard 1-grams where Swedish stop words have been removed, and extended with 2-grams and 3-grams. The matrix has then been transformed to a normalized tf-idf representation (using L2-norm) which is used for training the classifiers.

3.4 Models and parameters

The approach for this research will follow the conventional line of action where the data will be used to identify models and their suitable
parameters. Random search in the parameter space will be used in order to tune the final model choice based on performance. Notice that the "best" model might very well vary with the nature of the forms, due to their high variance in number of questions and type.

A highly adopted model for baseline purposes is NB, which has been used historically for classification and nowadays often appears as a trusted baseline. NB will similarly be used in this report for the same purposes. The other models that will be compared are SVMs, RF, kNN and MLP. All models have been considered for the feature groupings and SVMs, kNN and NB for the text data.

3.4.1 Testing and training

The division into train and test set has been done with ICD code frequencies in mind. Due to skewed ICD code frequencies, the training data has been subject to both over- and under-sampling. The data has been randomly sampled into test and training sets (25% and 75% respectively) on ICD code level. To achieve a uniform training, the average number of points across classes (ICD codes) has been calculated and then the training data for each class has been altered by under- or oversampling depending on its relative occurrence. However, when testing the classifier, the actual relative fraction has been left unaltered, which has been noted in the Results section to inform the actual number of test points for each class. This is believed to be a more representative approach for measuring, given the distribution of the dataset. The test and train sets have been re-sampled and classification has been done multiple times (10 times) and averaged. 10% of the training set was used for validation.

3.4.2 Random parameter search

As the content and quality between forms vary heavily, a random search has been performed for all forms and for all models with parameters (MLP, RF, kNN, SVM), for all feature groups (all features, selected features, text) to obtain the most just parameters (meaning parameters are form-, feature- and model specific). Due to the large feature space in the text classification setting, the choice was to perform a random search instead of a grid search.
Table 3.1: Table of the parameters used in the random search.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>batch sizes = [10, 50, 100]</td>
</tr>
<tr>
<td></td>
<td>activation = [identity, logistic, tanh, relu]</td>
</tr>
<tr>
<td></td>
<td>hidden layers = [(10-50, 1-2)]</td>
</tr>
<tr>
<td>Random Forest</td>
<td>nr. of estimators = [5, 10, 20, 25, 50]</td>
</tr>
<tr>
<td></td>
<td>max depth of tree = [None, 10, 20, 30]</td>
</tr>
<tr>
<td></td>
<td>min. samples for internal node split= [2, 6, 10]</td>
</tr>
<tr>
<td>kNN</td>
<td>nr. of neighbors = [5, 10, 15, 20]</td>
</tr>
<tr>
<td></td>
<td>weight function in prediction = [uniform, distance]</td>
</tr>
<tr>
<td>SVM</td>
<td>kernel = [linear, polynomial, rbf, sigmoid]</td>
</tr>
<tr>
<td></td>
<td>degree for polynomial kernel = [3, 4, 5]</td>
</tr>
<tr>
<td></td>
<td>gamma = [0.1, 0.05, 0.001, 0.5]</td>
</tr>
<tr>
<td></td>
<td>coef0 = [0.0, 0.001, 0.005, 0.01]</td>
</tr>
</tbody>
</table>

The randomized search has been run 10 times on randomly sampled test and training sets and averaged.

Criminisi et al (2011) reported that the forest size (number of estimators) and tree depth (max depth of tree) are among the most influential parameters for RFs [20]. For the kNN, the parameter of most importance is the number of neighbors, $K$ in 2.2 [17]. The kernel is of high importance when using an SVM, and polynomial and radial are commonly used as well [17]. Therefore, a range of kernels were of priority for the SVM. The linear was included to investigate whether a more complex kernel is necessary or if a linear kernel would suffice. Degree, gamma and coef0 are variable parameters for the kernels that are non-linear. The MLP contains a vast range of hyperparameters that could be explored and subject to extensive random search. However, due to the number of methods and the given time frame, the parameters that were chosen were those that were considered to be of most effect, which were the batch size, activation function, number of nodes and hidden layers (architecture). The additional parameters, min. samples for internal node split (RF) and weight function (kNN), were added to test their respective influence.
3.5 Validation and evaluation

F1-score will be used to evaluate performance of the models, which is a harmonic mean between recall and precision:

\[ F_1 = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} \]

where recall (sensitivity) is the fraction of times the correct diagnosis was predicted out of all the number instances for that diagnosis, precision is fraction of times that the correct diagnosis was predicted out of the times the diagnosis was predicted. Accuracy as the fraction of correctly predicted diagnoses will also be considered to indicate the overall performance in the forms:

\[ \text{Accuracy} = \frac{\# \text{correct predictions}}{\# \text{occurrences in test set}} \]

This is the same formula as the F1-scores recall, but this is instead considered for all diagnoses in a form. As the test set is skewed due to the imbalanced dataset, one should revise the confusion matrices that illustrates the overall distribution for the predictions and thereby all individual diagnoses. The F1-score is therefore of greater importance than accuracy. The F1-score will be considered both individually for all diagnoses as well as on average to investigate the classifiers ability to distinguish between diagnoses. To evaluate the statistical significance of the results, a 2-way ANOVA test will be performed on the results (F1-scores) to investigate both feature groups and models. Following this will be a post hoc multiple comparisons test (Tukey’s honest significant difference criterion) to investigate which groups have significant differences.
Chapter 4

Results

4.1 Data exploration

Initial work was done to gain better understanding of the data and to motivate the approach described in Methods. One relevant aspect to investigate was the distribution of features (questions) in the forms, which would motivate the single all-forms classifier and multiple form-specific classifiers approach. The feature distribution across forms is crucial in order to better understand the feature space and single-classifier, as well as to illustrate the variations between forms in terms of feature space and features.

Secondly, it was essential to investigate the distribution of ICD codes in the dataset, in order to better understand the diagnostic scope and the number of ICD codes that was relevant to include. Following this, the specific ICD code distribution within forms was also relevant, to investigate the variations between forms in terms of diagnostic span and specificity.

4.1.1 Feature distribution

In order to better understand the data and symptom forms, the features became a relevant aspect to dwell upon. As the number of shared features between the forms opens up for cross-form predictions, one would wish to see much overlap. At the same time, in order to attain specificity the forms should also be differentiable enough to target certain diagnostic domains, which requires form-specific questions. This is also important from the user’s perspective, as one wishes to opti-
mize the number of relevant questions for the user, which is a considerable trade-off.

As seen in Figure 4.1, only a few features are shared across all forms, while most are form-specific or shared between a small number of forms. Hence, there is little overlap between the forms, apart from the features age, gender, and symptom_form (corresponding labels are colored yellow in Figure 4.1). Apart from these protruding features, some features appear conditionally, e.g. earlier treatment/consultation, pregnancy and age. The distribution of features motivated the selection of features. In particular, age, gender & symptom_form appear across all forms. Therefore, the age & gender features are used in the multiple form-specific classifier approach and the age, gender & symptom_form features in the single all-form classifier approach. This implies a 2-dimensional and 3-dimensional feature space for respective approach, which is notably smaller than using all features. Note that these three features are included in the all features group, and therefore function as a val-
idation of how valuable the other features are. If using all features performs similarly to the selected features, one could argue that the other multiple-choice questions are redundant, and that these specific features might be a relevant feature selection. This can be seen as a comparison of how much information gain other features provide.

### 4.1.2 ICD code distribution

It is relevant to present the ICD code distribution both across forms and collected into a single dataset, in order to understand the limitations of the dataset. This is indicative of the possibility of cross-form prediction and the ICD specificity, regardless of the feature distribution and range in the various forms.

![Figure 4.2: ICD distribution in the dataset; a) cumulative percentage and b) area density (logarithmic scale).](image)

Figure 4.2: ICD distribution in the dataset; a) cumulative percentage and b) area density (logarithmic scale).
As seen in Figure 4.2b, the dataset is comprised of only a small set of diagnoses, as the top 200 most common diagnoses constitute over 90% of the data. The 200th most common diagnosis has 20 data points, which poses a serious challenge. The corresponding area plot, Figure 4.2b, reveals that a large majority of the data is found in a small set of the most common diagnoses.

Figure 4.3 illustrates several interesting features. Figure 4.3a sheds light on the fact that several ICD codes appear across forms, which opens up for cross-form diagnosis. Figure 4.3b reveals interesting aspects in relation to each individual form. Due to the great number of ICD codes, it is infeasible to illustrate all ICD codes over the forms. Therefore, ICD codes are grouped according to occurrence in forms. The graph illustrates that a large fraction of the ICD codes appear across many forms. Although several forms are more specific and contain less diagnoses, they still share a relatively large fraction of the diagnoses with other forms. This might indicate that the form identity and their associated features contain a modest level of information.
4.2 Single all-form classification approach

An initial approach was to have one classifier for the whole dataset, where the form-type merely would serve as a feature (as seen in Figure 4.1, the column symptom_form). As the single all-form classifier approach has a large group of outcomes, it is difficult to illustrate the result on a diagnosis-level. However, the averaged F1-score and accuracy for all feature groups (all features, selected features (age, gender & symptom_form) and text) can be investigated across the different models, as reported in Figure 4.4.

4.2.1 F1 score & accuracy

Figure 4.4 illustrates the results with the standard error for the F1 scores for the runs.

![Figure 4.4: Averaged F1-scores with standard error for the different feature groups across the different models. Dashed lines correspond to the average accuracy for all features (cyan), selected features (age & gender) (yellow) and text (dark gray).](image)

From Figure 4.4 we can conclude that the text performed substantially better than the other feature groups. A 2-way ANOVA test was performed on the results to investigate the effect of the two variables...
(models and feature groups) and their interaction. For the 2-way ANOVA test, only models that classify all feature groups (all features, selected features, and text) have been considered (SVM, kNN and NB). Table 4.1 holds the F-statistics and p-value for the 2-way ANOVA test with feature set and model as main factors.

<table>
<thead>
<tr>
<th>Source</th>
<th>F-statistics</th>
<th>p-value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature groups</td>
<td>430</td>
<td>6.4e-16</td>
<td>****</td>
</tr>
<tr>
<td>Models</td>
<td>122</td>
<td>3.0e-11</td>
<td>****</td>
</tr>
<tr>
<td>Interaction</td>
<td>308</td>
<td>2.7e-16</td>
<td>****</td>
</tr>
</tbody>
</table>

The results in Table 4.1 indicate that there are significant differences between models as well as feature groups. A post-hoc Tukey was performed for multi-comparison to determine what specific groups for the variables have significant differences (Figure 4.5).

![Figure 4.5](image-url)

**Figure 4.5:** Results of a post-hoc Tukey test on a) feature sets and b) models respectively. The significance level is set to 0.05%

The post-hoc Tukey tests visualised in Figure 4.5 reveal that there are pairwise significant differences between different feature groups and models. In particular, there is a significant difference in performance between all the feature groups, and there is a significant difference in terms of performance between the SVM and the other models.
4.3 Form-specific classification approach

Due to the great differences between forms in terms of number and quality of features and number of data points, a form-based classification approach emerged. The vast variations between forms (presented in Section 4.1) was an indication of why a shared classifier might not be as beneficial as several classifiers. The multiple form-specific classifiers approach results in less training points as the classifiers are only exposed to points relevant to a specific form. However, the suggested approach heavily reduces the feature space. Furthermore, this also implies that some ICD codes that appear across forms might not be frequent enough to be included here.

X group

As mentioned in the Methods, ICD codes present in less that 100 instances have been removed. In an attempt to preserve data and increase usability, all ICD codes with less than 100 points have been labelled X. This is a design choice and can be interpreted as a separate class in a prediction scenario.

4.3.1 F1 score & accuracy

Due to the great number of forms, a selection of forms are presented in this section. It is relevant to consider forms that have both text data and a reasonable number of outcomes (ICD codes). The forms that are presented are Allergies, Insect bite and Ingrown nail. Remaining forms can be seen in the Appendix. Table 4.2 lists important characteristics for the selected forms.

<table>
<thead>
<tr>
<th>Form</th>
<th># features</th>
<th># tot. p.</th>
<th># test p.</th>
<th>Text question &amp; avg #words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergies</td>
<td>6</td>
<td>1224</td>
<td>306</td>
<td>Describe your trouble (17)</td>
</tr>
<tr>
<td>Insect bite</td>
<td>9</td>
<td>380</td>
<td>95</td>
<td>Describe what happened (19)</td>
</tr>
<tr>
<td>Ingrown nail</td>
<td>7</td>
<td>832</td>
<td>208</td>
<td>Describe your issue (23)</td>
</tr>
</tbody>
</table>
Allergies

Figure 4.6 presents the main results from classification with the form-specific classification approach on the Allergies form.

**Figure 4.6**: F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are all features (cyan), selected features (age & gender) (yellow) and text (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the rightmost group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for all features (cyan), selected features (age & gender) (yellow) and text (dark gray).
In Figure 4.7, the confusion matrices for the best-performing model for each feature group (all features, selected features and text) are presented.

(a) Figure 4.7a is for all features where NB performed best with an accuracy of 12%. Figure 4.7b is for selected features (age, gender) where the RF performed best with 21% accuracy. Finally, Figure 4.7c is for text where the SVM performed best with 39% accuracy. (Random accuracy: number of classes is 5)

From Figure 4.6 and Figure 4.7, one can infer that different models performed well for different feature groups. In terms of accuracy and F1-score, the best performing model was SVM for the text feature, with 39% accuracy. When using all features, the results were very poor with only 12% accuracy, and slightly better with age&gender (21%).
Insect bite

Figure 4.8 presents the main results from classification with the form-specific classification approach on the *Insect bite* form.

**Figure 4.8:** F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are all features (cyan), selected features (age&gender) (yellow) and text (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the right-most group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for all features (cyan), selected features (age&gender) (yellow) and text (dark gray).
In Figure 4.9 the confusion matrices for the best-performing model for each feature group (*all features*, *selected features* and *text*) are presented.

![Confusion Matrices](image)

**Figure 4.9**: Confusion matrices for the best-performing models for the three feature groups. Figure 4.9a is for *all features* where SVM performed best with an accuracy of 36%. Figure 4.9b is for *selected features* (*age, gender*) where the NB performed best with 49% accuracy. Finally, Figure 4.9c is for *text* where the SVM performed best with 55% accuracy. (Random accuracy: number of classes is 3)

Figure 4.8 and Figure 4.9 informs us that SVM performed well for both *all features* (55%) and *text* (36%), whereas NB performed best when only using the *selected features* (*age&gender*) (49%). Overall, the best performing model was SVM for the *text* feature, however, the accuracy when using only *selected features* (*age&gender*) was not far behind. For these results, we should also consider the fact that there are only three possible outcomes. However, the results are considerably better than random.
Ingrown nail

Figure 4.10 presents the main results from classification with the form-specific classification approach on the Ingrown nail form.

Figure 4.10: F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are all features (cyan), selected features (age& gender) (yellow) and text (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the rightmost group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for all features (cyan), selected features (age& gender) (yellow) and text (dark gray).
In Figure 4.11 the confusion matrices for the best-performing model for each feature group (*all features, selected features* and *text*) are presented.

![Confusion Matrices](image)

**Figure 4.11**: Confusion matrices for the best-performing models for the three feature groups. Figure 4.11a is for *all features* where RF performed best with an accuracy of 31%. Figure 4.11b is for *selected features* (*age, gender*) where the NB performed best with 39% accuracy. Finally, Figure 4.11c is for *text* where the SVM performed best with 51% accuracy. (Random accuracy: number of classes is 5)

Figure 4.10 and Figure 4.11 illustrates that various models performed well depending on the different features used. For the *all features* the RF performed the best (31%), and for *selected features* (*age&gender*) NB outperformed the other models with 39% accuracy, and for *text* the SVM was best with 51% accuracy. Overall, the best performing model was SVM for the *text* feature. It it notable that *text* was the best performing feature group for all models that classifies text with the other feature groups (NB, kNN, SVM).

### 4.3.2 Collective statistics

A 2-way ANOVA test was performed on the results (F1-score) of each *form-specific classifier* to investigate the effect of the two variables (models and feature groups) and their interaction on the within-form performance. Only classifiers that classify all feature groups (*all features, selected features* and *text*) have been considered (SVM, kNN and NB). The individual results for each form are found in Appendix. The results where significant main effects were reported were subject to post-hoc Tukey tests for pair-wise comparisons, presented in Table 4.3.
Table 4.3: Results per form from a post-hoc Tukey tests where a significant main effect was reported from the 2-way ANOVA test. Each entry corresponds to a form and the pair-wise comparisons of feature groups followed by pair-wise comparisons of the models. An asterisk (*) indicate significantly different results in terms of performance between the two groups and NS indicate that no significant difference was reported. The significance level was chosen to 0.05. Only forms with multiple-choice questions and text data have been considered.

<table>
<thead>
<tr>
<th>Form</th>
<th>Text All features</th>
<th>Text Age&amp;gender</th>
<th>Age&amp;gender All features</th>
<th>NB SVM</th>
<th>kNN SVM</th>
<th>kNN kNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergies</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Insect bite</td>
<td>NS</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Ingrown nail</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Cold sore</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>NS</td>
</tr>
<tr>
<td>Sleeping problems</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Urinary tract infection</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Eye infection</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>General health questions</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>*</td>
<td>*</td>
<td>NS</td>
</tr>
<tr>
<td>Asthma</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Eczema</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>*</td>
<td>*</td>
<td>NS</td>
</tr>
<tr>
<td>Fever</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>*</td>
<td>*</td>
<td>NS</td>
</tr>
<tr>
<td>Constipation</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Wind</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Headache</td>
<td>NS</td>
<td>*</td>
<td>*</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

The results in Table 4.3 indicate that all the feature groups are significantly different in half of the forms. In total, the text feature is significantly different from all features and age&gender in 11/14 cases. The age&gender feature group presents significant differences from text and all features in 9/14 cases, and there is one instance where there is no significant difference between the feature groups. In terms of models, we see that the performance is significantly different in three cases, and that the SVM is significantly different from NB and kNN on eight occasions. In total, this implies that the SVM performs with significantly different accuracy compared to the other two models for the majority of the forms. In summary, all forms display significantly different results either in terms of feature group variations or significantly different results in terms of model variations, if not both. The results do, however, indicate that the feature grouping primarily affects results.
Chapter 5

Discussion

5.1 Single all-form classification vs. multiple form-specific classification

The single all-form classification approach performs reasonably well on text data given the large output space, and shows significant differences between all feature groups and for the SVM compared to the two other models. There is a clear advantage in utilizing text data instead of multiple-choice questions and SVM is the unambiguous winner for prediction.

While the single classifier approach performs well, the overall accuracy and F1-score is better when classification is done per form. There is, however, an advantage to only having one classifier as it is less complex and utilizes all the data points. On the other hand, as the features are not disseminated across forms, there is no real point in having one universal classifier, as the feature space grows unnecessarily complex. An argument for keeping the dimensionality as low as possible is to avoid the curse of dimensionality. Furthermore, while it is more complex to have several classifiers, they individually require less as they are computationally smaller.

5.2 Main findings in form-specific diagnosis

The majority of forms display significant differences between the text data and the other two feature groups, which is a clear indication that using text has significant differences in performance compared
to only using multiple-choice questions. There are also several instances where all features and age&gender do not present significant differences in terms of performance. The fact that age&gender have equal performance to that of all features is a clear indicator that the questions asked may be redundant and too general in most forms, as they do not present increased information gain compared to merely using age&gender. On the instances where all features display higher performance than age&gender, the most likely explanation is that the questions are of higher quality. However, one must also factor in the number of outcomes for the various form, as some forms have fewer diagnostic outcomes, which simplifies the diagnostic task. It is reasonable to believe that as the number of diagnoses grow so does the importance of well-designed questions. A few forms only have two outcomes (e.g. the Constipation form in Appendix), which explains the high accuracy. This might signal that the expected range of outcomes of the form is not complex enough to formulate as a classification task.

In summary, the text contributes with a high degree of information compared to other feature groups, and its performance is in the majority of forms significantly different from all features and age & gender. Where all features and age & gender do not have significantly different results, the quality of the multiple-choice question are expected to be low. There are however close to no occasions where the multiple-choice questions (either all features or age & gender) outperform text data. Clearly, the multiple-choice questions leave room for improvement. In terms of models, the overall tendency is that the performance of the SVM is significantly different from the other models where a significant main effect has been reported. While the SVM tends to be the winner in terms of accuracy, the performance of NB is still satisfactory. In fact, for some forms NB compares favorably to SVM. As the data varies between forms, the desired outcome might not be to determine the optimal classifier, but rather the preferred one on a form-basis.

5.3 Text data and multiple-choice questions

Text analysis has not been extensively used for medical diagnosis for a larger diagnostic field than specific diagnoses [37]. Most identified instances have been on datasets of small feature spaces with numerical, biometric data. A recent study attempted to reformulate the diag-
nostic problem as an information retrieval problem in order to utilize text data better. Their results were moderate but indicated that text mining can be used for diagnostic prediction [37]. These findings are similar to those presented in this thesis, yet different in their nature in terms of input. Instead of using keywords, as in the mentioned study, this thesis is based on the patient’s self-described assessment. The self-described assessment is believed to be more verbose and, in general, containing more words. One issue that was mentioned in the presented study is that text data in this context often suffers from circumlocution (overly verbose when a shorter description would suffice) [37]. Research has been done in attempt to combat the curse of circumlocution, which similarly had automated diagnosis in mind [38]. The mentioned research indicates that there is still work to be done in order to fully utilize the information in text, and that research within wide-domain diagnostic prediction on text data is limited. The results presented in this thesis support, however, that text data is promising as it holds much information.

One important aspect to consider is that the text data in this study is exhausted in terms of information (as the question is open-ended) and the true potential of well-designed questions is left unexplored. Therefore, there is still potential improvement left for the multiple-choice questions, as they still lack in performance comparing to text data and, on instances, using only age&gender results in similar performance. It is also noteworthy that handling multiple-choice questions is potentially preferred due to easier comparisons, which cannot be said about text that has a qualitative nature. Therefore, one interesting question to consider could be whether it is possible to construct questions that capture the current essence of the text data. How would one design such questions?

5.4 Model comparison

In terms of model choice, the results in this study is in line with the related work, where no algorithm can be eliminated before tested due to the uncertainty and peculiarity of the data [30]. SVMs have been successfully applied in several cases [27][32], so it is not a surprise that it performs well on several of the diagnostic tasks in this study. There is also heavy support that RFs perform well compared to other clas-
classifiers on a range of different datasets [32]. Although several models were tested, there are instances where the baseline model (NB) outperforms other models in this study. However, it has been reported that for certain diagnostic tasks the NB classifier has outperformed several other models [39]. A well-performing baseline model indicates that there is room for improvement. It may seem surprising that the MLP did not perform well in terms of F1-score or accuracy, given the extensive application in research. One of the possible explanations could be that the number of data points was not sufficient. As both RF and NB performed well and tend to do so on smaller datasets, other models might have had better performance with larger datasets.

5.5 Possible use

The primary use case for CDSSs is the application in primary health care, primarily as a tool for clinicians. This could be furthered extended in several ways, e.g. suggesting probable diagnoses or possible next questions. There are, however, further possible extensions for CDSSs beyond the perspective of the clinician, such as identifying diagnoses that require physical examination. This could be cost-effective as KRY currently do not assess whether a patient’s condition is suitable for digital consultation before a meeting takes place. A CDSS could also be implemented to evaluate responses as they are entered, and thus potentially have dynamic symptom form questions dependent on the specific answers. The format of digital health care, which is a digital platform for both clinicians and patients, is a good basis for the various possible applications of CDSSs as it can be integrated on both the clinician’s and patient’s end.

5.6 Limitations

Initially one should realize that the forms might not be as straightforward to compare as one would prefer. For all these comparison it is important to factor in the number of questions asked, the number of possible outcomes, the number of training points and the type of diagnoses it aims to target. One should also consider the nature of the text data and that it might be heavily dependent on the actual specificity of the form (in the form Allergies it might not cross a user’s mind to actu-
ally repeat the term "allergy" in the text description but rather mention the symptoms (e.g. snotty, itching eyes), which probably would not be the case in the form General health questions). One limitation is thus that the variations in the forms make it hard to generalize in terms of feature groups and models. Although it is tempting to have a larger group of shared features, it is impractical from a user’s perspective. If we, however, can conclude that something is true for a large set of forms, this might be a reasonable generalization given the dataset.

The size of the dataset might also be a limiting factor, as more data could have affected the performance of the MLP. Insufficient sizes of datasets are however a common case for machine learning tasks, and there are models in this study that are known to perform well to combat limitations like these. In addition to this, only random search was used to find optimal parameters. In the best-case scenario one would have wanted to perform an extensive grid search instead. While the parameters for most models were believed to have a sufficient range, the performance of the MLP might have been better if more architectures were explored and a larger set of parameters for training were tested (such as learning rate, initialization parameters, etc). However, as previously mentioned, the performance of the MLP is primarily thought to be due to the limitations of the dataset. It is also important to note that this study was performed with only one data provider. In an optimal setting, several sources of data would have been used and compared. This does, however, imply that the data would need to be of the same format for a valuable comparison, which is deemed unlikely.

5.7 Ethics & sustainability

The digitalization of health care will affect society in both economic and ecological aspects. Clinical decision support is only a component in the set of consequences that comes with a digital space for health care. As digital health care is expected to improve work environments and quality [2], this implies a more sustainable development of health care with improved conditions on several ends. The possible increased interest in personal health care also opens up for earlier detection of diseases, which could offer an increased economic sustainability. There are, however, studies that show how increased
prevention in some cases also add to health care costs \[40\]. One mentioned example is the potentially increased screening costs that would exceed the savings of avoided treatment, where only a tiny fraction of the population would have been subject to illness due to the absence of preventative care \[40\]. There has also been a reported relationship between the number of hospital visits and hospitalization rate \[41\]. As digital health care allows for increased access, this is an aspect that deserves serious consideration. It also implies a need for extensive investigation to target the objectives for the preventative care and access of digital health care, in order to have a positive economical development.

Decision support systems, as part of either digital or physical health care, requires consideration in terms of sustainability. From an economic viewpoint, the implications of a decision support system is heavily dependent on its reliability. If a CDDS has a negative effect on the clinician’s decision making or performance, this could potentially lead to economic disadvantages in the case of wrong diagnosis. This demands high requirements for CDDSs before put in practice, and naturally with a reasonable expectation on the physician’s capacity to evaluate the CDDS’s credibility. If CDDSs have a positive impact on the physician’s performance and the diagnostic outcome, this could lead to both economic and ecologic sustainability. The cost of a digital health care visit in terms of transportation is more cost-effective on both ends as neither the physician nor patient are not bound to a certain location. As previously mentioned, there are other possible use cases for CDDSs, such as detection of suitability for digital health care. This is an excellent opportunity for economical improvement. Implementing a CDDSs that runs while the patients fill out the form could also be beneficial as this could further prepare the clinician for the meeting, which is time-effective.

The ethical perspective on CDDSs, however, is most certainly in relation to the data that the models need to be exposed to in order to make qualitative predictions. To enable prediction in either a digital or physical space, the patient data needs to be exploited in some sense. Patients’ health care information accounts as sensitive personal data, and must therefore be protected more than regular personal data in Sweden \[42\]. The General Data Protection Regulation (GDPR) also imposes new regulations that must be considered. For instance, data must be subject to pseudonymity, which means that when handling
personal data it must be done so that it cannot be attributed to a specific individual without additional data \[^{13}\]. The text data in this study is a fitting example, as if it contained personal data such as personal numbers, it would not have adhered to the regulations. These types of considerations are critical, and the general use of personal health data is difficult as it is considered to be more sensitive than regular personal data. The ethical dilemma is in relation to the fact that there are competing interests, as one would expect the quality of CDSSs to be reliant on the type of data provided, whereas there is a simultaneous need for personal integrity. There are personal benefits to technologically advanced health care (such as highly developed tools for clinicians, like CDSSs), while there also is a personal interest in owning your personal data.

To summarize, CDSSs and digital health care could have both positive and negative impacts on many aspects in relation to sustainability and ethics, and the digital health care providers are advised to take great part in the public debate to share the benefits of digital health care and to gain increased knowledge of the potential risks.
Chapter 6

Conclusions

The results in this study indicate that the text data provides more diagnostic information than the multiple-choice questions. Furthermore, the findings on several symptom forms indicate that using all multiple-choice questions result in similar accuracy when compared to using only age & gender. The fact that only the combination of age & gender delivers comparable performance to all multiple-choice questions is a clear indication that the multiple-choice questions hold a potential of improvement, and that the subset of age & gender contains much information. The results also demonstrate that the SVM tends to perform particularly well in some cases, although NB and kNN oftentimes are suitable alternatives. Likewise, for prediction on multiple-choice questions, RF serves as an excellent choice as well.

6.1 Future work

As age & gender performs equally well as the whole feature set on certain forms, one could argue that they would be an appropriate choice for feature selection. One future investigation could be if other features holds the same potential and could be selected instead of age and/or gender. In the case of re-design one could ask the question if it would be better to stick with using text instead of questions, or to investigate how one can ask better questions. As previously mentioned, it would be of interest to investigate how multiple-choice questions could be improved in order to match the information level of the text data. The text data could also have benefited from feature selection in order to reduce the feature space, which is left unexplored. If one
seeks to optimize the overall accuracy, one might find it worthwhile to investigate how age & gender performs when aggregated with text data. Furthermore, while the optimal parameters were identified on a feature group level, the actual models were chosen with a comparative aspect in mind. Therefore, it seems valuable to research more complex models within the NLP field, such as deeper networks, for the text analysis. As SVMs have reported impressive results on text analysis, it might also prove fruitful to adapt methods from presented work [34][35]. It would also be interesting to include additional models that have been frequently used in research, such as BN, for comparative reasons. In terms of presentation, another possible extension would be multi-class classification where diagnoses are ranked according to probability, which provides additional information to physicians. This might better match the decision making process of the clinician and also present them with critical corner cases that have low probability, yet are important to rule out. It would in relation to this be interesting to compare how well the predictive system performs compared to physicians when provided with the identical information, as the majority of the information is believed to still primarily lie in the video chat.
Bibliography


[38] S. Ieong I. Stanton and N. Mishra. “Circumlocution in Diagnostic Medical Queries”. In: Microsoft Research (2014).


Appendix A

Results from form-specific classification
Cold sore

Figure A.1 presents the main results from classification with the form-specific classification approach on the Cold sore form.

**Figure A.1:** F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are all features (cyan), selected features (age&gender) (yellow) and text (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the rightmost group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for all features (cyan), selected features (age&gender) (yellow) and text (dark gray).
Sleeping problems

Figure A.2 presents the main results from classification with the form-specific classification approach on the Sleeping problems form.
Urinary tract infection

Figure A.3 presents the main results from classification with the form-specific classification approach on the Urinary tract infection form.

Figure A.3: F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are all features (cyan), selected features (age& gender) (yellow) and text (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the rightmost group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for all features (cyan), selected features (age& gender) (yellow) and text (dark gray).
Eye infection

Figure A.4 presents the main results from classification with the form-specific classification approach on the *Eye infection* form.

**Figure A.4:** F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are all features (cyan), selected features (age&gender) (yellow) and text (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the right-most group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for all features (cyan), selected features (age&gender) (yellow) and text (dark gray).
General health questions

Figure [A.5] presents the main results from classification with the form-specific classification approach on the General health questions form. Due to the large amount of outcomes in this form, labels have been removed in [A.5].

![Figure A.5](image-url)

**Figure A.5:** F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are *all features* (cyan), *selected features (age& gender)* (yellow) and *text* (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the rightmost group in each model plot is the averaged result (labelled *avg/total*). The dashed lines correspond to the average accuracy for *all features* (cyan), *selected features (age& gender)* (yellow) and *text* (dark gray).
Asthma

Figure A.6 presents the main results from classification with the form-specific classification approach on the *Asthma* form.

*Figure A.6:* F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are *all features* (cyan), *selected features* (*age* & *gender*) (yellow) and *text* (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the right-most group in each model plot is the averaged result (labelled *avg/total*). The dashed lines correspond to the average accuracy for *all features* (cyan), *selected features* (*age* & *gender*) (yellow) and *text* (dark gray).
Eczema

Figure [A.7] presents the main results from classification with the form-specific classification approach on the *Eczema* form.

Figure A.7: F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are *all features* (cyan), *selected features* (*age* & *gender*) (yellow) and *text* (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the rightmost group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for *all features* (cyan), *selected features* (*age* & *gender*) (yellow) and *text* (dark gray).
Fever

Figure A.8 presents the main results from classification with the form-specific classification approach on the Fever form.
Constipation

Figure A.9 presents the main results from classification with the form-specific classification approach on the Constipation form.

**Figure A.9:** F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are all features (cyan), selected features (age & gender) (yellow) and text (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the rightmost group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for all features (cyan), selected features (age & gender) (yellow) and text (dark gray).
Wind

Figure A.10 presents the main results from classification with the form-specific classification approach on the Wind form.

Figure A.10: F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are all features (cyan), selected features (age & gender) (yellow) and text (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the rightmost group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for all features (cyan), selected features (age & gender) (yellow) and text (dark gray).
APPENDIX A. RESULTS FROM FORM-SPECIFIC CLASSIFICATION

Headache

Figure A.11 presents the main results from classification with the form-specific classification approach on the Headache form.

Figure A.11: F1-score across diagnoses for the different feature groups and average accuracy for the feature groups across models. For each model result we can observe the F1-score for each individual diagnosis for the three feature comparisons that are all features (cyan), selected features (age & gender) (yellow) and text (dark gray). The diagnoses (ICD codes) are ordered by the number of test points, and the rightmost group in each model plot is the averaged result (labelled avg/total). The dashed lines correspond to the average accuracy for all features (cyan), selected features (age & gender) (yellow) and text (dark gray).