Detection of Health Problems based on Medical Historical Records

Alexandra Matei
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
d.matei@student.utwente.nl

Abstract

Demand for community home care services continues to increase due to demographic changes. Unfortunately, the home care organizations can not cope with this boost of people in need of home care due to a significant shortage of employees. The resulted heavy workload pressuring the professional care givers may have negative impact on the quality of the provided care. For instance, professional care givers postpone updating the medical records of the patients due to a lack of time. As a solution for speeding up the process, this study proposes a model capable of detecting health-problems based on the historical medical records of the patients and their personal profile (age and gender). Relationships between health-problems, as well as relationships between health-problems and personal profile were identified and integrated in a classification-machine-learning-based model. The obtained model is able to predict 31 different health-problems with an overall accuracy of 89.96%, the accuracy ranging from 99.3% to 67.2% per problem.

Keywords

Home Care, Detection of Health-problems, Machine Learning, The Omaha System

1. INTRODUCTION

For the past decades, the world is facing a fast aging population, caused by the decline in fertility rate and the rise of longevity [6]. For a long period in the human history, the proportion of elderly (individuals over 65) did not exceed more than 4% of a country's population. Currently, this proportion reached roughly 15% and it is expected to rise to 25% by 2050 [4]. Among this increasing old population, the trend of receiving home care has significantly spread in the past years [21]. The number of health care professionals was not able to keep up with this boost of elderly in need of home care, causing a shortage of employees in the home care industry. For instance, in 2018, a deficit of 100.000 health care employees was predicted in the Netherlands for the next four years [20].

To overcome this deficit, the professionals care givers are required to take care of more patients in a shorter period of time. However, this can not serve as a long-term solution since the high workload of health care professionals turned out to have a negative impact on the quality of care and safety of the patients [7]. Due to a lack of time, professional care givers bypass performing certain activities, mostly administrative tasks such as the development of assessments and update of the medical records [3].

In order to cope with these challenges raised by the heavy workload of professional care givers and the necessity to provide qualitative care, this study proposes a model capable of detecting health-problems. Nowadays, the professional care givers make use of electronic health care records applications when assessing the health condition of a patient [10]. Through the assessment process, they have to fill in a check-list with signals and symptoms of the patient, which helps them in identifying health-problems. This process usually takes about 20 minutes per patient [2], however the professional care givers also need to provide physical care to the patient and travel from a patient to another. Alternatively, the proposed model is identifying these health-problems based on the already existing health-problems in the medical records of the patients and their personal details such as gender and age. The identified problems by the model will suggest the health care professionals which health-problems might represent a risk for their patients. In this way, the health care professionals do not have to perform an assessment in order to find these health-problems. For example, problems concerning physical activity may raise risks for other health-problems such as cardiovascular diseases [9]. Linking the problems concerning physical activity to cardiovascular diseases would be of great benefit for the health care professionals by saving the time needed to perform an assessment. This study aims to investigate to what extent it is possible to create a model of detection health-problem based upon on the already existing health-problems in the assessments of the patients and their personal profile.

2. RESEARCH QUESTIONS

This research will answer the primary research question below. However, the proposed primary research question is built on the listed secondary research questions which should be answered first.

PRQ To what extent it is possible to identify health-related problems considering the historical medical records and the profile of a patient using data-dependent techniques?

SRQ1 To what extent do health-related problems correlate to each other?

SRQ2 To what extent does the profile of a patient influence the detection of health-related problems?

31th Twente Student Conference on IT July 5th, 2019, Enschede, The Netherlands.

SRQ3 What is the most optimal data-dependent technique in detecting health-related problems based on historical medical records and the profile of a patient?

SRQ3.1 What is the accuracy of predicting health-related problems using the most optimal data dependent technique?

3. DATA SET

The data set contains information from the electronic health records of the patients. This data has been provided by a company which produces the software for the electronic health records, the company having the consent of the health care organizations to use the data for research purposes. From these electronic health records, the following have been extracted: the health-problems, the dates of the assessments and the dates when the problems have been removed from the treatment plan of the patient. The data of the health records has been joined with the gender, date of birth and city of residence of the patient. The resulted data set contains 246,271 assessments registered between 2011 and 2019 for 81,863 patients (41% males and 59% females) in home care all around the Netherlands. The representation of the data model is depicted in **Figure 1**.

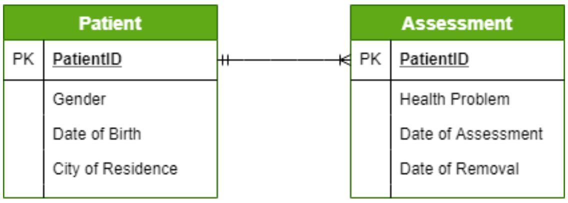


Figure 1: Representation of the Data set

3.1 Omaha System

The health-problems which are aimed to be identified in this paper are part of the Omaha System. The Omaha System is a classification system of health-problems designed to describe the health care of patients. It has been chosen for the scope of this study due to its wide adoption in the home care industry, being used by over 22,000 health care professionals [15]. The Omaha System identifies 42 different health problems divided in four main domains: psycho-social, psychological, environmental and health related behaviour. However, this proposed study is only focusing on 31 problems, several problems being omitted due to their sensitive nature. The list of problems and their definitions are listed in *Appendix B*.

4. RELATED WORK

There is limited research performed in the area of health-problems detection for patients provided with home care. Currently, when it comes to the home care industry, the main focus is on developing systems which are able to foster the independence of elderly people living at home. This type of systems include lifestyle monitoring, emergency detection and alarms systems for safety [17]. However, there is still some research related to our study in terms of either the home care industry, methodology or data input.

4.1 Detection of gait-related problems

Pogorelc et al. proposed a system capable of detecting gait-related problems of elderly people living at home by identifying anomalies in their manner of walking [18]. This detection system is an extension to a motion-capture system for elderly people containing body-worn tags and wall-mounted sensors. By classifying the movement data received from the sensors using seven different classification models, the model succeeded in detecting gait-related problems (Parkinson's disease, hemiplegia, pain in the leg, pain in the back) with an overall accuracy of 99%. Although this study aimed to detect health-problems for elderly people in home care, the input data was only collected from the sensors concerning movement patterns and the number of problems to be predicted was limited four.

4.2 Heart disease prediction

A research aiming to detect health problems with similar techniques to the proposed study is the one carried out by Soni et al [13]. For this project, three classification models (Decision Trees, Naive Bayes and KNN) were used to predict heart diseases based on the characteristics of an individual such as age, gender, weight etc. In order to improve the accuracy of these models, correlations between the individual's characteristics and heart diseases were identified using data mining techniques. The resulted model is able of predicting with 89.0% accuracy, however it can only predict whether an individual is suspected to have heart problems, without specifying the exact problem.

4.3 Chronic disease prediction

A similar research to our study is the one performed by Chen et al., but on hospital data [5]. Based on the medical records of over 20 million patients of hospitals in China, a convolutional neural network was created aiming to predict multiple chronic problems such as cerebral infarction, diabetes and hypertension. The medical records used for this research were more complex, in addition to gender, age, city and already existing problems, they also contained unstructured data about the habits of the patients and the notes of the doctor. It appeared that the extra data from the doctor's notes improved the accuracy of detecting health-problems from roughly 50% to 94%.

4.4 Contribution

The results of this research would be of great value especially in the home care industry, since the research in detection of health-problems is limited. Also, the proposed research is aiming to detect more than three or four health-problems (as in the examples above), covering 31 problems defined by the Omaha System.

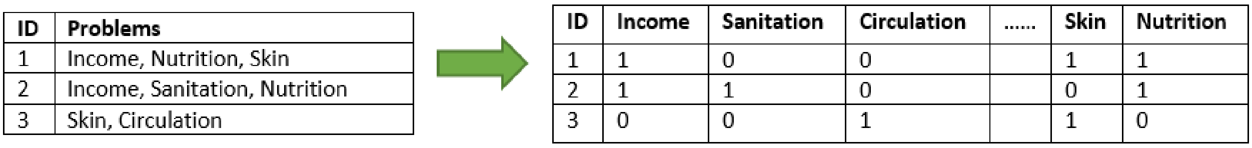
5. METHODOLOGY

To be able to answer the primary research question, the secondary research questions have to be answered first. A different method was used in order to answer each secondary research question. However, the feature selection algorithm used for answering **SRQ1** has also been integrated in the classification model built to answer **SRQ3**. For answering **SRQ2** different charts have been plotted. Every method is discussed in one of the following sections.

5.1 Feature Selection

Starting with **SRQ1**, a feature selection algorithm has been applied on the data set aiming to find possible correlations between health-problems. Usually, feature selection is used in building machine-learning models, for selecting a subset of features which are the most relevant and discriminative in predicting. For the purpose of this study, we used a correlation-based feature selection, this method being capable of identifying correlations between features (in our case health-problems) [8].

Figure 2: Data Set Transformation



ID	Problems
1	Income, Nutrition, Skin
2	Income, Sanitation, Nutrition
3	Skin, Circulation

ID	Income	Sanitation	Circulation	......	Skin	Nutrition
1	1	0	0		1	1
2	1	1	0		0	1
3	0	0	1		1	0

In order to apply this correlation-based feature selection method, the data set (**Figure 1**) had to be manipulated. Hence, an one-hot encoding array has been created for each assessment containing all the health-problems listed in *Appendix B.* The health-problems present in the assessment were mapped to 1 and the ones not present to 0. A representation of this transformation of the data set is illustrated in **Figure 2**.

Afterwards, the data has been imported in RapidMiner, an environment designed for machine learning purposes [16] which also provides the selection algorithm ready to be applied on the data set. This algorithm has been applied 31 times, in each repetition a different health-problem represented the target in identifying health-problems to correlate with. For each health-problem, the algorithm calculated the correlation coefficient with every other problem and created a rank correlation in a descending order.

For evaluating the identified relationships between health-problems, we interpreted the obtained correlation coefficients according to the *Rule of Thumb for Interpreting the Size of a Correlation Coefficient* [11] (illustrated in **Table 1**). To validate the discovered relationships between the health-problems, we conducted an online survey with 41 health care professionals around the Netherlands. Through the survey they had to specify to what extent they agree with the correlations considering their expertise. The survey can be found in the *Appendix A.*

Table 1: Rule of Thumb for Interpreting the Size of a Correlation Coefficient

Size of Correlation	Interpretation
0.90 to 1.00 (-0.90 to -1.00)	Very high correlation
0.70 to 0.90 (-0.70 to -0.90)	High correlation
0.50 to 0.70 (-0.50 to -0.70)	Moderate correlation
0.30 to 0.50 (-0.30 to -0.50)	Low correlation
0.00 to 0.30 (0.00 to -0.30)	Negligible correlation

5.2 Data analysis

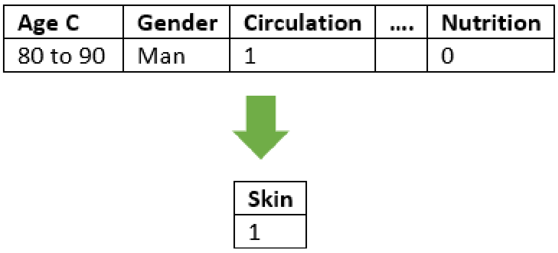
In order to answer **SRQ2** and find possible relationships between health-problems and the personal profile of a patient, graphic analysis combined with descriptive data analysis have been performed with respect to gender, age and region of residence. Graphic analysis is used to visualize the collected data in graphs such as histograms, line graphs, while descriptive analysis is used to provide summaries and explanation about the plotted data. For this, the data set had been uploaded to Qlik Sense, a query-based business intelligence application which provides charts ready to be populated with data [1].

At first, a bar chart was created with the scope of analyzing distribution of gender over health-problems. For each problem, the percentage of females and males diagnosed with that specific problem has been calculated and plotted in the bar chart. Secondly, in order to determine the relationship between age and health-problems, we created an extra attribute "*age at the date of assessment*" based on the date of birth and the date of assessment. To reduce the complexity generated by the variety of ages (alternating from 10 to 110 years old), the age attribute has been matched to an age category with a range of 10 years (for example: "80 to 90"). Hereafter, using this *age category* attribute, we computed the distribution of age over the assessments. Considering this distribution as baseline, the same distribution has been computed for each problem and plotted in a line graph, each line representing one of the age categories.

The last visualization aimed to discover possible relations between the region of residence and health-problems. In order to simplify the interpretation of the data, the city of residence has been mapped to a *region category.* This category was created according to the official division of regions of the Netherlands established by the European Union. Further, for each problem, the percentage of patients assessed has been plotted on the map of the Netherlands per region. Later on, a descriptive analysis has been provided for all three created visualizations (bar chart, line graph and the map) summarizing the observed results.

5.3 Classification models

In order to answer **SRQ3**, several classification models have been built with different machine learning methods aiming to detect health-problems based on the past assessments of a patient. For this, the data set used for feature selection (**Figure 2**) has been been merged with the gender and the age category (computed before for data analysis) of the patients. This data was loaded again in RapidMiner where the classification models were created.



Age C	Gender	Circulation	....	Nutrition
80 to 90	Man	1		0

Skin
1

Figure 3: Data Input target Output

Generally, a classification model attempts to draw conclusions from observed values. By providing the input, the model will try to predict the value of the outcome. For each health-problem, a different classification model has been created, where the assessments have been split in input and output. The input is represented by the gender, the age category and the health-problems detected in the past. Based on this input, the model will output either "1" if the targeted problem might represent a risk for the patient or "0" if the problem does not represent a risk. For example, considering the third assessment from **Figure 2** and assuming that the model aims to predict the health

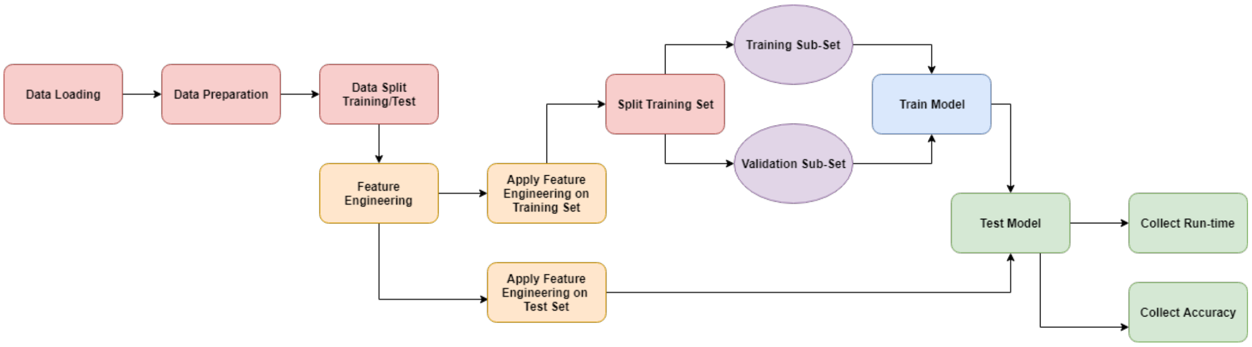


Figure 4: Classification Model Process

problem *Skin*, then: the input data is represented Circulation mapped to ”1”, all the other health-problems mapped to ”0”, the gender and age category, while the output data is ”1”. This example is illustrated in **Figure 3**.

A total of 31 such classification models have been built, each model aiming to predict a different health-problem listed in *Appendix B*. They were created applying nine different machine learning algorithms usually used for classification purposes, listed in **Table 2**. All these algorithms receive the same input and provide the same type of output, however they use different functions in order to come up with the result.

Table 2: Machine-learning algorithms used for classification

Name	**Type of classification**
Naive Bayes	based on Bayes' theorem
Deep Learning	based on artificial neural networks
Logistic Regression	based on logistic functions
Generalized Linear Model	based on generalization of multiple regression models
Decision Trees	based on tree representations
Random Forest	based on multiple Decision Trees
Gradient-Boosted Trees	based on loss function and weak learners
Support Vector Machine	based on creation of a separating hyper plane
Fast Large Margin	optimization of Support Vector Machine

The process of creating the models started with loading and cleaning the data, the rows containing missing values being excluded. Once the data was ready for modelling, it has been split into a training set used for training the model and a testing set used for evaluating the performance of the trained model. This division was in a ratio of 80 to 20 motivated by the Pareto's principle [22]. Next, automatic feature engineering has been applied on this data set which integrates feature selection (already used for SRQ1) and feature generation. Through feature generation, new possible features are discovered based on the existing features in the data set [14]. Through the automatic feature engineering, the most discriminative features are detected which will help in speeding up the training process.

Subsequently, the training process starts by first splitting the training set using the stratified sampling method. This method is commonly applied on nominal data as in our case '0' for *no-risk* and '1' for *risk* for the targeted health-problem to occur. This method ensures that the testing set and the validation set contain roughly the same proportion of the two values [23]. Once the training set is divided, one of the aforementioned algorithms (each at the time) is applied on the training set.

Lastly, the trained model is tested using the testing data set and for evaluation purposes, the run-time for training the data and the accuracy of the model have been collected. The steps through the creation process of the classification models are illustrated in **Figure 4**, for each step RapidMiner providing an algorithm ready to apply on the data set.

6. RESULTS

By applying the feature selection method, results have been obtained for **SRQ1**. The algorithm calculated the correlation coefficients of each health-problem with all the other problems defined by the Omaha System. Further, these correlation coefficients have been ranked by the algorithm from highest to lowest. The correlation coefficients have been analyzed and interpreted according to **Table 1**.

Table 3: Symmetrical correlations between health-problems (”CC” stands for ”correlation coefficient”)

Problems	**CC**
High Correlation	
Income, Care taking, Social Contact, Communication with community resources	≈ 0.90
Moderate Correlation	
Residence, Sleep, Role Change, Physical Activity, Health care supervision	≈ 0.55
Urinary function, Bowel function	≈ 0.50
Infectious/Communicable condition, Sanitation, Health care supervision	≈ 0.50
Low Correlation	
Nutrition, Medication Regimen, Sleep	≈ 0.45
Vision, Hearing, Oral health	≈ 0.40
Mental health, Cognition, Role Change, Health care supervision	≈ 0.40

After investigating all the correlations, we came up with three categories which would categorize them: *symmetrical correlations, unsymmetrical correlations and* ***no*** *correlations.* The first category includes associations of health-problems which are symmetrically placed in the rank correlation. Such a correlation is the one between the health-problems *Residence and Sleep*, where in the rank correlation obtained for *Residence*, *Sleep* is placed first with a

Table 4: Unsymmetrical correlations between health-problems (”CC” stands for ”correlation coefficient”)

Problem	**Problems which correlate**	**CC**
Moderate Correlation		
Infectious condition	Health care supervision, Sleep and rest	≈ 0.55
Reproductive function	Oral health, Infectious condition, Health care supervision,	≈ 0.55
Substance use	Health care supervision, Sleep and rest	≈ 0.55
Sanitation	Role Change, Infectious condition, Health care supervision,	≈ 0.50
Low Correlation		
Respiration	Sleep and rest Infectious condition, Digestion and hydratation	≈ 0.45
Digestion Hydratation	Health care supervision, Sleep and rest	≈ 0.45
Pain	Sleep, Bowel function, Infectious condition	≈ 0.40

correlation coefficient of 0.53. *Symmetrically*, in the rank correlation obtained for *Sleep*, *Residence* is placed first as well. All correlations which fulfill this *mirror condition* are listed in **Table 3**. However, we mostly identified associations of such symmetrical correlations which means that all health-problems in such an association have a symmetrical correlation between each other.

However, we also identified correlations between problems where the obtained ranking is not symmetrical (second category). Such a correlation is the one between *Digestion and Hydration* and *Sleep*, where in the correlation ranking obtained for *Digestion and Hydration*, *Sleep* is placed as second. On the other hand, in the rank correlation for *Sleep*, *Digestion and Hydration* is placed on the 8th position. All detected non-symmetrical correlations are listed in **Table 4**. Lastly, for some of the health-problems, no resulted coefficient correlation was higher than 30% which according to **Table 1** makes the correlations negligible. The problems for which no real correlations were found are:

- Skin
- Personal Care
- Circulation
- Neuro-muscolo-skeletal function

While aiming to find possible influence of the patient’s profile on the arising of health-problems (answer **SRQ2**), three characteristics of a patient have been investigated: gender, age and city of residence. For analyzing possible differences of health-problems per gender, the bar chart depicted in **Figure 5** has been plotted. For each health-problem on the X-axis, the percentage of males and females out of the total population of males and females who had been diagnosed with the problem is calculated (on the Y-axis). However, not all 31 health-problems have been illustrated, only the ones where differences in the distribution of gender over health-problems were approximately 3% or higher than 3%.

The bar chart demonstrates that there is an quite equitable distribution, however when it comes to the socio-economical problems, namely *Income, Social Contact, Care taking, Interpersonal Relationships* the percentages of the females is generally higher than the ones for males. On the other hand, men are more likely to be diagnosed with *Urinary* and *Bowel* functions problems (difference between genders is of 7% and 3%), while women seem to have more *Pain* and *Neuro-muscolo-skeletal* problems than men (the difference being about 3%). Likewise, females appear to encounter more problems when it comes to *Personal Care* than males, the percentages of females being 6% higher.

Looking at the age, the distribution of ages over the assessments has been calculated. According to this distribution, the largest proportion (40.7%) of assessed patients had ages between 80 and 90, followed by patients with ages between 70 and 80 (26.1%). The percentage of patients with ages between 60 to 70 and 90 to 100 is quite comparable, 12.6% and 10.8%, the rest of 9.8% being covered by the other age categories. In order to find any clues about the possible relations between the age and health-problems, the same distribution of ages has been calculated over health-problems. Comparing the basic distribution of ages with the distribution of ages over problems, the categories of ages which proved more fluctuations were 80 to 90 and 90 to 100. Noticeably, when it comes to *Hearing* and *Vision* problems, the categories of ages 90 to 100, but also 80 to 90 have a significant deviation from the baseline. Also, problems such as *Cognition, Circulation and Medication Regimen* are more characteristic to the

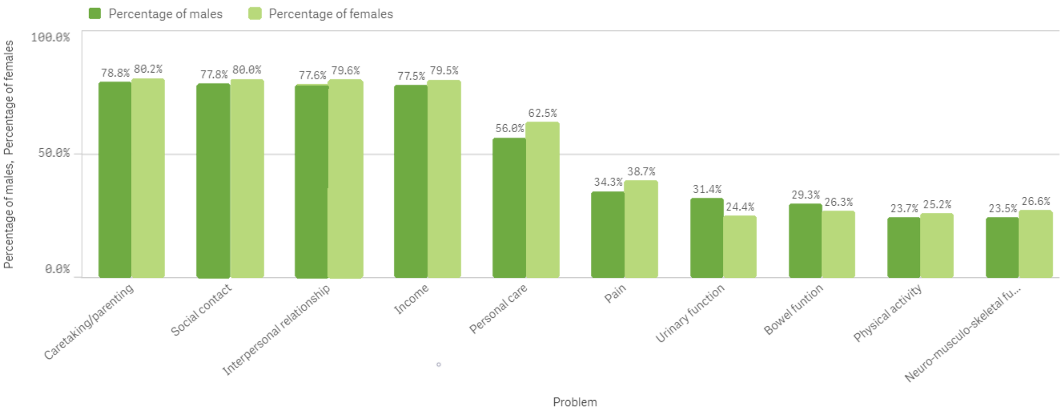


Figure 5: Gender distribution over health-problems

patients with ages between 80 to 90.

Last aspect to be considered from the patient-profile was the city of residence. Generally, the distribution of health-problems over the regions of the Netherlands appeared to be evenly distributed with some small exceptions. For each region the health-problems which seem to be of high predominance were as follows:

- **North Netherlands**: Bowel function, Oral health, Reproductive function, Substance Use, Vision
- **East Netherlands**: Cognition, Consciousness, Hearing, Oral health, Substance use, Vision, Sleep
- **West Netherlands**: Bowel function, Consciousness, Hearing
- **South Netherlands**: Reproductive function

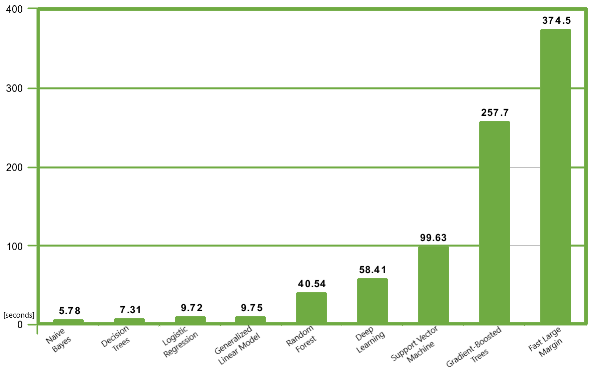


Figure 6: Average duration for training the model per technique

Intending to find possibilities to predict health-problems, based on the presence of other health-problems and answer **SRQ3**, nine different data-dependent techniques have been applied on the data set. We collected the run-time needed to train the model for assessing the efficiency of the technique. The accuracy of the prediction was calculated to assess the quality of the results. The run-times and the accuracy have been collected after applying every technique in the prediction of every problem (each technique being applied 31 times).

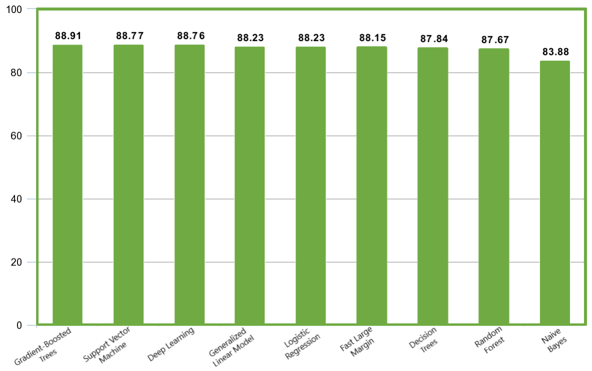


Figure 7: Accuracy of the predictive model per technique

The obtained results in terms of duration are depicted in **Figure 6**, where per each technique, the average time required for training the model (expressed in seconds) is shown. According to these results, *Naive Bayes* benefits from the fastest run-time (5.8 seconds), followed by *Decision Trees* with an average run-time of 7.31 seconds. Oppositely, *Fast Large Margin* technique required the highest amount of time for training (average of 374.5 seconds). In terms of accuracy of the prediction, the values are close to each other as illustrated in **Figure 7**. Although *Naive Bayes* proved to be the fastest, it also proves to provide the lowest accuracy in prediction with a 83.3% accuracy. And despite the fact that *Gradient Boosted Trees* demanded longer time for training the model, it accomplished the highest accuracy in prediction (88.91%).

As a follow up question to **SRQ3**, we looked into the highest accuracy to be achieved in detecting each health-problem, also as an answer to **SRQ3.1**. Using all the techniques, as **Table 5** shows, the highest accuracy is over 99% and it has been obtained for the problems *Income, Interpersonal relationships and Communication with community resources.* The lowest accuracy achieved is of 67.2% for problem *Skin*, followed by the problem *Circulation* with an accuracy of 69.7%. Overall, the median accuracy obtained for the health-problems is 89.96%.

Table 5: Accuracy of prediction obtained per problem

Accuracy	**Problems**
>99	Income, Interpersonal relationships Communication with comunity resources
98 to 99	Social contact, Reproductive function
97 to 98	Care taking, Conciousness, Substance use
96 to 97	Sanitation, Health care supervision
94 to 96	Oral Health, Role Change, Infectious Condition
93 to 94	Residence, Hearing, Sleep
91 to 93	Digestion-hydratation
89 to 91	Vision
87 to 88	Respiration, Mental health, Physical activity
80 to 84	Urinary function, Bowel function, Pain Cognition, Nutrition
78 to 79	Neuro-muscolo-skeletal function
75 to 76	Personal Care, Medication Regimen
67 to 70	Skin, Circulation

,

7. VERIFICATION BY PROFESSIONAL CARE GIVERS

We conducted a survey with professional care givers in order to verify the correlations obtained by the selection algorithm. As the the pie charts depicted in **Figure 11** show, all the health care professionals could recognize the strong relationship between *Income, Care taking, Social Contact and Communication with community resources* (70.7% could entirely recognize them and the rest of 29.3% only to some extent). However, when it comes to the *Moderate* correlations, 46.3% of the respondents could entirely confirm the associations, while 42.7% could do it only to some extent. Notably, the *Low* correlations had more positive responses than the *Moderate* ones. Besides, being asked to what degree they agree with the correlations, they were also required to indicate which health-problems should be added to or removed from the associations. The answers of the health care professionals and the results of the survey can be found in *Appendix A*. Additionally, the health care professionals provided suggestions of correlations for the health-problems for which no significant correlations could be identified by the selection algorithm.

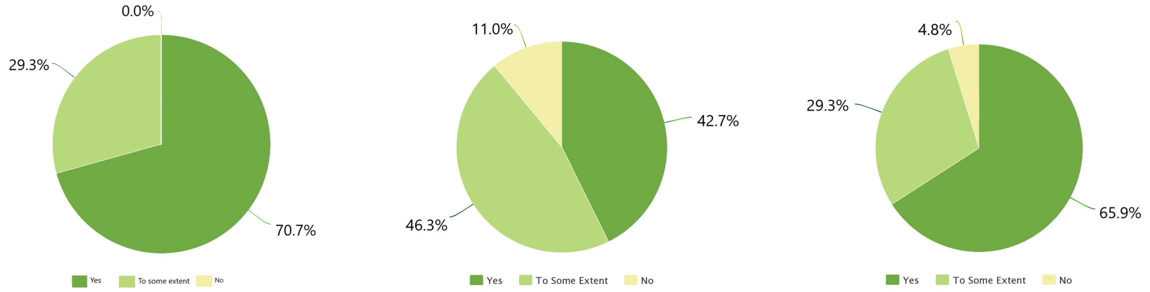


Figure 8: High Correlations

Figure 9: Moderate Correlations

Figure 10: Low Correlations

Figure 11: The confirmation of health care professionals of the identified correlations

8. DISCUSSION

First, this study showed that there exist relationship between health-problems. Additionally, we also found relationships between health-problems and the profile of patients. Using these relationships together with different data-dependent techniques, this study proved that it is possible to detect health-problems based on past assessments of the patients.

Looking back at the results obtained for **SRQ1**, the feature selection algorithm has identified high, moderate and low correlations between health-problems. On the domain level, we noticed that the environmental problems correlate with the behavioural and psycho-social ones. Also, the behavioural problems have some moderate correlations with the psycho-social. However, for the physiological problems, there are only some low correlations with the behavioral problems. Also, the responses of the professional care givers were positive in supporting the resulted correlations. All the professional care givers indicated that they could recognize the identified strong correlations. Likewise, for the moderate correlations, most of the health care professionals could recognize them, but only to some extent. This proves that the results obtained by the algorithm generally match with the responses of the experts in the domain.

Additionally, we also identified relationships between health-problems and the personal profile of a patient (**SRQ2**). As noticed, women are more likely to be diagnosed with *Neuro-muscolo-skeletal function* problems than men. The research performed by Regitz-Zagrosek [19] also proves that especially skeletal problems are more typical to women rather than men. Also, looking at the division of age per health-problem, the highest proportion of patients diagnosed with *Hearing and Vision* were between 80 and 100 years old. In the same direction, Jaul and Barron [12] confirm that people start to have sensory impairments at the age of 80.

For evaluating which data-technique is the most optimal and answering **SRQ3**, both accuracy and efficiency of the models have to be taken into consideration. While *Naive Bayes* is the fastest technique to apply, it also provides the lowest accuracy. On the other hand, *Gradient Boosted Tress* provides the best accuracy, however it is one of the slowest techniques. For example, a better balance between the duration of the training and the accuracy is achieved by Deep Learning, being the third performing technique in high accuracy with a difference of 0.15% and an average duration of 58 seconds.

Looking at the possibilities of detecting health-problems based on past assessments of the patients (our research question), 13 of the health-problems could be satisfactory (comparing with the related work) predicted with an accuracy of over 95%. The other health-problems have a lower accuracy which most likely can be increased by optimizing the classification model. A possible reason for these fluctuations in the accuracy could be the strength of the correlation between health-problems. It is interesting to note that for the problems with a high correlation such as *Care taking, Income* the accuracy is between 96 and 99%, while for the problems with insignificant correlations such as *Skin and Circulation* the accuracy drops under 70%.

8.1 Limitations

One limitation of the selection algorithm is that it could not find correlations for four of the health-problems. Furthermore, for determining the correlations between two health-problems, the algorithm is checking the amount of assessments in which both problems have been diagnosed. Considering this, it might be likely that those two problems have similar occurrences in the assessments, but it does not necessarily imply that they are somehow related to each other. On the other hand, comparing the results of the algorithm with the responses of the health care professionals proves that these limitations are minimal.

When it comes to the city of residence, one limitation might be the division of patients in the data set. As noticed, most of the health-problems seemed to be predominant in the North and East areas of the Netherlands. Also, most of the patients in our data set have a city of residence in these areas. Therefore, it is more difficult to draw conclusions when it comes to the influence of the region of residence over the health-problems.

9. CONCLUSION

The heavy workload pressuring the professional care givers make them bypass the development of new assessments which may affect the quality of care. As an alternative of the assessment process, this study proposed a model capable to detect health-problems based on the historical medical data of the patients. Correlations between health-problems have been discovered on different levels (high, moderate and low). Furthermore, the profile of a patient proved to have influence over prospective of health-problems. Nine different data techniques were used in combination with the aforementioned findings to detect health-problems based on past assessments of the patients. The resulted model accomplished to detect the health-problems with on overall accuracy of 89.96%, the accuracy ranging between 99.3% (for detecting Income problems) to 67.2% (for detecting Skin problems). For future work, the classification model should be improved for obtaining a higher accuracy.

10. FUTURE WORK

The classification model should be improved before it can be used in practice by health care professionals. One first step in improving the classification model would be to analyze the responses of the health care professionals and improve the correlations accordingly. However, a more comprehensive investigation should be performed in order to understand why the health care professionals do not fully agree with the results obtained by the algorithm and identify where the algorithm fails. Alternatively, a different selection method than the correlation-based method could be applied on the data set. This could improve the accuracy of problems for which no strong correlations have been found.

Another possibility for improving the classification model would be to integrate the notes made by the health care professionals during the assessment process. As the study performed by the Chen et al. [5] shows, the accuracy of a convolutional neural network which predicts chronic diseases based on the patient's medical file has been improved from roughly 50% to 94.5% by integrating the notes of the doctors. Integrating these notes requires more complex methods capable of natural language processing, but they might improve the accuracy by providing more insights in arguing why a patient has been diagnosed with a certain problem.

11. ACKNOWLEDGEMENT

First of all, I would like to thank my supervisor, Nirvana Meratnia for guiding me during this research, for making me understand how a research should be performed, for the time and the constructive feedback. Secondly, I would like to thank Ecare Services for providing the data set and facilitating a place to perform my research.

12. REFERENCES

[1] Qlik sense:data analytics. https://www.qlik.com/us/products/qlik-sense.

[2] B. Ahmad, K. Khairatul, and A. Farnaza. An assessment of patient waiting and consultation time in a primary healthcare clinic. *Malaysian Family Physician: the Official Journal of the Academy of Family Physicians of Malaysia*, 12(1), Apr 2017.

[3] J. E. Ball, T. Murrells, A. M. Rafferty, E. Morrow, and P. Griffiths. 'Care left undone' during nursing shifts: associations with workload and perceived quality of care. *BMJ Quality & Safety*, 23(2):116–125, Feb. 2014.

[4] M. Chand and R. L. Tung. The Aging of the World's Population and Its Effects on Global Business. *Academy of Management Perspectives*, 28(4):409–429, Nov. 2014.

[5] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang. Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. *IEEE Access*, 5:8869–8879, 2017.

[6] D. A. Etzioni, J. H. Liu, M. A. Maggard, and C. Y. Ko. The aging population and Its Impact on the Surgery Workforce. *Annals of Surgery*, 238(2):170–177, Aug. 2003.

[7] L. Fagerstrom, M. Kinnunen, and J. Saarela. Nursing workload, patient safety incidents and mortality: an observational study from Finland. *BMJ Open*, 8(4), Apr. 2018.

[8] M. A. Hall. Correlation-based Feature Selection for Machine Learning. page 198, Mar 1999.

[9] W. L. Haskell. Cardiovascular disease prevention and lifestyle interventions: Effectiveness and efficacy. *Journal of Cardiovascular Nursing*, 18(4):245, Oct 2003.

[10] K. Hayrinen, K. Saranto, and P. Nykanen. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International Journal of Medical Informatics*, 77(5):291–304, May 2008.

[11] D. E. Hinkle, W. Wiersma, and S. G. Jurs. *Applied statistics for the behavioral sciences*. Boston Houghton Mifflin, 2nd edition, 1988.

[12] E. Jaul and J. Barron. Age-Related Diseases and Clinical and Public Health Implications for the 85 Years Old and Over Population. *Frontiers in Public Health*, 5, Dec. 2017.

[13] S. Jyoti, A. Ujma, S. Dipesh, and S. B. Sunita. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. volume 17, Mar 2011.

[14] G. Katz, E. C. R. Shin, and D. Song. Explorekit: Automatic feature generation and selection. In *IEEE 16th International Conference on Data Mining (ICDM)*, Dec 2016.

[15] K. S. Martin. *The Omaha System: A Key to Practice, Documentation, and Information Management*. Health Connections Press, 2nd edition, 2009.

[16] I. Mierswa and R. Klinkenberg. Rapidminer studio. https://docs.rapidminer.com/, Dec 2018.

[17] C. Pattichis, E. Kyriacou, S. Voskarides, M. Pattichis, R. Istepanian, and C. Schizas. Wireless telemedicine systems: an overview. *IEEE Antennas and Propagation Magazine*, 44(2):143–153, Apr. 2002.

[18] B. Pogorelc, Z. Bosnic, and M. Gams. Automatic recognition of gait-related health problems in the elderly using machine learning. *Multimedia Tools and Applications*, 58(2):333–354, May 2012.

[19] V. Regitz-Zagrosek. Sex and gender differences in health. *EMBO Reports*, 13(7):596–603, July 2012.

[20] M. Solanki. 130.000 jobs available in the Dutch healthcare sector. https://www.iamexpat.nl/career/employment-news/130000-jobs-available-dutch-healthcare-sector, Mar 2018.

[21] I. Stuart-Hamilton. *An Introduction to Gerontology*. Cambridge University Press, Mar 2011.

[22] B. S. Yaochu Jin. Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Journals*, Apr 2008.

[23] Y. Ye, Q. Wu, J. Zhexue Huang, M. K. Ng, and X. Li. Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition*, 46(3), Mar 2013.

APPENDIX

A. SURVEY

In this section, a part of the survey conducted with healthcare professionals and the results are presented. It is importance to notice the structure of the survey. The first part is about the identified symmetrical correlations, the second one about the identified unsymmetrical correlations and the third part is about the problems for which no correlations have been identified.

Part 1

During its prediction process, our algorithm identified several correlations between problems which we would like to be assessed by you.

Statement 1

The algorithm that we developed found a strong correlation between the problems *Social Contact, Interpersonal Relationship, Income, Communication with community resources and Caretaking.*

Question 1.1: Do you agree with the proposed association?

- Yes (69%)
- To some extent (31%)
- No (0%)

Question 1.1.1 (Only for users whose response was not "Yes" for Question 1.1): If you do not fully agree with the proposed association, in your opinion which problem does not belong in the association: (multiple choices are possible)

- Social Contact (0%)
- Interpersonal Relationship (6.7%)
- Income (73.3%)
- Communication with community resources (20%)
- Caretaking (46.7%)

Question 1.2: Do you think that there are problems missing in the proposed association. Which ones?

- No (88.5%)
- Yes
  - Pain (2.3%)
  - Residence (2.3%)
  - Health care supervision (6.9%)

Statement 2

The algorithm that we developed found a strong correlation between the problems *Residence, Sleep, Physical Activity, Role Change and Health care supervision.*

Question 2.1: Do you agree with the proposed association?

- Yes (60.5%)
- To some extent (34.9%)
- No (4.7%)

Question 2.1.1 (Only for users whose response was not "Yes" for Question 2.1): If you do not fully agree with the proposed association, in your opinion which problem does not belong in the association: (multiple choices are possible)

- Residence (33.3%)
- Sleep (3.7%)
- Physical activity (3.7%)
- Role Change (66.7%)
- Health care supervision (40.7%)

Question 2.2: Do you think that there are problems missing in the proposed association. Which ones?

- No (79.1%)
- Yes
  - Neuro-muskolo-skeleto function(2.3%)
  - Pijn (2.3%)
  - Cognition (2.3%)
  - Nutrition (2.3%)
  - Other (11.7%)

Part 2

Question 1: In your opinion, which problems are highly correlated with *Sanitation*: (multiple choices are possible)

- Health care supervision (48.8%)
- Communicable/Infectious condition (74.4%)
- Role Change (25.6%)
- Others, namely:
  - Sleep (27.9%)
  - Digestion and Hydration (20.9 %)
  - Residence (90.7%)

Question 2: In your opinion, which problems are highly correlated with *Pain*: (multiple choices are possible)

- Sleep and rest patrons (74.4%)
- Communicable/Infectious condition (34.9%)
- Bowel function (48.8%)
  - Physical activity (81.4%)
  - Residence (11.6%)
  - Health care supervision (48.8%)
  - Neuro-muskolo-skeleto function (93.0%)

Part 3

For some of the problems, our algorithm could not find relevant correlations with other problems, therefore the accuracy of predicting them is somewhat lower. The next questions are related to these problems trying to find these missing similarities based on your expertise.

Question 1: In your opinion, are there problems which correlate/cause *Skin* problems?

- I cannot associate Skin with any other problems (7%)
- Yes, namely:
  - Oral health (23.3%)
  - Infectious condition (72.1%)
  - Circulation (86%)
  - Bowel function (23.3%)
  - Others (13.8%)

B. PROBLEMS DEFINED BY THE OMAHA SYSTEM

Table 6: Health Problems identified by the Omaha System

Domain	Problem Name	Target of the problem
Environmental	Income	Available money sources for living and health care expenses
	Sanitation	Environmental cleanliness against infection and diseases
	Residence	Living Area
Psycho-social	Communication with community resources	Interaction between the individual and community and social service organizations
	Social Contact	Interaction between the individual and others outside the immediate living area
	Role Change	Additions to or removal of a set of expected behavioral aspects
	Interpersonal Relationships	Associations or bonds between the individual and others
	Mental health	Development and use of mental/emotional abilities to adjust life situations, interactions with others and engage in activities
	Care taking/Parenting	Providing support, stimulation and physical care for dependent child or adult.
Physiological	Hearing	Perception of sound by the ears
	Vision	Act or power of sensing with the eyes
	Oral health	Condition of the mouth and gums and the number, type and arrangement of the teeth
	Cognition	Ability to think and use information
	Pain	Unpleasant sensory and emotional experience associated with actual or potential tissue damage
	Consciousness	Awareness of and responsiveness to stimuli and the surroundings
	Skin	Natural covering of the body
	Neuro-muscolo-skeletal function	Ability of nerves, muscles and bones to perform or coordinate specific movement, sensation or regulation
	Respiration	Inhaling and exhaling air into the body and exchanging oxygen
	Circulation	Pumping blood in adequate amounts and pressure throughout the body
	Digestion-hydratation	Process of converting the food into forms that can be absorbed andassimilated and maintaining fluid balance
	Bowel function	Transporting food through the gastro-intestinal tract to eliminate wastes
	Urinary function	Production and excretion of urine
	Reproductive function	Condition of genital organs and breasts and the ability to reproduce
	Communicable/infectious condition	State in which organisms invade/infest and produce superficial or systematic illness with the potential for spreading and transmission
Health-Related Behaviors	Nutrition	Select, consume, and use food and fluids for energy, maintenance, growth, and health
	Sleep and rest patterns	Periods of suspended motor and sensory activity and periods of inactivity
	Physical activity	State or quality of body movements during daily living
	Personal care	Management of personal cleanliness and dressing
	Substance use	Consumption of medicines, recreational drugs, or materials likely to cause mood changes and/or psychological/physical dependence, illness
	Health care supervision	Management of health care treatment plan by health care providers
	Medication regimen	Use or application of over-the-counter and prescribed medications