Medical Decision Support Using Machine Learning

Daniëlle Hooijenga

▶ **To cite this version:**

Daniëlle Hooijenga. Medical Decision Support Using Machine Learning. Other. Université de Lyon, 2020. English. NNT : 2020LYSEM029 . tel-03358245

HAL Id: tel-03358245
https://theses.hal.science/tel-03358245

Submitted on 29 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre NNT: 2020LYSEM029

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de

École des Mines de Saint-Étienne

École Doctorale N° 488

(Sciences, Ingénierie, Santé)

Spécialité de doctorat: Génie Industriel

Discipline: Machine Learning

Soutenue publiquement le 11/12/2020, par:

Daniëlle Hooijenga

Medical Decision Support Using Machine Learning

Devant le jury composé de:

VERDIER, Christine	Prof	Université Grenoble Alpes	Présidente du Jury
JOURDAN, Laetitia	Prof	Université de Lille	Rapporteure
GONZALEZ, Javier	Prof	Universidad del Rosario, Bogotá	Rapporteur
VERDIER, Christine	Prof	Université Grenoble Alpes	Examinatrice
AUGUSTO, Vincent	Prof	Mines Saint-Etienne	Directeur de thèse
XIE, Xiaolan	Prof	Mines Saint-Etienne	Co-directeur de thèse
PHAN, Raksmey	Dr	Mines Saint-Etienne	Co-encadrant de thèse
HEUDEL, Pierre-Étienne	Dr	Centre Léon Bérard, Lyon	Membre invité
SARAZIN, Marianne	Dr	Clinique Mutualiste Chirurgicale	Membre invité

Spécialités doctorales	Responsables :	Spécialités doctorales	Responsables
SCIENCES ET GENIE DES MATERIAUX	K. Wolski Directeur de recherche	MATHEMATIQUES APPLIQUEES	O. Roustant, Maître-assistant
MECANIQUE ET INGENIERIE	S. Drapier, professeur	INFORMATIQUE	O. Boissier, Professeur
GENIE DES PROCEDES	F. Gruy, Maître de recherche	SCIENCES DES IMAGES ET DES FORMES	JC. Pinoli, Professeur
SCIENCES DE LA TERRE	B. Guy, Directeur de recherche	GENIE INDUSTRIEL	N. Absi, Maitre de recherche
SCIENCES ET GENIE DE L'ENVIRONNEMENT	D. Graillot, Directeur de recherche	MICROELECTRONIQUE	Ph. Lalevée, Professeur

EMSE : Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'État ou d'une HDR)

ABSI	Nabil	MR	Génie industriel	CMP
AUGUSTO	Vincent	CR	Image, Vision, Signal	CIS
AVRIL	Stéphane	PR2	Mécanique et ingénierie	CIS
BADEL	Pierre	MA(MDC)	Mécanique et ingénierie	CIS
BALBO	Flavien	PR2	Informatique	FAYOL
BASSEREAU	Jean-François	PR	Sciences et génie des matériaux	SMS
BATTON-HUBERT	Mireille	PR2	Sciences et génie de l'environnement	FAYOL
BEIGBEDER	Michel	MA(MDC)	Informatique	FAYOL
BLAYAC	Sylvain	MA(MDC)	Microélectronique	CMP
BOISSIER	Olivier	PR1	Informatique	FAYOL
BONNEFOY	Olivier	PR	Génie des Procédés	SPIN
BORBELY	Andras	MR(DR2)	Sciences et génie des matériaux	SMS
BOUCHER	Xavier	PR2	Génie Industriel	FAYOL
BRODHAG	Christian	DR	Sciences et génie de l'environnement	FAYOL
BRUCHON	Julien	MA(MDC)	Mécanique et ingénierie	SMS
CAMEIRAO	Ana	MA(MDC)	Génie des Procédés	SPIN
CHRISTIEN	Frédéric	PR	Science et génie des matériaux	SMS
DAUZERE-PERES	Stéphane	PR1	Génie Industriel	CMP
DEBAYLE	Johan	MR	Sciences des Images et des Formes	SPIN
DEGEORGE	Jean-Michel	MA(MDC)	Génie industriel	Fayol
DELAFOSSE	David	PR0	Sciences et génie des matériaux	SMS
DELORME	Xavier	MA(MDC)	Génie industriel	FAYOL
DESRAYAUD	Christophe	PR1	Mécanique et ingénierie	SMS
DJENIZIAN	Thierry	PR	Science et génie des matériaux	CMP
BERGER-DOUCE	Sandrine	PR1	Sciences de gestion	FAYOL
DRAPIER	Sylvain	PR1	Mécanique et ingénierie	SMS
DUTERTRE	Jean-Max	MA(MDC)		CMP
EL MRABET	Nadia	MA(MDC)		CMP
FAUCHEU	Jenny	MA(MDC)	Sciences et génie des matériaux	SMS
FAVERGEON	Loïc	CR	Génie des Procédés	SPIN
FEILLET	Dominique	PR1	Génie Industriel	CMP
FOREST	Valérie	MA(MDC)	Génie des Procédés	CIS
FRACZKIEWICZ	Anna	DR	Sciences et génie des matériaux	SMS
GARCIA	Daniel	MR(DR2)	Sciences de la Terre	SPIN
GAVET	Yann	MA(MDC)	Sciences des Images et des Formes	SPIN
GERINGER	Jean	MA(MDC)	Sciences et génie des matériaux	CIS
GOEURIOT	Dominique	DR	Sciences et génie des matériaux	SMS
GONDRAN	Natacha	MA(MDC)	Sciences et génie de l'environnement	FAYOL
GONZALEZ FELIU	Jesus	MA(MDC)	Sciences économiques	FAYOL
GRAILLOT	Didier	DR	Sciences et génie de l'environnement	SPIN
GROSSEAU	Philippe	DR	Génie des Procédés	SPIN
GRUY	Frédéric	PR1	Génie des Procédés	SPIN
HAN	Woo-Suck	MR	Mécanique et ingénierie	SMS
HERRI	Jean Michel	PR1	Génie des Procédés	SPIN
KERMOUCHE	Guillaume	PR2	Mécanique et Ingénierie	SMS
KLOCKER	Helmut	DR	Sciences et génie des matériaux	SMS
LAFOREST	Valérie	MR(DR2)	Sciences et génie de l'environnement	FAYOL
LERICHE	Rodolphe	CR	Mécanique et ingénierie	FAYOL
MALLIARAS	Georges	PR1	Microélectronique	CMP
MOLIMARD	Jérôme	PR2	Mécanique et ingénierie	CIS
MOUTTE	Jacques	CR	Génie des Procédés	SPIN
NAVARRO	Laurent	CR		CIS
NEUBERT	Gilles			FAYOL
NIKOLOVSKI	Jean-Pierre	Ingénieur de recherche	Mécanique et ingénierie	CMP
NORTIER	Patrice	PR1	Génie des Procédés	SPIN
O CONNOR	Rodney Philip	MA(MDC)	Microélectronique	CMP
PICARD	Gauthier	MA(MDC)	Informatique	FAYOL
PINOLI	Jean Charles	PR0	Sciences des Images et des Formes	SPIN
POURCHEZ	Jérémy	MR	Génie des Procédés	CIS
ROUSSY	Agnès	MA(MDC)	Microélectronique	CMP
ROUSTANT	Olivier	MA(MDC)	Mathématiques appliquées	FAYOL
SANAUR	Sébastien	MA(MDC)	Microélectronique	CMP
SERRIS	Eric	IRD		FAYOL
STOLARZ	Jacques	CR	Sciences et génie des matériaux	SMS
TRIA	Assia	Ingénieur de recherche	Microélectronique	CMP
VALDIVIESO	François	PR2	Sciences et génie des matériaux	SMS
VIRICELLE	Jean Paul	DR	Génie des Procédés	SPIN
WOLSKI	Krzystof	DR	Sciences et génie des matériaux	SMS
XIE	Xiaolan	PR0	Génie industriel	CIS
YUGMA	Gallian	CR	Génie industriel	CMP

Mise à jour : 30/08/2018

Medical Decision Support Using Machine Learning

Daniëlle Hooijenga

11/12/2020

Abstract

Abstract

The main objective of this thesis is to develop innovative machine learning techniques to aid in medical decision making using available health databases such as PMSI, and local databases of the CLB (Center Léon Bérard, Lyon, France). In the thesis, several case studies were examined, concerning readmission to the emergency room, readmission to the hospital and decision support for the treatment of breast cancer.

Firstly, we consider a case study concerning emergency department readmission prediction. Readmission to the emergency department may be a sign of insufficient treatment at the first visit of the patient. For this goal, we combine a classification model (DAMIP) with tabu search for feature selection. Moreover, we combine this with a sampling method for speeding up the computations. This method has been shown to give slightly better results than conventional machine learning methods found in literature when applied to the emergency department readmission data set. However, the results for all the tested methods were not very satisfying, which was probably due to a lack of information in the data. The data consisted mainly of administrative data about the patient. Because of this, we considered another case study concerning hospital readmission. This data set contains a large number of features, including diagnoses and medical acts. On this data set the results were significantly better and we again showed that our Tabu/DAMIP framework outperforms other methods.

After, we developed a method that combines an autoencoder for dimensionality reduction with DAMIP for classification. In this method, we make use of a discretization method in order to be able to combine the two parts. This method was tested on CLB breast cancer treatment data. The goal of the case study is to be able to aid the clinician in making a decision on the treatment of an elderly breast cancer patient. We aim to do this by several approaches. Either we try to predict 5-year survival given the treatment the patient receives, or we try to predict whether a patient needs a treatment or not, given that the patient survives at least five years. The results are similar to those of the Tabu/DAMIP framework, but they are obtained much faster by using an autoencoder. Besides, we also combined the autoencoder with other classification algorithms, where the best result was obtained by the autoencoder with linear discriminant analysis.

Finally, we developed a simulation model to show the impact of our methods when used in a real application regarding hospital readmission. In this model, we apply the Tabu/DAMIP framework for prediction on whether a patient will return or not. If we predict that a patient is likely to be readmitted, we make a new prediction with an extended length of stay. Based on our predictions a decision on an extended stay is made. The goal of this approach is to reduce the number of readmissions. The results show that we can indeed manage to decrease the number of readmissions, even though in this case the cost may increase.

Résumé

L'objectif principal de cette thèse consiste à développer des techniques innovantes d'apprentissage automatique pour aider à la décision médicale à l'aide des bases de données de santé disponibles telles que PMSI, et des bases de données locales du CLB (Centre Léon Bérard, Lyon, France). Dans la thèse, plusieurs études de cas ont été examinées, concernant la réadmission aux urgences, la réadmission à l'hôpital et l'aide à la décision pour le traitement du cancer du sein.

Tout d'abord, une combinaison de DAMIP avec la recherche tabou a été développée, avec DAMIP pour la classification et la recherche tabou pour la sélection des variables. Il a été démontré que cette méthode donne de meilleurs résultats que les méthodes classiques d'apprentissage automatique. Ces résultats ont été obtenus sur les données de réadmission à l'hôpital. Ensuite, nous avons développé une méthode qui combine un autoencodeur pour la réduction de dimensionnalité avec DAMIP pour la classification. Cette méthode a été testée sur les données de traitement du cancer du sein au CLB. Les résultats sont similaires à la méthode précédente, mais sont obtenus beaucoup plus rapidement. Nous avons aussi combiné l'autoencodeur avec d'autres algorithmes de classification : le meilleur résultat a été obtenu par l'autoencodeur avec une analyse discriminante linéaire.

Enfin, nous avons développé un modèle de simulation afin d'évaluer l'impact de nos méthodes lorsqu'elles sont utilisées dans une application réelle de réadmission à l'hôpital. Dans ce modèle, nous appliquons la méthode Tabu / DAMIP pour prédire si un patient reviendra ou non. Si nous prédisons qu'un patient est susceptible d'être réadmis, nous faisons une nouvelle prédiction avec une durée de séjour prolongée. Sur la base de nos prévisions, une décision sur un séjour prolongé est prise. Le but de cette approche est de réduire le nombre de réadmissions. Les résultats montrent que l'on parvient effectivement à diminuer le nombre de réadmissions, même si dans ce cas le coût peut augmenter.

Acknowledgements

Writing the acknowledgements means that I have reached the point where my PhD is coming to an end. I definitely have some people to thank for being part of my journey.

First of all, I would like to thank all my supervisors, Vincent Augusto, Xiaolan Xie, and Raksmey Phan, for their guidance, their ideas, their time, and the opportunities they gave me.

Besides, I'd like to thank the practitioners I worked with, specifically Marianne Sarazin and Pierre Heudel for their interesting insights and their endless patience in explaining.

I'd like to thank the colleagues of CIS for the great atmosphere to work in, with a special thanks to the entire I4S team. Without them the three years wouldn't have been the same.

Furthermore, I would like to thank Omar, Vinicius, Solmaz, Jamal, Asma, Nilson, Cyriac, Ozge, Zeineb, Laure, Jay, and anyone I might be forgetting. They helped me make my stay in France a good and memorable experience.

Lastly, a special thanks to my parents and my brother, who have always supported me in all my decisions. They have always encouraged me to choose my own path and they have always supported along the way.

Contents

List of Figures

List of Tables

Introduction

The availability of health care data has been growing exponentially in recent times. Similar to other fields, digitalization made its appearance in healthcare, advancing the collection and storage of healthcare related data. A large variety of data can be found in healthcare. Data may consist of images, signals, administrative data, biological data, free text, and so on.

With the growing availability of data, the field of data-driven approaches in healthcare also continues to expand. The huge amounts of data may be overwhelming and hard to analyze for physicians. Using machine learning methods, patterns might be found of which physicians are unaware. Machine learning methods in the healthcare domain are generally used to make classifications, which can be used as predictions. If the outcome of, for example, a treatment can be predicted with a certain level of certainty, this could help the physician in making a better informed decision on the next step in the treatment of a patient. Eventually, this can lead to an improved clinical pathway for the patient.

In this thesis we will consider data-driven approaches in healthcare. We consider multiple data-driven approaches with the goal of providing medical decision aid to practitioners. If better-informed decisions can be made, this will lead to improvements in patient care and in the overall healthcare system. We will look at several data sets concerning healthcare, which all carry different characteristics, ranging from only containing administrative data, to containing many different kinds of data from different sources. Besides different data sets, we also consider different methods. We look how different classification methods perform on the different data sets and we consider several methods to reduce the dimensionality of our data sets. Moreover, we provide insight in the effect of applying our method a in real-life situation. The objectives of this thesis are described below.

Scientific objectives

The main objective of this thesis is to develop a methodology, based on machine learning and mathematical models, which can be used in the field of clinical decision aid. The resulting model should be able to be used in helping clinicians make their decisions. Besides, an objective is to verify the impact of the developed models. More specifically, those objective can be split up in sub-objectives:

— **Provide a framework combining feature selection with classification for medical decision aid.** The goal of the framework is to support physicians in their decision making. For this purpose we have to be able to predict what will happen to a specific patient. However, for making this classification we also need to be able to select the most relevant subset of information from a data set, as not all variable may be relevant.

— **Establish a new framework consisting of dimension reduction and classification.** Instead of selecting a subset of features, we can also focus on using all features in a linear combination before doing the classification we are interested in. This approach might better capture all information we have in a data set.

— **Provide decision aid to physicians using our established frameworks.** Once a framework is established, it should be easily usable by physicians in order to aid them in making their decisions. This implies that the whole framework should be automated and the results easily understandable.

— **Give insight in the impact of medical decision aid.** Once a framework for classification has been established, we are interested to see how this would impact a healthcare system in a real-life situation. For this goal, we will make a simulation in which we impact certain decisions made in the process based on our classification model.

Outline

This thesis consists of four chapters.

In Chapter 1 an overview of existing literature is given. We look at different aspects such as data-driven approaches and machine learning algorithms in healthcare, but also at limitations and open challenges. The overview of literature given in this chapter is highly general, as each chapter also has its own specific literature review.

Chapter 2 focuses on the problem of hospital readmissions. The goal is to predict those readmissions, such that measures can be taken in order to avoid readmission. For this purpose, we propose a classification model in combination with a local search heuristic for feature selection. The developed framework is tested on two different data sets, where one consists mainly of administrative data, whereas in the other data set a lot of information on diagnosis and treatment is known.

Chapter 3 concerns breast cancer treatment decision aid for elderly patients. In this chapter we look at the option of dimensionality reduction by using an autoencoder. Instead of choosing a subset of features, in this chapter, we aim to reduce the dimensionality of the data by creating a linear combination of all features present. This is applied to a data set concerning breast cancer patients. This data set has a wide variety of data available, including administrative data, biological data, and treatment data. We try out different approaches with the goal of providing decision aid.

In Chapter 4 we investigate the effect on hospital readmission if we use our developed methodology from Chapter 2 in a real-life scenario. For this we created a simulation model, where we consider patients coming to the hospital, staying for a specified amount of time and then based on our prediction they either stay longer in the hospital or they are discharged.

Chapter I

Literature Review

Contents of the chapter

Abstract of the chapter

In this chapter we look at the existing literature on the topic of data-driven methods in healthcare. As more and more data is generated in hospitals and other healthcare institutions, the need for data-driven methods is also growing.

Firstly, we look at different types of data which are present in healthcare. Data can come in many forms such as administrative data, free text, biological data, and so on. We distinguish between structured and unstructured data. We look at what has been done on both types of data in a healthcare context.

After, we take a look at different machine learning methods which are applied to healthcare problems in literature. We present an overview of which methods have been applied and to what kind of problems.

Finally, we take a look at some limitations and open challenges on the topic of big data and data-driven methods in healthcare. Those limitations include imbalanced data, limitations in the data, interpretability, and feature selection.

This chapter gives a very general overview of the literature. In each of the following chapter a more specific literature review relevant to the topic of the chapter is presented.

I.1 Introduction

Nowadays, more and more data are being recorded in hospitals, care centers and other healthcare organizations due to advances in data collection methods such as physiological monitoring data and insurance claims data [Adibuzzaman et al., 2017].

With the growth of quantity of data in healthcare, analysis and machine learning algorithms can help in early disease detection, patient care, and community services [M. Chen et al., 2017]. Different approaches of prediction in healthcare have been investigated and different methods have been tested.

Many different types of data can be found in healthcare. Data can be in the form of images, signal data, free text, biological data, administrative data, and so on. The first part of the literature review shows what has been done on several types of data. This shows the variety in data sources possible.

The big amounts of data generated allow for different data-driven approaches including machine learning. The second part of this literature review shows an overview of different machine learning methods which have been tested on healthcare data.

In the last part of the literature review the focus is on some challenges and limitations in the field of data-driven methods in the healthcare domain. Examples of challenges include the quality of data, interpretability of machine learning methods, imbalanced data, and the choice of the most relevant information within a dataset.

In this literature review a general overview of existing literature on machine learning in healthcare is given. This overview is very general on purpose. In the following chapters in this thesis, a more specific literature review is given in each technical chapter.

I.2 Big data in healthcare

Typically, in healthcare large amounts of data are generated, driven by record keeping, regulatory requirements, and patient care [Kudyba, 2010]. According to the authors of [Y. Wang et al., 2018] the healthcare industry does not yet leverage the full potential of the benefits which can be gained from big data analytics. The authors of [Yang et al., 2017] mention that physicians nowadays can be overwhelmed by the huge availability of data sources and that machine learning models can be capable of handling such data in order to help physicians make their decisions. In this section some approaches that were developed and tested in the field of big data applied to healthcare are discussed.

Different types of data can be distinguished that are typically found in the healthcare domain. The authors of [Kamesh et al., 2015] make the distinction between four levels of data in healthcare: molecular level, tissue level, patient level, and population level. In this thesis, the focus will be on patient level data. At this level we can distinguish multiple forms, of which the two principal forms are structured and unstructured data. Structured data follows some kind of standard, such as, for example, ICD-10 codes for medical diagnoses. Unstructured data on the other

hand includes free text, images, and signals. In Table I.1 an overview is given of articles which use data-driven approaches applied to healthcare problems. They are ordered by the type of data that is used. When the authors use both structured and unstructured data, the main form used is chosen. After, for each article a short description is given.

Type of data	**References**
Structured data	[Finkelstein and Jeong, 2017], [Gilbank et al., 2019], [Baig et al., 2016], [Valdes et al., 2017], [Mylona et al., 2019], [Ambale-Venkatesh et al., 2017], [Thottakkara et al., 2016], [Hussain and Junejo, 2019], [Bardhan et al., 2011], [Feng et al., 2017], [L. Wang et al., 2016], [Ross et al., 2008], [Eswari et al., 2015], [M. Chen et al., 2017], [Downing et al., 2017], [Swain, 2016], [Ow and Kuznetsov, 2016], [Koppad and Kumar, 2016], [H. Chen et al., 2013], [J. Hu et al., 2016]
Unstructured data	[Horng et al., 2017], [Lambin et al., 2013], [Amirian et al., 2017], [Forkan et al., 2017], [Szlosek and Ferrett, 2016], [C. Hu et al., 2016], [Shin and Markey, 2006], [El Naqa et al., 2006], [Ding et al., 2017], [Ali et al., 2019], [Geerts et al., 2016], [Jiang et al., 2014], [Khalifa and Meystre, 2015], [L. Wang et al., 2012], [Chang, 2018], [Lo'ai et al., 2016], [Jagadeeswari et al., 2018], [Abbas et al., 2016]

Table I.1 – Overview of references using structured and unstructured data

In [Finkelstein and Jeong, 2017] the authors consider predicting asthma exacerbations from self-monitoring of patients. Their dataset consists of just over 7000 patients who performed home telemonitoring. The authors applied a naive Bayesian classifier, an adaptive Bayesian network, and support vector machines to this problem and it was found that the adaptive Bayesian network gives the most promising results. The authors mention that this system can aid in personalized decision support for asthmatic diseases and that their method can be easily extended to other chronic health problems.

The goal of the authors in [Horng et al., 2017] is to show the usefulness of using free text data for identifying patients with an infection in the emergency department. The authors developed a method using a representation method for the free text and after support vector machines for prediction of an infection in a patient at the emergency department. The model is tested on a dataset covering five years of data containing 230936 patients. The results of the model using free text is then compared to using only vital sign and demographic data. The authors show that using free text in addition increases the area under the receiver characteristic curve value from 0.67 to 0.86, showing that the use of free text is valuable in predicting infections.

In [Lambin et al., 2013] an overview is given of existing methods for clinical decision-support systems based on prediction models of treatment outcome for radiation oncology. The authors focus on models combining predictive and prognostic data factors from clinical, imaging, molecular and other sources. The authors state that a truly useful predictive model should be continuously re-evaluated on different data sets from different regions.

The data used in [Amirian et al., 2017] comes from Point of Care (POC) devices, which are devices used to obtain diagnostic results while the doctor is with the patient. Such devices can be used for example to measure cholesterol levels. The authors perform thorough analysis using this data, without having any patient-specific information, making sure privacy is not a concern. The approach chosen by the authors is used to get a clear overview of the healthcare system and to identify high-risk populations. The authors mention this is useful for national health authorities in order to optimize resource allocation.

In [Forkan et al., 2017], big data produced from wearable and wireless sensors technology is studied. The authors consider vital signs such as heart rate and blood pressure, with the goal of identifying dangerous clinical events. The approach of the authors consist of a prognostic model where the idea is to compare the vital functions of a patient to historical data of similar patients. The proposed approach is shown to achieve a promising accuracy and it is mentioned that this implies that vital sign big data can be used to identify health problems ahead of time.

The authors of [Gilbank et al., 2019] focus on machine learning technologies for decision aid in the field of radiation oncology. The authors mention that decision aid tools may make mistakes that practitioners would not but that tools might also be able to find patterns which practitioners are unable to detect. In the article the authors focus on the viewing point of practitioners and develop an UI based on their feedback.

In [Szlosek and Ferrett, 2016] the focus is on evaluating the effectiveness of clinical decision support systems using electronic medical record data. The authors make use of natural language processing (NLP) and manual evaluation. The dataset consists of artificial patient records with free text.

In [C. Hu et al., 2016] a clinical decision support tool is developed for the early diagnosis of Alzheimer's disease. For this purpose the authors investigate the use of deep learning. The initial data consists of raw images, which are then converted into matrices, representing different regions of the brain. From those matrices, correlation matrices can be constructed which shows correlations between the different regions in the brain.

The aim in [Baig et al., 2016] is to offer support for identifying physiological events. The system the authors developed is self-organizing, consisting of preprocessing, clustering, and diagnosing. The clustering is done based on fuzzy logic modeling. The data used in the system consists of the vital signs of patients. The method was tested on 30 patient datasets. It is mentioned that the system can be useful but should be further tested in real-life situations.

The authors of [Shin and Markey, 2006] developed a clinical decision support system which can be used for early detection of cancer. This system makes use of

mass spectrometry, which gives insight in the levels of proteins, for identifying cancer biomarkers. It is mentioned that this data needs preprocessing as mass spectrometry data is typically noisy. After, methods for feature extraction and feature selection are used as the number of features is about 15000. The authors make use of an SVM framework for classification. It is argued that before the system can be deployed in real-life applications, studies need to be done on the accuracy of the method.

A clinical support system for treatment of lung cancers was developed in [Valdes et al., 2017]. The authors have used a dataset consisting of 104 patients who have been treated and for whom the outcome of the treatment is known. It is mentioned that those patient can be used to make the decision of treatment for future patients. This decision support system gives an advice to practitioners and patients about several different types of treatment and also an advice on the dose of the treatment is provided by the system.

The authors of [Mylona et al., 2019] consider the problem of prediction of urinary toxicity after radiotherapy for prostate cancer. Besides applying several machine learning algorithms, the authors combine those methods with oversampling techniques for imbalanced data. It was found that those oversampling techniques always improve the performance of the algorithm used, regardless of which algorithm was used. The best results were found using SMOTE for oversampling followed by a combination of Edited Nearest Neighbor algorithm and Regularized Discriminant Analysis.

The topic considered by the authors of [Ambale-Venkatesh et al., 2017] concerns the prediction of a cardiovascular event in a multi-ethnic group. The study concerns 6814 participants from different ethnicities, initially free of cardiovascular disease. The authors apply a so-called random survival forest to predict six different possible cardiovascular events in a time-span of twelve years. The proposed method is not only used for prediction, but also for recognizing the top-20 biomarkers based on which the predictions were made. In total the authors possessed 735 different variables. The authors state that their method outperforms other prediction methods.

In [Thottakkara et al., 2016] the authors try to predict postoperative sepsis and acute kidney injury. Several models are applied to this problem. Besides prediction with all available information, the authors also consider feature reduction techniques. It was shown that feature extraction by the use of principal component analysis results in an improved performance of the models.

An extensive systematic review on prediction of readmission to the hospital after heart failure is given in [Ross et al., 2008]. While one of the authors´ goals was to identify studies designed to compare hospital rates of readmission, no such model was identified. On the other hand, many studies about the prediction of hospital readmission were considered in which heterogeneous approaches were used and the predictive patient characteristics were found to vary widely.

In [El Naqa et al., 2006] the authors focus on prediction of radiotherapy outcome. In the paper, data sets of esophagitis and xerostomia are used. The authors propose a logistic regression framework for the approximation of the treatment-response function. The robustness of the performance is evaluated using Spearman's coefficient. Besides, the authors consider a bootstrap variable selection technique. It is

stated that this technique improves the model building. The authors conclude that the mentioned variable selection technique increases the reliability of the used models. Furthermore, the authors mention that the prediction of treatment response can be improved by combining clinical and dose-volume factors.

For patients with tuberculosis (TB) it is important that they finish their treatment. In [Hussain and Junejo, 2019] the authors explore a machine learning approach to predict whether a patient will finish his treatment or not. Based on results from literature, the authors decided to test three approaches for their problem: Random Forest, Support Vector Machine, and neural network. The used data set consists of the medical records of 4213 patients collected from their initial screening and registration to the end of TB treatment. Records of patients who died during treatment (due to external causes) were excluded from this study. The data contains 84 attributes, of which 52 were used after feature selection. The target variable is whether the treatment was completed or not, where 64.37% of patients in the data set completed their treatment and 35.62% did not. Feature selection was done with a two-step approach. In the first step, a chi-squared test is performed. All insignificant features were removed in this step. In the second step, for each individual feature, the authors learn a model. If the prediction accuracy is less than the probability of the more frequently occurring class, then the feature is removed. The idea behind this is that the same accuracy can be achieved by predicting every entity to be in the majority class. The authors achieve a prediction accuracy of approximately 76%. The best results are achieved by random forests.

In [Bardhan et al., 2011] the authors seek to predict preventable thirty-day readmissions. For this purpose, they develop a novel method called the beta geometric Erlang-2 (BG/EG) hurdle model. The goal of this model is to predict the propensity, frequency, and timing of readmissions of patients diagnosed with congestive heart failure. The model is tested on a data set consisting of patient demographic, clinical, and administrative data from 67 hospitals in North Texas over a four-year period. The authors state that the model can be used as a clinical decision support tool to identify high-risk patients. Specific variables that are significantly associated with readmission risk are health IT, patient demographics, visit characteristics, payer type, and hospital characteristics.

A system called CAC was developed by the authors of [Ding et al., 2017]. This system integrates Clustering, Association analysis, and Collaborative filtering to predict patients´ future conditions. All the conditions of several visits of a patient are combined to find frequent patterns. The patient records are clustered. Association analysis is used to find strong disease correlations among patients where each item represents a historical condition. Collaborative filtering is needed because in association analysis may overlook some rare diseases, while common diseases, like influenza, will be contained in frequent patterns. The data set used to test the method includes 151237 patients from a provincial capital city of China. For 71% of acute patients and 82% of chronic patients their future conditions were shown to be predictable.

In [Ali et al., 2019] voice recordings are used for prediction of Parkinson's disease. The authors propose a two dimensional data selection method for sample and feature

selection. This method makes use of a chi-square model to rank features. After, it searches for the optimal subset of those ranked features and iteratively selects samples. It is shown in the article that the proposed method outperforms other methods in terms of accuracy.

In [Feng et al., 2017] the authors propose a framework called the Intelligent Perioperative System (IPS). This is a real-time system which purpose is to indicate the risk of postoperative complications and to interact with physicians to improve predictive results. The developed framework consists of periodically collecting the EHR data of patients and after performs data integration, variable generation, surgical risk scores prediction, and risk scores visualization. The authors state this is the first of its kind and it would help physicians make data-driven decisions.

In [L. Wang et al., 2016] a review of feature selection methods is provided. The authors mention that data are becoming bigger, both in terms of instances and in terms of dimensionality of features. This growth can significantly degrade the accuracy and the efficiency of learning algorithms. Firstly, the authors mention that optimal feature selection is theoretically possible by means of exhaustive search. However, because of an exponential growth of possibilities this becomes practically impossible if there are more than 30 features. Proposed heuristic algorithms include a genetic algorithm (GA), ant colony optimization (ACO), particle swarm optimization (PSO), chaotic simulated annealing, tabu search, noisy chaotic simulated annealing, and branch-and-bound.

The authors of [Eswari et al., 2015] investigate the use of big data and data analysis for diabetic patients. The aim of the analysis is to predict the type of diabetes prevalent, possible complications, and the type of treatment a patient should receive. The authors mention that using predictive analytics on the available data patient care can be improved, being more affordable and available.

In [M. Chen et al., 2017] the problem of prediction of chronic disease outbreak in disease-frequent communities using big data is considered. The data is collected from central China and the authors mention that the data is incomplete, which can reduce the quality of analysis accuracy. For this purpose, a latent factor model is used to reconstruct the missing data. It is shown that the proposed convolutional neural network achieves high accuracy with a fast convergence speed.

The authors of [Downing et al., 2017] address the problem of hospital performance qualification. They mention that with the big amounts of data available and many resulting measures of quality, it is difficult to provide an overall characterization of hospital performance. The authors developed an approach with the goal of identifying similarities and differences between hospitals in order to describe common patterns of hospital performance. The conclusion of the authors is that their approach has revealed differences in performance that are hidden in existing systems of hospital rating.

The authors of [Geerts et al., 2016] consider big data concerning Alzheimer patients. They are interested in the development of analytical methods to advance clinical research and drug development. In the article, the authors mention the specific challenges in predictive analytics for Alzheimer's disease, such as the heterogeneity of the disease.

In [Jiang et al., 2014] the challenge of big data from wearable sensors is addressed. The data is collected from elderly people living an independent lifestyle in their own homes. The challenge the authors mention is to recognize human behaviour patterns for which they developed a hidden Markov model.

The authors of [Khalifa and Meystre, 2015] focus on big data from health records of diabetic patients. The challenge they consider is to use natural language processing (NLP) to identify health problems which may lead to cardiovascular problems such as high blood pressure, high cholesterol levels, obesity, and smoking status.

The goal of the authors in [Swain, 2016] is to find a profile of an adult population group being at risk of obesity. In this study, big data from healthcare is used in the United States, where obesity has a large impact on the costs of healthcare. The authors developed two predictive models which, is mentioned, can aid in early intervention strategies.

Big data can also be found in signals like EEGs. Analysis of this data can aid in detecting and diagnosing brain disorders. The authors of [L. Wang et al., 2012] aim to improve the performance of neural signal processing. Test results prove their approach to be efficient.

In [Chang, 2018], data analytics and visualization of cancerous tumors is discussed. The other developed a method which can inspect the status of malignant tumors and a simulation approach in which the tumor can be inspected in 360 degrees.

The authors of [Ow and Kuznetsov, 2016] consider the field of personalized medicine, which is mentioned to be growing due to big data becoming more and more important. Specifically, the authors consider patients with ovarian cancer. Using a prognostic method, the patients have been classified to three survival-significant risk groups. Besides, the authors combine their approach with classical machine learning techniques and they conclude that this approach, using a multi-test voting system, provides a more precise patient stratification.

In [Koppad and Kumar, 2016] the aim is to use big data for the benefit of Chronic Obstructive Pulmonary Disease (COPD). The authors apply a decision tree with the purpose of improving COPD diagnoses. The big data set used in this study contains many details about each individual patient including previous treatments. The method is found to be promising for its purpose.

The authors of [Lo'ai et al., 2016] look at a practical side of big data, namely the techniques and tools used to obtain the data from mobile devices. The authors mention that mobile devices are nowadays an indispensable part of everyday life and that this technology can be used to obtain data useful for the development of healthcare applications and systems. It is mentioned that the considered technologies can aid in advancing personalized medicine, reduced healthcare costs, and better clinical and operational processes.

Personalized healthcare is also a goal of the authors of [H. Chen et al., 2013]. Their focus is on diabetic patients. A system called DiabeticLink was developed, with the goal of address the needs of patients, caretakers, nurse educators, physicians, pharmaceutical companies, and researches. The authors make use of mining algorithms with the goal of healthcare decision support.

The authors of [J. Hu et al., 2016] focus on a concept called Learning Health Systems (LHS) which consists of learning from electronic healthcare data with the goal of improving healthcare quality as personalized decision support. To achieve this goal the authors provide visualization techniques which can be used to discover clusters of similar patients. With this information, more precise healthcare can be provided.

The study in [Jagadeeswari et al., 2018] concerns the use of big data stemming from Internet of Things (IoT). The authors look to use this data for early detection and diagnosis of diseases. They discuss how the data is collected and how this data can be used for several applications such as continuous monitoring of ICU patients, diagnosis of cardiac diseases, and prediction and prevention of Zika virus.

The authors of [Abbas et al., 2016] make use of data from social media for the purpose of identifying the risk of diseases. A cloud based framework was developed which makes use of a filtering approach. This approach searches for similarities between profiles of users. The risk assessment results were compared to classical machine learning approaches and the proposed method was shown to perform significantly better.

I.3 Machine learning in healthcare

As more data is available nowadays, the field of machine learning in healthcare is also expanding. Machine learning combines many different fields, including computer science, statistics, and optimization, in order to create methods for identifying patterns in data. Using those patterns, the understanding of a current situation may be better understood, or predictions about a future situation can be made [Wiens and Shenoy, 2018].

In this section we take a broad look at machine learning methods applied to healthcare problems. In Table I.2 an overview is given of different machine learning methods and relevant references. Below, a short description of each of those references is given in order to give insight in existing literature on machine learning in healthcare.

The authors in [Waljee et al., 2018] make use of logistic regression and random forest for the purpose of predicting the disease course of inflammatory bowel disease. Their method was tested on a dataset containing over 20000 patients. It was shown that random forest achieves the best results with an area under the receiver operating characteristic curve value of 0.85. The authors mention that this model is useful to distinguish between patients at high and low risk of a disease flare, which can aid practitioners in individualizing the treatment.

Like with many diseases, also for thyroid diseases early detection is crucial for proper treatment. In [Banu, 2016], a linear discriminant analysis classifier is used with the goal of early detection of thyroid disease. On a data set consisting of more than 3700 instances, it is shown that this approach achieves an impressive accuracy.

The authors of [Devinsky et al., 2016] explore a machine learning approach for the choice of the antiepileptic drug (AED) for individual patients. In their approach, the authors use machine learning methods to create a predictive algorithms which

Method	**References**
Logistic Regression	[Waljee et al., 2018], [Hall et al., 2016], [Parikh et al., 2019], [Goldstein et al., 2017], [Weng et al., 2017], [Thottakkara et al., 2016], [Yang et al., 2017], [Buettner et al., 2009], [Chu et al., 2008]
Random Forest	[Waljee et al., 2018], [Devinsky et al., 2016], [Hall et al., 2016], [Senders et al., 2018], [Parikh et al., 2019], [Weng et al., 2017], [Chu et al., 2008]
Decision Tree	[Hashi et al., 2017], [Hall et al., 2016], [Szlosek and Ferrett, 2016], [Senders et al., 2018], [Lynch et al., 2017], [Asri et al., 2016], [Goldstein et al., 2017]
Neural Network	[Hall et al., 2016], [C. Hu et al., 2016], [Senders et al., 2018], [Goldstein et al., 2017], [Yang et al., 2017], [Su et al., 2005], [Chu et al., 2008]
Naive Bayesian	[Hall et al., 2016], [Gultepe et al., 2014], [Asri et al., 2016], [Thottakkara et al., 2016], [Munley et al., 1999]
Support Vector Machine	[Hall et al., 2016], [Szlosek and Ferrett, 2016], [Gultepe et al., 2014], [Senders et al., 2018], [Lynch et al., 2017], [Asri et al., 2016], [Thottakkara et al., 2016], [Chu et al., 2008]
Nearest Shrunken Centroid	[Lusa et al., 2013], [Chu et al., 2008], [Khoshhali et al., 2015], [Thottakkara et al., 2016]
Linear Discriminant Analysis	[Senders et al., 2018], [Lee and Yang, 2016], [Peterson et al., 2008], [Banu, 2016], [Chu et al., 2008]
K-Nearest Neighbours	[Hashi et al., 2017], [Szlosek and Ferrett, 2016], [Garmendia et al., 2019]
Gaussian models	[Gultepe et al., 2014]
Markov models	[Gultepe et al., 2014]
Gradient Boosting	[Parikh et al., 2019], [Lynch et al., 2017], [Weng et al., 2017]
Linear Regression	[Lynch et al., 2017]

Table I.2 – Overview of references using different machine learning methods

estimates the probability of success for an individual patient and a specific treatment regimes. The AED regimen predicted by the model is the treatment with the highest success probability for the patient. The authors apply a basic method of feature selection, based on correlation between both the variables between each other as well as with the outcome variable. After a random forest approach is applied. There are large differences found between the model predicted treatment and the actual prescribed treatment, where the model predicted results show better results for patients: longer time to subsequent treatment modification and reductions in predicted health-care resource utilization (hospitalizations, AED use, specialist visits).

In [Lusa et al., 2013] an approach using Nearest Shrunken Centroid is proposed. The authors focus on problems in which the data is highly imbalanced, which is often the case in healthcare. The proposed methods shows good results on a breast cancer data set.

The authors of [Hashi et al., 2017] have developed a system which can be used by practitioners for the prediction of diagnosis of diabetic patients. In this system, decision tree and k-nearest neighbors are used for prediction. It is shown that the best results were achieved by decision tree with an accuracy of over 90%.

A clinical decison support system is proposed in [Hall et al., 2016]. The goal of this system is to make use of machine learning techniques in order to predict insulin resistance. For the prediction an ensemble classifier is used consisting of a stack of multilayer perceptron, decision tree, naive Bayes, SVM, logistic regression and random forest. The authors show to achieve an overall accuracy of 78% using their method and they mention that using their method, invasive blood testing can be avoided.

The goal in [Chu et al., 2008] is to develop a method which is capable of predicting and identifying patients with acute gastrointestinal bleeding, which is a highly unpredictable event. The authors apply several methods, including neural network, support vector machine, linear discriminant analysis, nearest shrunken centroid, random forest, and logistic regression. The authors state that using random forest excellent results were achieved.

In [Szlosek and Ferrett, 2016], the authors make use of natural language processing (NLP) and manual evaluation. The machine learning methods applied are support vector machines, decision tree, and k-nearest neighbors. The best results were found using the support vector machines method. The methods argue that using natural language processing in combination with machine learning techniques can be a useful application in electronic medical records.

The authors in [Peterson et al., 2008] are interested to find a relationship between burnout and self-reported physical and mental health factors. For this purpose a linear discriminant analysis approach is used. With this approach several indicators for burnout could be distinguished.

In [C. Hu et al., 2016] a clinical decision support tool is developed for the early diagnosis of Alzheimer's disease. For this purpose the authors investigate the use of deep learning. It is shown by the authors that their method achieves better classifications than more traditional methods and they mention that their developed

system can aid practitioners in recognizing Alzheimer's disease in an early stage in which case measures can be taken to slow down the process.

Decision support for identification of patients with a high risk of developing hyperlactatemia is the topic in [Gultepe et al., 2014]. The goal is to predict hyperlactatemia early on such that the clinical staff can respond to this and patient health can be positively impacted. In this study the authors make use of the electronic health records of 741 patients. Different classification methods were used, including naive Bayes, support vector machines, Gaussian models, and Markov models. It was shown that reliable predictions can be made using only three features, namely the median of the lactate levels, the mean arterial pressure, and the median absolute deviation of the respiratory rate.

The authors of [Senders et al., 2018] investigate the use of machine learning for medical decision aid concerning the outcome of neurosurgery. The decision aid system should be able to help in identifying patients who will benefit from surgery before the actual intervention. The authors perform a systematic review of the topic, for which thirty articles were found useful. It is mentioned that machine learning models can perform significantly better than existing methods. The authors therefore argue that such decision aid system can be very useful for practitioners but that more research is needed on how the machine learning methods should be implemented in a practical tool.

In [Parikh et al., 2019] 26525 adult cancer patients were studied. The goal of the authors is to predict six month mortality. The methods applied by the authors are random forest, gradient boosting, and logistic regression. The gradient boosting algorithm has also been applied in real time, where it was used to classify patients to be at high risk. The results of this real-life application were in line with the expectations of the practitioners.

The authors of [Khoshhali et al., 2015] use a nearest shrunken centroid approach to predict categories of colon cancer. It is mentioned that this method is very successful and that an accuracy of 97.7% is achieved.

Several machine learning techniques are applied in [Lynch et al., 2017]. The authors apply linear regression, decision tree, gradient boosting machines, support vector machines, and a custom ensemble to data concerning lung cancer patients. The goal using those techniques is to predict survival. The variable to be predicted is treated as a continuous variable, in order to give better insight in the probability of survival. It is shown that the predicted values are in accordance with actual values for short to moderate survival times. This concerns the major part of the data. The best model was found to be the gradient boosting machine, whereas decision tree was found to be inapplicable.

In [Asri et al., 2016] several data mining and classification algorithms are applied for the prediction of breast cancer risk and diagnosis. The authors compare Support Vector Machine (SVM), decision tree, Naive Bayes, and k-Nearest Neighbors. Those methods are applied on the well-known Wisconsin Breast Cancer dataset. It is shown that the best results are achieved using SVM when comparing by accuracy.

Cardiovascular health problems is the topic considered in [Goldstein et al., 2017]. The authors apply several methods, namely classification trees, a regression tech-

nique, neural networks, and nearest neighbours. Those methods are applied to health data in order to predict mortality after diagnosis of cardiovascular problems. The authors mention that for different goals, different methods should be used. Furthermore, they state that care is needed as machine learning methods are often a black box, which may raise concerns is healthcare applications.

The authors of [Weng et al., 2017] performed a study using a clinical data set concerning 378256 patients. The goal of the study is to assess machine learning techniques with the goal of identifying heart failure. The authors compare four machine learning methods: random forest, logistic regression, gradient boosting, and neural networks. Besides, they also compare the results of those methods to the results of an algorithm widely used in the United States. The best results were found using a neural network. In this case 355 more patients were recognized who developed heart failures as compared to the established algorithm.

In [Thottakkara et al., 2016] the authors try to predict postoperative sepsis and acute kidney injury. Several models are applied to this problem, namely logistic regression, generalized additive models, naive Bayesian, nearest shrunken centroid, and support vector machines. The different models are compared based on the area under the receiver operating characteristic curve, accuracy, and positive predicted value. It is found that logistic regression, generalized additive model, and support vector machines achieve better results than the naive Bayesian model, reach AUC scores of up to 0.858 for acute kidney injury and up to 0.909 for severe sepsis.

The authors of [Yang et al., 2017] make use of a combination of a recurrent neural network encoder and a multinomial hierarchical regression decoder for predictive modeling of treatment decision for metastatic breast cancer. The authors have shown that the proposed method outperforms more traditional approaches.

The authors of [Buettner et al., 2009] propose a framework which uses Bayesian logistic regression with high-order interactions for prediction of radiation-induces toxicities in cancer patients. The developed framework is shown to achieve area under the ROC curve scores of 0.72 and 0.64 for two different toxicities.

The goal of the authors of [Su et al., 2005] is to predict radiation-induced pneumonitis (RP). The authors propose an artificial neural network (ANN) for this objective. The used data set consists of clinical data from 142 patients, of which 26 with RP and 116 without RP. The input of the ANN was limited to the patient lung dose-volume data. Different training and testing procedures of the ANN are used for experimentation. The predictive accuracy was verified as the area under a receiver operator characteristic (ROC) curve. The authors consider their approach to be a useful tool for the prediction of RP. The best results are achieved using the ANN_ 1 method, which was trained and tested by using the leave-one-out method. In the leave-one-out method, one patient's data from the total data set is excluded to predict the network performance. The network was trained by the remaining data (n - 1) and tested by that excluded patient data. The process was repeated n times. This method maximizes the data available to train and test a predictive model for a limited data set.

The problem considered in [Munley et al., 1999] concerns the prediction of symptomatic lung injury. For this purpose the authors developed a nonlinear neural

network. As input the model takes pre-radiotherapy pulmonary function, three-dimensional treatment plan doses, and demographics. The model outputs a value between 0 and 1 indicating the likelihood that the concerned patient would become symptomatic. The model is trained on 97 patients, where the mean-squared error is minimized. The developed model reaches an area under the ROC curve of 0.833, which is shown to be higher than other methods.

The prediction of time until hospital readmission is studied in [Garmendia et al., 2019]. Their study is in the framework of survival analysis, which is generally used to study the time until death, however it can be generalized to time of readmission. The authors experimented using several neural and statistical prediction model. They find the approaches weighted k-NN and regression tree based rule system providing smooth approximations of the observed survival function. Besides, using an exploratory survival analysis with the use of Cox regression, the authors found that the age and motive of admission are the most significant variables in predicting the readmission. Especially fever, cough and certain kinds of abdominal pain are significant motives of admission.

The authors of [Lee and Yang, 2016] present a discriminant analysis approach to prediction. They mention the objective of this approach to be to derive rules that can be used to classify entities into groups. Many applications are mentioned including: identifying early predictive signatures of vaccine responses, early detection of mild cognitive impairment and Alzheimer's disease, and prediction of aberrant CpG island methylation in human cancer. The authors mention that their approach shows promising results both on real-life data as well as on simulated data.

I.4 Limitations and open challenges

So far in this chapter we have seen that many possibilities have arisen due to a higher availability of data in healthcare. Several machine learning algorithms have been applied to a variety of problems. However, there are also still some limitations and open challenges, of which a few are discussed in this section.

Imbalanced data

One challenge in healthcare data is that datasets tend to be imbalanced, the authors of [Razzaghi et al., 2019] experiment with different techniques for dealing with imbalanced data. The data set considered in this article concerns patients having had bariatric surgery, where patients having complications after surgery are a small minority. The proposed techniques are synthetic minority oversampling technique (SMOTE), random undersampling, random forest, bagging, and AdaBoost. Besides, the authors try to improve the classification performance by using feature selection techniques (chi-squared, information gain, and correlation-based feature selection). The different methods are tested on the most common complications, including diabetes, angina, heart failure, and stroke. It is shown that the ensemble learning-based classification techniques using any of the mentioned feature selection methods results in the best results.

The focus in [D. Dai and Hua, 2016] is on classification performance on imbalanced datasets. The authors mention that this is a common problem in healthcare, such as for example for rare disease classification. In the article, several experiments are done using different techniques for random under-sampling on a real data. It is shown that random forest for random under-sampling achieves a particular benefit in the classification.

Under-sampling is also the topic in [Zhao et al., 2018]. Similarly, the authors perform experiments on a data set concerning rare healthcare events using different methods for under-sampling. Other than the authors in [D. Dai and Hua, 2016], it is found here that the best results are achieved using logistic regression with synthetic minority oversampling technique (SMOTE). When looking at recall, a 45.3% increase is found when using SMOTE as when using only logistic regression.

Data limitations

The authors in [Farahani et al., 2018] discuss the use of Internet of Things (IoT) for medical purposes. They mention that IoT can be useful in healthcare to meet the increasing demands in an increasingly aging population. The challenges in IoT are mentioned to concern data management as a lot of data is generated, scalability, regulations, security, and privacy.

In [Shilo et al., 2020] an overall overview of challenges of big data in healthcare is given. The authors consider different characteristics of data, which they call axes. The authors mention that the axes often come at the cost of another. For example, increasing the population size of a data set may be limited by financial or organizational contraints. Another challenge the authors mention is the heterogeneity of the data. It is important to represent the full population, but this is not always obvious. It may occur in studies that the population is too homogeneous for the results to apply to the general population. Finally, the authors argue that interpretability of a predictive model is very important and not always straightforward.

The authors of [Dinov, 2016a] consider the complexity problem of big data. As data nowadays contains imaging, genetic and other complex data, new automated classification techniques are necessary. The authors state that even though the field of big data in healthcare is rapidly evolving, new techniques are still much needed in order to improve and scale the processing of large data sets.

A particular problems regarding sensor data is discussed in [Pike et al., 2019]. It is mentioned that such data, when collected from different sources, can give different values and different results. The authors suggest that this may be due to atmospheric variables or air pollution and propose that those variables should be taken into account.

The authors of [Dinov, 2016b] argue that even though there are many promising possibilities with machine learning in healthcare, those promises might not be fully realized without big technological advancement and a commitment to open science. It is pointed out that healthcare data is specifically often incomplete or inconsistent. Furthermore, the typical problem of big data is pointed out, namely the impracticality of high computation times and need of resources. The authors claim that data

analytics may be a more rapid approach towards estimation and prediction of big data.

In [Johnson et al., 2016] the focus is on challenges regarding data in healthcare. One problem addressed is that in order for the data to be useful, its quality must be of high standards. So, the data should be carefully archived and retrieved. Besides, the preprocessing of data is also mentioned to be an important point. The authors claim that machine learning algorithms in healthcare lag behind those in different fields of application. This may be partly due to a lack of consistent and reliable data management of hospitals.

In [F. Wang et al., 2019] the authors consider the challenges of deep learning in healthcare. The first challenge they mention is that generally still a lot of feature engineering has to be done before a deep learning model can be applied. Furthermore, they mention the large amount of data that is needed for deep learning models to function well. Besides the quantity of data, the quality of data is also important as otherwise it becomes hard for the model to recognize reliable patterns. Finally, the common problem of model interpretability is mentioned. Deep learning methods are often seen as black boxes, not giving clear explanations for a classification. The authors mention that there remains a big challenge in interpreting the results of models.

Interpretability

In [Shilo et al., 2020], the authors argue that interpretability of a predictive model is very important and not always straightforward.

One big concern on machine learning models is discussed in [Vellido, 2019]. Machine learning models tend to be complex and nonlinear, which make the interpretability and explainability difficult. The authors argue that especially in healthcare is this a concern and that this might lead to a limitation of the use of machine learning models in practice. It is proposed to make use of data and model visualization. Furthermore, the authors argue that the expertise of medical experts should be an integral part of the design of data-driven approaches.

The authors of [Thesmar et al., 2019] see the advantages of machine learning in healthcare, specifically in the detection of disease patterns. They also mention that in order for this to work, important issues that need to be taken into account include the confidence from patients, the transparency of the applied methods, and potential discrimination by algorithms.

In [F. Wang et al., 2019] the authors consider the challenges of deep learning in healthcare. The first challenge they mention is that generally still a lot of feature engineering has to be done before a deep learning model can be applied. Furthermore, they mention the large amount of data that is needed for deep learning models to function well. Besides the quantity of data, the quality of data is also important as otherwise it becomes hard for the model to recognize reliable patterns. Finally, the common problem of model interpretability is mentioned. Deep learning methods are often seen as black boxes, not giving clear explanations for a classification. The authors mention that there remains a big challenge in interpreting the results of

models.

Feature selection

In [Chandrashekar and Sahin, 2014] the authors offer an overview of several feature selection methods available. They consider filter, wrapper, and embedded methods and apply the studied techniques on standard datasets. The filter methods considered are correlation criteria (using the Pearson correlation coefficient) and Mutual Information. The wrapper methods discussed in the article are: sequential selection algorithms (sequential feature selection, sequential backward selection, sequential floating forward selection) and heuristic search algorithms (genetic algorithm). In the field of embedded algorithms, the authors discuss several possibilities among which Lazy Feature Selection. The authors mention that feature selection techniques show that more information is not always beneficial for machine learning applications and that feature selection provides benefits such as a better insight in the data and identification of irrelevant variables.

In [Y. Wang et al., 2009] the authors propose a tabu search algorithm for feature selection. In this algorithm a long-term memory is used to decrease the risk of getting trapped in a cycle around a local optimal solution. Besides, the method makes use of probabilistic neural networks. By experiments, the authors show on real-world data sets that their method achieves higher classification accuracy than previous studies, while selecting an equal number or fewer features.

The authors of [Zhang and Sun, 2002] experiment with using tabu search for feature selection. They test this algorithm on synthetic data and compare the results to classic algorithms, such as several sequential methods, and a branch and bound method. The authors state that the results are promising, often the tabu search algorithm finds the optimal or a near-optimal solution. This is in contrast to the sequential algorithms, which are more likely to get trapped in a local optimum. The authors mention that the application of tabu search for feature selection should be further explored by experiments on real data sets.

In [Vergara and Estévez, 2014] a review is given on methods for feature selection based on mutual information. Mutual information is a measure of the amount of information that a variable has about another variable. The authors state that feature selection should go beyond the concepts of relevance and redundance to include complementarity. A recently proposed unifying framework is presented by the authors, mentioning that this framework is able to retrofit successful heuristic criteria. Finally, the authors suggest some open problems in the field of feature selection to the readers.

The authors of [Janecek et al., 2008] investigate the relationship between attribute space reduction techniques and classification accuracy. Attribute space reduction includes both feature selection and dimensionality reduction. Attribute space reduction in this article is done by filter and wrapper techniques for feature selection, and principle component analysis (PCA) for dimensionality reduction. It is shown by the authors that wrapper approaches for feature selection tend to produce the smallest feature subsets, achieving competitive classification accuracy.

However, these approaches tend to be more computationally expensive than other feature selection methods. Furthermore, the results show that the accuracy based on PCA is very sensitive to the data type. The authors also note that the variance captured by the principal components is not necessarily an indicator for the classification performance.

With this thesis we have the intention to address several of the mentioned challenged. Specifically, we aim to address open challenges in feature selection, interpretability, dealing with imbalanced data, and evaluation of data-driven methods in a healthcare environment.

I.5 Conclusion

In this chapter we have seen an overview of existing literature concerning data and machine learning in the healthcare domain.

The availability of data in healthcare is ever growing, as data collection became easier due to more digitalization. Data is collected in hospitals, as well as in care centers and in other healthcare organizations. This data may include biological data, administrative data, insurance claims data, among others. As the amount of data increases, the use of data-driven methods for healthcare related problems increases as well.

In this literature review firstly, we looked at the growing amount of data generated in healthcare and the different types of data that exist in healthcare. The types of data in healthcare vary greatly, it may exist of images, signals, free text, and administrative data, among others. We presented several studies which were done on different types of data, where a distinction was made between structured data and unstructured data.

Secondly, we investigated different methods used in the healthcare domain. The most common machine learning methods applied were found to include Logistic Regression, Random Forest, Decision Tree, Neural Network, and Support Vector Machine. A variety of applications were presented. In literature different methods are shown to provide good results in different areas in healthcare.

Finally, some open challenges and limitations were discussed. One important point is the data which is produced in healthcare. Clearly, data acquired in the healthcare system are very personal and very sensitive. It is thus of great importance to make sure the data is secured properly. Besides, the quality of the data is crucial. If there is a lot of missing data or many values are incorrect, this will directly influence the quality of the results from machine learning algorithms. Furthermore, we have seen that imbalanced data creates a challenge in the use of data-driven methods. Data in healthcare tends to be highly imbalanced, for example when looking at a rare disease. Several methods have been tested in literature. Another challenge in machine learning is the interpretability of the results. Many methods are, at least to an extent, a black box, where it is unclear what happens inside the algorithm. Especially in healthcare this may pose problems as it is important to know why, for example, a certain treatment is suggested. Lastly, the challenge of

feature selection is discussed. In healthcare generally many variables are kept track of, which may not all be relevant to the considered problem. It is shown to still be a challenge how to find the optimal set of variables for making the wanted prediction or decision.

As was mentioned before, this literature review was kept generic on purpose. In the remaining chapters, each chapter will have a more specific literature review.

Chapter II

Classification and Feature Selection for Readmission Prediction

Contents of the chapter

Abstract of the chapter

In this chapter we consider the problem of emergency department readmission and hospital readmission. Those are often considered as a lack of quality of care and should be avoided as much as possible. We look at the cases as classification problems. We present a framework in which we combine an optimization model with a feature selection method.

First, we look at relevant literature concerning hospital readmissions. We consider what has been done already on the topic and where improvements can still be found.

Next, we present the Tabu/DAMIP framework, in which we consider the DAMIP model for classification, which we combine with tabu search for feature selection. Besides, we present a data sampling technique which can be used in our framework in order to significantly decrease computation times.

After the methods have been described, we take a look at two case studies. First, we look at emergency department readmission prediction. In this case study, we only have a small amount of features, mostly concerning administrative data. It is shown that our framework achieves the highest F1-score, but the scores are all rather close to each other. Moreover, even the highest F1-score is still unsatisfactory. It seems that the range of data in this case study is not enough to make a reliable prediction. Therefore, in the next case study, we look at general all-cause hospital readmission. In this case study we have more data available. This data not only concerns administrative data, but also many medical features, mostly related to the medical diagnoses and acts of a patient. We can see in the results that indeed, having more information available gives us better performance of all methods. We compare our framework to the results presented in literature on similar problems. Our framework reaches the highest F1-score in all cases. Finally, we do some additional performance analysis. We show that using our sampling technique, we significantly decrease the computation time, while the F1-score does not decrease substantially.

II.1 Introduction

Hospital readmission rates are often considered as a measure of quality of care. Readmission events are costly and highly inconvenient to both the hospital and the patient. In France, around 15% of patients are readmitted to the hospital within 30 days of discharge [Gusmano et al., 2016]. By estimating the risk of readmission, measures may be taken and rehospitalization avoided.

The described problem can be seen as a classification problem where we use patient characteristics to make a prediction on the readmission of a patient. Classification is a widely studied problem in machine learning. The goal of classification is to determine to which sub-population a new observation belongs, based on data points where the sub-population is known. An observation is typically characterized by multiple variables or features. Not all known features may be relevant to the classification, which shows the necessity of feature selection, in which the most

pertinent features are chosen.

In this chapter we present the so-called Tabu/DAMIP framework. In this framework we combine tabu search for feature selection with the an optimization-based discriminant analyis model (DAMIP). Both parts will be explained later in this chapter. The framework is applied to two real-life case studies concerning emergency department readmission and general hospital readmission.

The remainder of this chapter is structured as follows: an overview of relevant literature is given in Section II.2. Then, the DAMIP model is presented in Section II.3 and Tabu search for feature selection in Section II.4. An overview of the complete framework is presented in Section II.5 and the framework including our data sampling method is presented in Section II.6. The first case study, concerning emergency department readmission, is given in Section II.7. The second case study, on general hospital readmission, is shown in Section II.8. Finally, concluding remarks are given in Section II.9.

II.2 Literature review

In this section we take a look at what has been done in literature on the topic of readmission prediction. Firstly, we take a look at the prediction of hospitalization in general. After, we examine the literature specifically on the prediction of hospital readmissions.

II.2.1 Hospitalization prediction

The goal of the authors of [Brisimi et al., 2019] is to develop a model to predict hospitalization of patients due to complications attributed to type 2 diabetes, during the following year. For each patient, the features are taken from the EHR data. The goal is to, among those patients, differentiate between patient who will be hospitalized within a year and those who will not be hospitalized within a year. The authors test several methods, including support vector machines, random forest, and gradient tree boosting. Besides, the authors propose a new framework using a statistical procedure. The different methods were tested on data from the Boston Medical Center. The framework proposed by the authors achieves an area under the ROC curve score of 89%. Even though other methods can increase this number to 92%, the authors state that those methods carry a higher computational cost and a lack of interpretability.

The authors of [Brisimi, Chen, et al., 2018] try to predict hospitalizations for cardiac events. For this goal they apply the soft-margin l1-regularized sparse support vector machine classifier. This is extended by the development of an iterative cluster primal dual splitting algorithm for solving the large-scale problem in a decentralized way. The authors show that the proposed framework converges faster than centralized methods and achieves similar area under the ROC curve accuracy.

In [Brisimi, Xu, Wang, Dai, Adams, et al., 2018] the authors aim to predict hospitalizations due to two common chronic diseases: heart disease and diabetes. The predictions are made based on the data from the patients' electronic health records

(EHR). Several methods are tested including spare support vector machines, sparse logistic regression, and random forests. Besides, the authors propose two new methods: a likelihood ratio test-based method and a joint clustering and classification method. It is shown that the best results are achieved by the joint clustering and classification method.

The goal of the authors of [W. Dai et al., 2015] is to predict hospitalization of patients with heart diseases. The data used in this prediction comes from the Electronic Health Records (EHRs) of the patients. In this paper five machine learning algorithms are applied to the data: support vector machines, AdaBoost, logistic regression, naive Bayesian, and a variation of a likelihood ratio test. The authors show that the results from all five models are consistent, which, according to the authors, indicates the limit on the possible prediction accuracy. The detection rate is shown to be able to reach 82%.

II.2.2 Readmission prediction

In the systematic review [Artetxe et al., 2018] the authors explore existing literature on hospital readmission. The authors show that the most commonly considered readmission delay is 30 days. Moreover, the authors state that readmission is intrinsically an imbalanced classification problem, as fewer people are readmitted than not. However, only four out of the 77 studies use any time of imbalance addressing technique. The most used method for dealing with imbalanced data is resampling. For classification the different studies use techniques varying from logistic regression or other regression techniques to machine learning techniques, such as tree-based methods and support vector machines. The results are compared based on Area Under the ROC Curve (AUC), which values range between 0.54 and 0.92 in the different studies. The authors of [Shankar and Manikandan, 2019] consider the problem of 30-day hospital readmission. They state that, according to the Agency for Healthcare Reasearch and Quality (AHRQ), the United States alone has spent 41.3 billion dollars between January and November 2011 to treat patients readmitted within 30 days of discharge. The authors create a baseline using SVM and random forest. As their proposed method, they developed a deep neural network based on an optimized sequential architecture. The authors show that their method outperforms the baseline methods based on accuracy. In [Lai et al., 2018] the authors use a wrapper method integrating genetic algorithm and support vector machine for 30-day readmission prediction. They test their method on hospital data from Taiwan and use four different objectives: accuracy, sensitivity, specificity, and AUC. The results for those objectives are, respectively, 69.33-71.44%, 66.27-69.41%, 69.32-72.24%, and 0.7518-0.7601. The authors of [T. Wang and Paschalidis, 2019] propose a new method, called Prescriptive Support Vector Machine (PSVM). This method is based on the well-known SVM method, but with three features added. First of all, a regularization constraint is introduced to induce a sparse classifier. Besides, the authors devise a method that partitions the positive class into clusters and selects a sparse SVM classifier for each cluster. Thirdly, a method is developed for optimizing the values of controllable variables with as goal reducing the number of data points

predicted to be in the undesirable group. The proposed method is tested on a data set of 2.28 million patients over a four year period, with the objective being predicting and preventing 30-day hospital readmissions. The authors show that their method reduces the readmission rates by an average of 1.24%.

In [Ashfaq et al., 2019] the problem of 30-day readmission for patients with Congestive Heart Failure (CHF) is considered. In this research a Long Short-Term Memory (LSTM) neural network is presented, which makes use of expert features and contextual embedding of clinical concepts.

In [Salzman et al., 2019] the authors consider hospital readmission of older patients. They consider the Probability of Repeat Admission (PRA), the Vulnerable Elders Survey (VES-13), and a provider estimate of likelihood of hospitalization, to try to identify patients at high risk for emergency department visits or hospitalization at 6 and 12 months. The goal of the study is to determine the feasibility of this risk identification from a sample of 60 adults aged 65 and older. The authors found that PRA and provider estimate were not significant predictors of hospitalization at 6 months, but they were at 12 months. Similarly, a hospitalization during the year before was not a significant predictor of hospitalization at 6 months, but it was at 12 months. None of the tools was a significant predictor of ED visits, independent of the time.

The authors of [Jain et al., 2019] focus on using large amounts of data fundamental to readmission prediction analysis. A framework is proposed based on High Performance Computing Cluster (HPCC) for big data readmission risk analysis. This framework makes use of the Naive Bayes classification algorithm. The authors show that their framework can decrease the evaluation time significantly while maintaining model performance.

In [Ramirez and Herrera, 2019] hospital readmission of patients with diabetes is considered. The authors apply simple machine learning models of which the best results are achieved by a random forest model. This model outperforms deep learning techniques, while it requires significantly less computing power.

The authors of [Schwab et al., 2019] focus on readmission of elderly patients. This article gives a systematic review of studies on the given subject. In total, 12 studies were included in the review. In those studies the area under the receiving operating characteristic curve is shown to be between 0.45 and 0.69. The studied patients are in some studies 65 years and older and in other studies 75 years and older. Readmission rates vary between 12.1% and 28.4%.

The authors of [Futoma et al., 2015] compare several models for the prediction of early hospital readmissions. The tested methods are: logistic regression, logistic regression with multi-step variable selection, penalized logistic regression, random forest, and support vector machine. All methods show similar results, but the best one is achieved by random forest with an AUC score of 0.684.

[Flaks-Manov et al., 2019] researches the timing of readmission risk prediction. They state that generally readmission prediction is done at the time of discharge of the patient, but that often this is too late, as intervention to prevent readmission is not possible at this time anymore. The authors state that at-admission models allow for early identification of possible readmissions and thus allowing intervention.

However, this type of model may miss patients who are at high risk of readmission caused by factors accrued during hospitalization. In conclusion, the authors recommend an approach applying readmission risk detection at both admission and at discharge. The goal in [Lee et al., 2012] is to classify readmission to the emergency department within 72 hours. For this intention the authors combine particle swarm optimization (PSO) for feature selection with a classification model (DAMIP). The data used in this study includes 96 factors for each of the patients, including chief and secondary complaint, physician diagnosis, 5 factors related to demographic information, 8 factors related to patient arrivals, 44 factors related to the treatment and procedures received, and 35 factors related to the hospital environment. The results are compared to linear discriminate analysis, naïve Bayesian classifier, support vector machine, logistic regression, decision tree, random forest, and nearest shrunk centroid. It is shown that the proposed framework achieves the best results.

In this literature review we have seen that prediction of hospitalization and readmission has widely been researched. Many methods have been tried out with the goal of prediction a hospital (re)admission. The focus in research seems to be on hospital readmission. On the other hand, emergency department readmission is a lot less studied. In this chapter we will provide two case studies, where one is on emergency department readmission and the other on general hospital readmission.

Moreover, we can see that in general quite classical methods are applied to this classification problem, such as SVM and logistic regression. We will experiment by using a non-classical classification model, which has been shown useful on emergency department readmission prediction. Besides, we will combine this model with a feature selection method, which works together with the classification model.

II.3 DAMIP classification model

All classifications, both final and intermediary, in the framework are done using the DAMIP classification model. This model was introduced in 2012, in [Lee et al., 2012]. The DAMIP classification model is an optimization-based discriminant analysis model, which has as goal to optimize the total number of correctly classified entities. As an additional asset, this model provides the option to limit the number of misclassifications. For each class the upper bound on misclassifications can be set separately. By using the DAMIP model for classification we can make a clear distinction between classification and feature selection, which will be discussed later this chapter.

The relevant notation for the mathematical model is given below.

Sets

$\mathcal{G}$	Groups to which an entity can be classified	$k \in \mathcal{G} = \{1, 2, ...\}$
$\mathcal{O}$	Entities	$i \in \mathcal{O} = \{1, 2, ...\}$

Parameters

π_k	Prior probability of group k
$f_k(x)$	Conditional probability function of group k
α_{hk}	Upper bound on misclassification where the observations of group k are classified to group h
y_i	Group to which entity i belongs

Variables

λ_{hk}	Non-negative constants giving the optimal decision rule
u_{ki}	Equals one if entity i is classified to group k and zero otherwise
L_{ki}	Loss functions

Anderson [Anderson, 1969] proposes to seek for a partition $\{R_0, R_1, \ldots, R_K\}$, where R_k is the region assigned to group k and R_0 is a region for "deferred judgment". This region is introduced to be able to put a restriction on the probability of misclassification. The model proposed by Anderson to find the described partition is as follows.

$$\text{Max} \quad \sum_{k \in \mathcal{K}} \pi_k \int_{R_k} f_k(x)dx \tag{II.1}$$

$$\text{s.t.} \quad \int_{R_h} f_k(x)dx \leq \alpha_{hk} \qquad \forall h, k \in \mathcal{K}, h \neq k \tag{II.2}$$

where π_k is the prior probability of group k, $f_k(x)$ is the conditional probability density function of group k and α_{hk} is the predetermined limit on the misclassifications where the observations of group k are classified to group h.

Anderson showed that there exist non-negative constants $\lambda_{hk}, h, k \in K, h \neq k$, such that the optimal decision rule is given by:

$$R_k = \{x \in \mathbb{R}^m : L_k(x) = \max_{h \in \{0\} \cup \mathcal{K}} (L_h(x))\}, k \in \{0\} \cup \mathcal{K} \tag{II.3}$$

where

$$L_0(x) = 0 \tag{II.4}$$

$$L_k(x) = \pi_k f_k(x) - \sum_{h \in \mathcal{K}, h \neq k} \lambda_{hk} f_h(x), k \in \mathcal{K} \tag{II.5}$$

With decision rules given in (3)-(5), the classification model (1)-(2) can be transformed into linear mixed integer programming models.

Below the formulation of the Discriminant Analysis Mixed Integer Programming (DAMIP) model, as presented by [Lee et al., 2012], is given.

$$\text{Max} \quad \sum_{i \in \mathcal{O}} u_{y_i i} \tag{II.6}$$

$$\text{s.t} \quad L_{ki} = \pi_k f_k(\mathbf{x}_i) - \sum_{h \in \mathcal{G}, h \neq k} f_h(\mathbf{x}_i)\lambda_{hk} \qquad \forall i \in \mathcal{O}, \forall k \in \mathcal{G} \tag{II.7}$$

$$u_{ki} = \begin{cases} 1 & \text{if } k = argmax\{0, L_{hi} : h \in \mathcal{G}\} \\ 0 & \text{otherwise} \end{cases} \qquad \forall i \in \mathcal{O}, \forall k \in \{0\} \cup \mathcal{G} \tag{II.8}$$

$$\sum_{k \in \{0\} \cup \mathcal{G}} u_{ki} = 1 \qquad \forall i \in \mathcal{O} \tag{II.9}$$

$$\sum_{i : i \in \mathcal{O}_h} u_{ki} \leq \lfloor \alpha_{hk} n_h \rfloor \qquad \forall h, k \in \mathcal{G}, h \neq k \tag{II.10}$$

$$u_{ki} \in \{0, 1\} \qquad \forall i \in \mathcal{O}, \forall k \in \{0\} \cup \mathcal{G} \tag{II.11}$$

$$L_{ki} \in \mathbb{R} \qquad \forall i \in \mathcal{O}, \forall k \in \mathcal{G} \tag{II.12}$$

$$\lambda_{hk} \geq 0 \qquad \forall h, k \in \mathcal{G}, h \neq k \tag{II.13}$$

Objective

The decision variable u_{ki} equals one if entity i is classified to group k and zero otherwise. The parameter y_i gives the group to which entity i truly belongs. That is, $u_{y_i i}$ equals to 1 if entity i is classified correctly. The objective (II.6) of the model is to maximize the number of correctly classified entities.

Constraints

— π_k is the prior probability of an entity belonging to group k. $f_k(\mathbf{x}_i)$ represents the conditional probability density function. It is defined to be the probability of the features having the values given in $\mathbf{x}_i$, given that the entity is in group k. λ_{hk} are the non-negative constants such that the optimal decision rule is determined, this is a decision variable. Constraints (II.7) and (II.8) determine the classification of the entities, the entity is classified to the group k, for which L_{ki} is maximum, and to the unclassified group if the maximum is negative.

— Constraint (II.9) makes sure that every entity is assigned to exactly one group.

— Constraint (II.10) puts an upper bound on the allowed rate of misclassification, where α_{hk} is a parameter which is to be set by the user, and n_h is the number of entities in group h.

The non-linearity of the model makes it more impractical to solve. [Yuan, 2015] proposes a linear version of the model, which is given below.

$$\text{Max} \quad \sum_{i \in \mathcal{O}} u_{y_i i} \tag{II.14}$$

$$\text{s.t} \quad L_{ki} = \pi_k f_k(\mathbf{x}_i) - \sum_{h \in \mathcal{G}, h \neq k} f_h(\mathbf{x}_i)\lambda_{hk} \qquad \forall i \in \mathcal{O}, \forall k \in \mathcal{G} \tag{II.15}$$

$$a_i - L_{ki} \leq M(1 - u_{ki}) \qquad \forall i \in \mathcal{O}, \forall k \in \mathcal{G} \tag{II.16}$$

$$a_i \leq M(1 - u_{0i}) + \epsilon \qquad \forall i \in \mathcal{O} \tag{II.17}$$

$$a_i - L_{ki} \geq \epsilon(1 - u_{ki}) \qquad \forall i \in \mathcal{O}, \forall k \in \mathcal{G} \tag{II.18}$$

$$a_i \geq \epsilon u_{ki} \qquad \forall i \in \mathcal{O}, \forall k \in \mathcal{G} \tag{II.19}$$

$$\sum_{k \in \{0\} \cup \mathcal{G}} u_{ki} = 1 \qquad \forall i \in \mathcal{O} \tag{II.20}$$

$$\sum_{i : i \in \mathcal{O}_h} u_{ki} \leq \lfloor \alpha_{hk} n_h \rfloor \qquad \forall h, k \in \mathcal{G}, h \neq k \tag{II.21}$$

$$u_{ki} \in \{0, 1\} \qquad \forall i \in \mathcal{O}, \forall k \in \{0\} \cup \mathcal{G} \tag{II.22}$$

$$L_{ki} \in \mathbb{R} \qquad \forall i \in \mathcal{O}, \forall k \in \mathcal{G} \tag{II.23}$$

$$\lambda_{hk} \geq 0 \qquad \forall h, k \in \mathcal{G}, h \neq k \tag{II.24}$$

$$a_i \geq 0 \qquad \forall i \in \mathcal{O} \tag{II.25}$$

In this model, ϵ and M represent a small and a large number, respectively. The constraint set (II.8) of the non-linear model is replaced by constraint sets (II.16) - (II.19) in the linear model. These constraints make sure that the considered entity is classified into the group with the highest value of L_{ki} or in the group of reserved judgment if all values of L_{ki} are negative.

II.4 Feature selection

In the DAMIP classification model we presented, patients are compared to each other based on their characteristics, which we call features. However, not all features may be of interest to the desired prediction and providing too many features to the model only causes noise. Therefore, it is important to decide on which features should be used in the prediction. Optimizing the subset of features will result in an optimized classification. As the number of possible combinations increases quickly with the number of features, a full enumeration algorithm will not be possible within reasonable computation times. We will thus resort to a metaheuristic, tabu search.

Tabu search, first introduced in [Glover, 1989], is a local search algorithm, which tries to escape possible local minima by keeping track of a list of forbidden solutions (the tabu list). The specific notation for tabu search is given below, in Table II.1.

Symbol	Description
$\mathbf{x}$	Initial solution
c	Size of candidate set
$C(\mathbf{x})$	Candidate set of solution $\mathbf{x}$
$\mathbf{y}$	Best solution in $C(\mathbf{x})$
l	Tabu list length
a	Objective value of best found solution
a'	Objective value of current solution
$\mathbf{p}$	Features of best found solution

Table II.1 – Notation for Tabu search

In the algorithm, we start with an initial solution $\mathbf{x}$, which represents a specific set of features. For this initial solution, the criterion function is evaluated. Next, a set of candidate moves is considered. A candidate move is the move from one subset of features to another subset of features, where exactly one feature differs in presence or absence. We take into consideration c candidate moves, where c is a parameter specified by the user. The candidate moves are selected at random, by randomly selecting one of the features, each with equal probability, and changing the presence or absence of this feature. If the best of these moves is not in the tabu list, this solution is now considered to be the current solution and this solution is placed in the tabu list (TL), which has length l. The tabu list prevents moves to be reversed within l iterations, l to be set by the user. The procedure is repeated for a specified number of iterations. Similar to before, $\mathbf{p}$ represents the features used in the best found solution so far and a is the corresponding objective value.

Algorithm

Algorithm 1 Tabu search

Input: All features and variable to be predicted
Output: Selection of features

generate an initial solution $\mathbf{x}$;
TL $\leftarrow \mathbf{x}$;
$\mathbf{p} \leftarrow \mathbf{x}$;
a $\leftarrow evalf(\mathbf{x})$;
for k from 1 to i **do**
  Form candidate set C($\mathbf{x}$) ;
  $a' \leftarrow 0$;
  for $\mathbf{z}$ in C($\mathbf{x}$) **do**
    if ($\mathbf{z}$ not in TL) and $evalf(\mathbf{z}) > a'$ **then**
      $\mathbf{y} \leftarrow \mathbf{z}$;
      $a' \leftarrow evalf(\mathbf{z})$;
    end if
  end for
  if $a' > a$ **then**
    $a \leftarrow a'$;
    $\mathbf{p} \leftarrow \mathbf{y}$;
  end if
  push $\mathbf{y}$ in TL ; //Add $\mathbf{y}$ to the end of TL
  if size of TL $>$ l **then**
    shift(TL) ; //Remove the head element of TL
  end if
  $\mathbf{x} \leftarrow \mathbf{y}$;
end for

II.5 Tabu/DAMIP framework

In Figure II.1 an overview of the Tabu/DAMIP framework is given. In summary, first a random set of features is chosen and evaluated. After, in each iteration, the neighborhood of the subset of features is evaluated. From this subset the best performing subset is chosen as the next solution, even if it is worse than the current solution. After a fixed amount of iterations, the overall best subset of features is chosen, from which we can get the actual classification using DAMIP.

II.6 Sampling

Besides the standard Tabu/DAMIP framework as described before, we also propose an extra option in the framework. We add data sampling to the framework with the goal of speeding up the complete process. The fact that we use the DAMIP

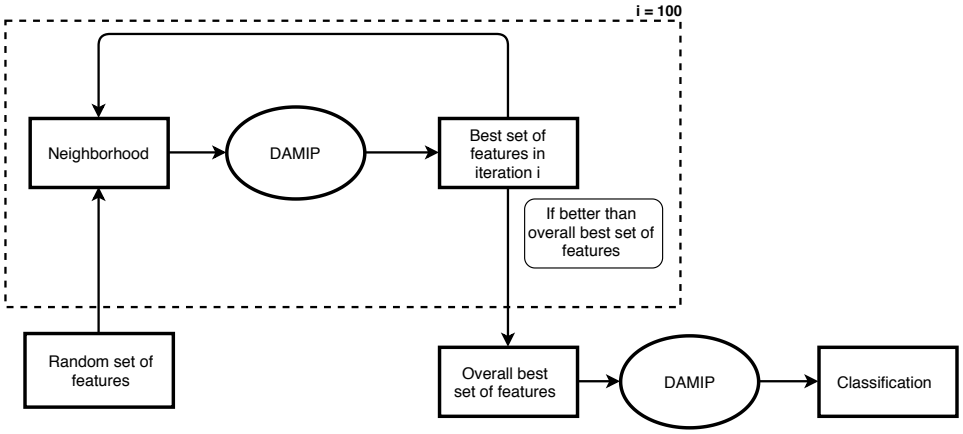


FIGURE II.1 – Overview of the proposed method

model in each iteration has a big impact on the computation time. The classification model is called many times in the process, resulting in possibly high computation times. In the proposed sampling approach, in each iteration we take a random sample of the data set. This sample consists of a specified percentage of the original data, which can be set by the user. Using this sample we call DAMIP for a quick performance evaluation of the neighborhood. Because of the smaller quantity of data, this runs significantly faster. After the best subset of features within an iteration is determined, we run DAMIP once more with the full data set, this result is compared to the best known solution so far. In the next iteration a new subset of data is selected and again the best solution in the neighborhood is tested using all data, which is compared to the best known solution. An overview of the framework including sampling is given in Figure II.2.

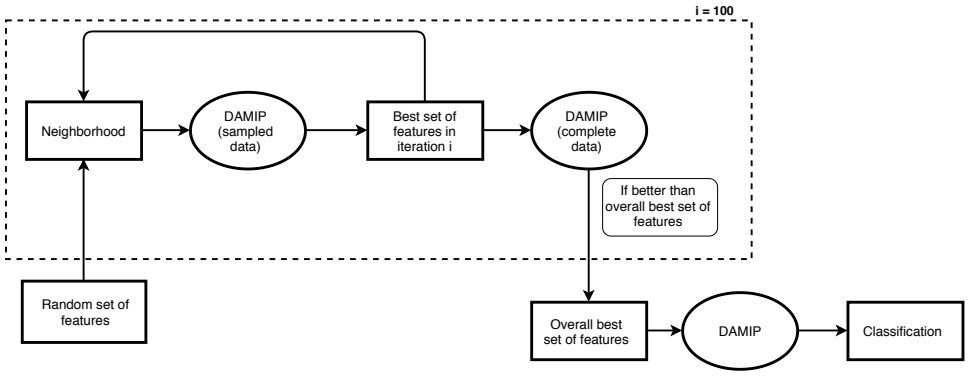


FIGURE II.2 – Overview of the proposed method with sampling

II.7 Case study: emergency department readmission prediction

In this section we consider a classification problem, with the goal of predicting whether a patient will visit the emergency department (ED) again after being discharged from the ED earlier. Readmission to the emergency department is often

considered a measure of quality in healthcare [Benbassat and Taragin, 2000]. A considerable part of readmissions have been judged to be preventable. By estimating the risk of readmission, measures may be taken and revisits avoided.

II.7.1 Data description

The complete data set concerning emergency department admissions consists of almost 12 million admissions in France. However, this is too much for most classification algorithms; the training time would be impractically high. Therefore, a subset of the data was selected. The data used in this study consists of 91 000 emergency department admissions, which represents approximately one month of emergency department admissions. Among these admissions, around 5% of the patients return within 72 hours. Features known for each of the admissions include the age of the patient, the gender, the arrival mode, the urgency level, and the main diagnosis. In total 44 features are taken into account. Among the 91 000 admissions, 54.7% are men. The vast majority (97.1%) of the patients arrived from home (in contrast to arriving from a medical unit). Moreover, after treatment at the emergency department, most patients (78.1%) return home. The most common main diagnosis is a traumatic injury (44.6%).

II.7.2 Experiments

In this section we provide a benchmark to compare our developed framework to classical classification methods listed in Table II.2. The machine learning algorithms to which we compare our results are shortly explained below. A more extensive description can be found in Appendix A.

Linear discriminant analysis is a commonly used technique for data classification. The method tries to maximize the ration of between-class variance to the within-class variance, guaranteeing maximal separability [Balakrishnama and Ganapathiraju, 1998]. A Naive Bayesian classifier based on Bayes' theorem is a probabilistic statistical classifier [Yoo et al., 2012]. This classifier is based on the assumption that all features are independent of each other. Fundamentally, support vector machines search for the optimal separating hyperplane, where the margin between two different objects is maximal. To find this maximal margin, support vectors are used [Yoo et al., 2012]. Logistic regression is a statistical regression model, which has as an advantage that it provides the user explicitly with probabilities and not only the class label information [Shevade and Keerthi, 2003]. Classification tree classifiers construct a tree structure, where at every step an attribute is sought whose sorting result is closest to the pure partitions by the class in terms of class values [Yoo et al., 2012]. Random forest was introduced in [Breiman, 2001]. This algorithm uses a group of classification tress, each of which is built using a bootstrap sample of the data [Diaz-Uriarte and De Andres, 2006]. In the nearest shrunken centroid algorithm for classification, shrunken centroids are used for each class and test samples are classified to the class whose shrunken centroid is nearest to it [Tibshirani et al., 2003]. Neural network attempts to mimic the neurological functions of the

brain [Yoo et al., 2012]. Neural network consists of nodes mimicking the functions of neurons in the brain. The nodes are interconnected via links with adjustable weights. The weights are adjusted by learning.

II.7.3 Performance measures

For the outcomes of the experiments, we look at the performance measures accuracy, precision, recall, and F1-score. Note that in the next definitions a positive entity implies a hospital stay for which the target variable is positive, implying a readmission.

— True Positive (TP): positive entities correctly classified as positive.

— True Negative (TN): negative entities correctly classified as negative.

— False Positive (FP): negative entities wrongly classified as positive.

— False Negative (FN): positive entities wrongly classified as negative.

Using above definitions, we can define the following performance measures:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{II.26}$$

Accuracy shows the percentage of correctly classified instances over all instances.

$$Precision = \frac{TP}{TP + FP} \tag{II.27}$$

Precision represents the percentage of correctly classified entities among all instance classified as being positive.

$$Recall = \frac{TP}{TP + FN} \tag{II.28}$$

Recall is the percentage of all positive instances being classified correctly.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{II.29}$$

The F1-score is the harmonic mean between precision and recall.

The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate. The Area under the ROC curve (AUC) is a commonly used performance measure in machine learning, which we will also provide.

Because of the imbalanced data used in this study, we will not look only at accuracy as this gives a false indication of performance. Both precision and recall focus on the recognition of the minority class. We will use F1-score for comparison between results as this combines the precision and recall into a single score.

II.7.4 Results

The results for the presented case study are shown in Table II.2. Those results were obtained using the following parameters: the candidate size is equal to the number of features in the data set and the tabu list length is set to 50. The algorithm

Method	Accuracy	Precision	Recall	F1	AUC
Random Forest	0.630	0.068	0.528	0.121	0.582
Support Vector Machine	0.635	0.074	0.570	0.131	0.604
Naive Bayesian	0.896	0.112	0.170	0.135	0.551
Nearest Shrunken Centroid	0.474	0.058	0.652	0.106	0.558
Linear Discriminant Analysis	0.706	0.075	0.449	0.129	0.584
Logistic Regression	0.677	0.073	0.489	0.128	0.588
Neural Network	0.624	0.073	0.591	0.130	0.608
Classification Tree	0.667	0.068	0.465	0.119	0.571
DAMIP (without tabu)	0.644	0.069	0.525	0.121	0.585
DAMIP + tabu	0.693	0.082	0.485	**0.140**	0.594
DAMIP + tabu with sampling	0.652	0.074	0.536	0.131	0.597

Table II.2 – 72 hour emergency department readmission

runs for 100 iterations before terminating. In our data sampling approach, we select 10% of the original data set.

From the results it can be seen that in general the F1-score is very low for any of the algorithms. The highest F1-score is achieved by our Tabu/DAMIP framework, but this value is still very low and also very close to all other algorithms. Even if we apply only DAMIP to the complete data set, without any feature selection, the F1-score is rather close to the best found score. We can deduce from this that feature selection is not very relevant in this case study. It seems that the readmission prediction is very complicated and that possibly the data set does not contain the right information to explain a readmission.

The results also show why we do not use accuracy as our main performance measure. We can see that for Naive Bayesian the accuracy reaches 89.6%, which makes our result seem like a reasonable one. However, when we look at the other measures, we get a different image. This is an effect caused by the imbalanced nature of the data. By classifying a large part of the entities to the majority class, we can easily reach a high accuracy score.

In this case study on emergency department readmission prediction we had quite a low number of features, containing mostly administrative data of patients. Assuming that this causes the low quality results, we will take a look at a second case study in which we have a lot more data about each patient. This case study concerns general hospital readmission and in this case study we also have, besides administrative data, medical data, such as the diagnoses and medical acts.

II.8 Case study: hospital readmission prediction

Similar to emergency department readmission, hospital readmission is often considered as an indication of lack of quality of care. Generally those readmissions are seen as avoidable events. If we can predict those readmissions, it will have a

financial benefit and it will prevent a big burden on the hospital and the patient. In this case study, we have access to more detailed data for each patient, containing a large variety of diagnoses and medical acts.

II.8.1 Data

The data used in this study is from a French hospital and covers 5 years of hospitalizations. The data set contains a total of 75239 hospital admissions. For each, we have the information about the patient, the diagnosis and the treatments: age, gender, medical speciality of admission, IGS2 score (in French: *Indice de Gravité Simplifié*, indication of gravity used in the intensive care unit), ICD-10 codes for diagnoses, CCAM (in French: *Classification Commune des Actes Médicaux*) codes of medical procedures, mode of arrival, mode of departure, and time of arrival and departure.

The most common age group among the admitted patients is from 51 to 70 years, with an occurrence of 44.7%. The age groups 71 to 80 years and 26 to 50 follow in order of frequency with occurrences of 21.3% and 19.9% respectively. Less frequent age groups are 80 years and older (10.1%), 16 to 25 years (3.8%), and 6 to 15 years (0.2%). Patients between the age of 0 and 5 years do not occur in the data set used in our study. The gender of the patients is quite balanced with 51.3% women and 48.7% men.

It is interesting for the results section (Section II.7.4) to note that large majority of patients (88.6%) return home after discharge from the hospital. A much smaller part of the patients (10.9%) go to a convalescent hospital, and 0.5% of patients deceased in the hospital.

The diagnoses are represented by ICD-10 codes, it contains one letter, followed by one to five numerals, and optionally a seventh character, which is a letter. To reduce the number of possible diagnoses, the codes have been reduced to in total three characters: one letter followed by two digits. In total our data set contains 654 different diagnoses. The three most common diagnoses are M17 (Osteoarthritis of knee), M23 (Internal derangement of knee), and M16 (Osteoarthritis of hip) with frequencies of 5.2%, 5.2%, and 5.0%, respectively. In Figure B.1 we can see an overview of the occurrences (in %) of the 80% most common diagnoses.

Our data set contains 1906 different medical procedures, represented by CCAM codes. A procedure is represented by a code consisting of four letters followed by three digits. The three most frequently applied medical procedures are DEQP007 (Continuous monitoring of the electrocardiogram by oscilloscopy and / or telemonitoring), GLLD017 (Oxygen therapy with continuous oximetry monitoring), and GELD005 (Nebulization for bronchial use with monitoring of oxygen saturation), with frequencies of 4.8%, 4.0%, and 4.0%, respectively. In Figure B.2 we can see an overview of the occurrences (in %) of the 80% most common medical acts.

Note that for both diagnoses and medical procedures it holds that a patient can have multiple of either. Every diagnosis and every medical procedure is represented by a binary variable, indicating its presence or absence.

We consider readmission within different time periods. In our data set there are

9.4% readmissions within 30 days, 15.6% readmissions within 90 days, and 20.2% readmissions within 180 days. Some studies also consider shorter periods of readmission, such as 3 days or 7 days, however, in the data set used in this study there are no readmissions within those time periods. In case of multiple readmissions of a single patient in the considered time period, we consider the readmission after each individual stay, similar to single readmissions.

II.8.2 Data preparation

Because of the high number of medical procedures and diagnoses, we select only the 80% (116) most frequent medical procedures and the 80% (67) most frequent diagnoses as features in our data set. Those features are represented in a binary way, the feature gets value 1 if the concerning patient has received the medical procedure or diagnosis, and 0 otherwise.

Similarly to the medical procedures and diagnoses, most features are represented by binary variables. The binary variables include the gender of the patient, the age group (6-15, 16-25, 26-50, 51-70, 71-80, 81+), IGS2 gravity score group (0, 1-40, 41-70, 71-80, 81-90, 91-100), 13 medical units, and a variable indicating whether the patient has been hospitalized in the previous 30, 90 or 180 days.

Besides the binary variables, our data set consists of four categorical variables indicating how and from where the patient arrives to the hospital, and how and to where the patient leaves the hospital. Another categorical variable consists of the first letter of the main diagnosis, which indicates the medical speciality. Finally, we have 3 variables to represent the homogeneous group of patients the specific patient belongs to. First of all, one categorical variable indicating whether a (minor) surgery is performed. Next, a categorical variable for a score of gravity (which can be both a number or a letter), and finally, a categorical variable which gives an indication of the group of diseases the patient is affected by.

Taking into account the variables as described, our data set contains in total 287 variables and 1 target variable, for readmission.

II.8.3 Experiments

Using the data as described before, we did experiments for readmission delays of 30 days, 90 days, and 180 days. The readmission delay is calculated as the time of admission minus the time of discharge of the previous hospitalization. With the given data we predict any type of readmission within the given time periods.

The goal of the performed experiments is to classify the readmission of patients within a specified time period, that is, the target variable is a binary variable that equals 1 if the patient is readmitted and 0 otherwise. The parameters to be decided are chosen by means of trying different combinations and choosing the best performing. Similar to the previous case study, we make use of a tabu list of length 50, we perform 100 iterations, and in our sampling approach we use 10% of the original data set.

As was mentioned before, the used data sets are imbalanced, with 20.2% read-

missions in a period of 180 days. Logically, the data sets with shorter readmission delays have lower rates of readmission. To improve the learning part of classification, we apply a balancing technique, undersampling. This means that after splitting the complete data set into training data (80%) and test data (20%), we delete a part of the majority class in the training data such that the distribution becomes more balanced. However, we do test on imbalanced data, as we are interested to see how the algorithms perform on a case which reflects reality.

II.8.4 Results

In this section, we present the results obtained for the described experiments. First, we will look at the classification performance of our framework compared to those in literature. After, we consider the features which have been selected and which seem to be important in the prediction of hospital readmission. Finally, we do some additional optimization performance analysis, where we look at the different options our framework offers and how it affects the results.

II.8.4.1 Classification performance analysis

The results for our classification problem are shown in Tables II.3, II.4, and II.5 for return delays of 30 days, 90 days, and 180 days, respectively.

In order to be able to compare our results, we look at articles from the literature in which different machine learning algorithms were applied to similar cases as ours. Those algorithms have all been programmed using the *scikit-learn* package in python. When the information is known, we use the configurations as are indicated in the mentioned literature, to make the comparison as reliable as possible.

As mentioned before, we will compare the different solutions based on F1-score. The results are established using 5-fold cross validation.

In Table II.3 we can see that for 30-day readmission the F1-scores of all the classic machine learning algorithms are similar, all achieving a value of approximately 0.3. When we apply the DAMIP model without any feature selection, achieve the lowest score, namely 0.158. This can be explained by the fact that all features are used, whereas not all of them will be relevant to the prediction. However, when we combine the DAMIP classification model with tabu search for feature selection, we manage to predict readmission better than all algorithms and we achieve an F1-score of 0.416.

In Table II.4 we can see that the performance of the standard classification algorithms slightly improves. Most likely this is due to the fact that in a 90-day period there will be more readmissions and thus the data is less imbalanced. However, the performance of the DAMIP classification model remains stable. On the other hand, like before, the F1-score does improve when the classification model is combined with tabu search and we achieve an F1-score higher than the other algorithms as well as for the 30-day readmission case.

For a readmission period of 180 days, the results are shown in Table II.5. The results are again slightly better than those shown in the previous two tables. The result obtained by DAMIP in combination with tabu search is the best F1-score we managed among the three data sets considering all algorithms.

Method	Accuracy	Precision	Recall	F_1	AUC
[Alajmani and Elazhary, 2019]					
Naive Bayesian	0.445	0.134	0.885	0.233	0.642
[Lee et al., 2012]					
Nearest Shrunken Centroid	0.785	0.214	0.473	0.294	0.645
Linear Discriminant Analysis	0.731	0.220	0.720	0.337	0.726
DAMIP + PSO	0.739	0.214	0.548	0.308	0.655
[Ramirez and Herrera, 2019]					
Random Forest	0.757	0.223	0.635	0.330	0.702
Support Vector Machine	0.899	0.454	0.314	0.371	0.637
Neural Network	0.712	0.204	0.704	0.316	0.708
[Sharma et al., 2019]					
Logistic Regression	0.737	0.224	0.723	0.342	0.731
Classification Tree	0.686	0.183	0.667	0.287	0.678
Proposed method					
DAMIP	0.898	0.592	0.091	0.158	0.542
DAMIP + tabu	0.832	0.328	0.568	**0.416**	0.715
DAMIP + tabu + sampling	0.831	0.287	0.524	0.370	0.694

Table II.3 – 30-day readmission classification results

II.8.4.2 Feature selection analysis

As we can see from the previous three tables, the best result is achieved on the data set concerning 180-day readmission, with an F1-score of 0.507. The features that were used in this classification are given in Table B.1.

From the features selected by tabu search, we can see that there are several features that are selected in all three instances of DAMIP with feature selection, without sampling. Those are indicated in boldface in Table B.1.

First of all, the feature *IGS2_71_80* seems to be important. IGS2 is an indication of gravity used in the intensive care unit. This score has a value between 0 and 100, where 100 is the worst state a patient can be in. This feature, which was chosen for the prediction of 30-, 90-, and 180-day readmission prediction, is a binary variable indicating whether a patient has an IGS2 score between 71 and 80.

The principal diagnoses *M16* and *K42* are also both chosen in all three cases. They represent osteoarthritis of the hip and an umbilical hernia, respectively. When we check the statistics in the data, we can see that for diagnosis *M16* the return rate is lower than average (7.8%). This also holds true for diagnosis *K42*, with a readmission rate of 14.1%.

Finally, four different medical acts were chosen as a feature in all three cases. The first, *EQQP011* represents continuous monitoring of intraarterial pressure, *BFGA004* represents Extracapsular lens extraction, with implantation of an artificial lens in the posterior chamber of the eye. The act *JCLE002* concerns the placement of a ureteral stent Finally, *HSLF002* represents the medical act of par-

Method	Accuracy	Precision	Recall	F_1	AUC
[Alajmani and Elazhary, 2019]					
Naive Bayesian	0.507	0.226	0.862	0.358	0.650
[Lee et al., 2012]					
Nearest Shrunken Centroid	0.739	0.286	0.425	0.341	0.612
Linear Discriminant Analysis	0.719	0.322	0.690	0.439	0.707
DAMIP + PSO	0.589	0.222	0.577	0.320	0.584
[Ramirez and Herrera, 2019]					
Random Forest	0.724	0.318	0.644	0.426	0.692
Support Vector Machine	0.832	0.456	0.284	0.350	0.610
Neural Network	0.703	0.306	0.684	0.423	0.695
[Sharma et al., 2019]					
Logistic Regression	0.727	0.330	0.695	0.448	0.714
Classification Tree	0.666	0.272	0.654	0.384	0.661
Proposed method					
DAMIP	0.838	0.620	0.087	0.153	0.538
DAMIP + tabu	0.752	0.356	0.631	**0.455**	0.704
DAMIP + tabu + sampling	0.739	0.294	0.459	0.358	0.625

TABLE II.4 – 90-day readmission classification results

enteral nutrition with an intake of 20 to 35 kilocalories per kilogram per day.

From the features that were chosen, we can see that quite many features were used in the prediction of readmission. However, only few of them were used in all three readmission delays. It might be caused by the large amount of features that are in the data set.

II.8.4.3 Optimization performance analysis

Additional information about the achieved results is given in Table II.6. Note that the F1-scores in this table are repeated from the three preceding tables. Besides the F1-score, the table shows the number of features chosen by each method.

In the case of applying only the classification model the number of features is always equal to the total number of features, 285, as no feature selection is performed. For the cases where we do apply feature selection, we can see that in all cases approximately a quarter of the features are selected. An example of such selection was shown before in Table II.6.

Furthermore, we show the percentage of unclassified entities in Table II.7. This is the percentage of entities in the test data, which have been notified by the model as undecided. This option is a big advantage of the DAMIP model, as we prefer unclassified entities over wrongly classified entities. This information is specifically interesting in combination with the F1-score. The results shown in the table are produced using the data of 180-day return. In our previous results, no entities were left unclassified, in order to make a fair comparison to the other algorithms. The

Method	Accuracy	Precision	Recall	F_1	AUC
[Alajmani and Elazhary, 2019]					
Naive Bayesian	0.511	0.277	0.855	0.418	0.638
[Lee et al., 2012]					
Nearest Shrunken Centroid	0.719	0.341	0.398	0.367	0.600
Linear Discriminant Analysis	0.708	0.379	0.668	0.484	0.693
DAMIP + PSO	0.635	0.265	0.639	0.374	0.637
[Ramirez and Herrera, 2019]					
Random Forest	0.694	0.355	0.605	0.447	0.661
Support Vector Machine	0.784	0.455	0.277	0.344	0.596
Neural Network	0.693	0.365	0.680	0.475	0.688
[Sharma et al., 2019]					
Logistic Regression	0.710	0.383	0.685	0.492	0.701
Classification Tree	0.648	0.318	0.632	0.423	0.642
Proposed method					
DAMIP	0.785	0.521	0.077	0.134	0.529
DAMIP + tabu	0.737	0.424	0.631	**0.507**	0.699
DAMIP + tabu + sampling	0.710	0.378	0.554	0.449	0.653

Table II.5 – 180-day readmission classification results

Method	Data set	F1-score	#Features
DAMIP	**30 days**	0.158	285
	90 days	0.153	285
	180 days	0.134	285
DAMIP + tabu	**30 days**	0.416	75
	90 days	0.455	73
	180 days	0.507	66
DAMIP + tabu with sampling	**30 days**	0.370	78
	90 days	0.358	82
	180 days	0.449	63

Table II.6 – # features for each method and data set

exact same features as in the previous results were used. Only the parameter *alpha* was adjusted in order to allow for fewer misclassifications. Our results show the trade-off between number of unclassified entities and F1-score. It can be seen that when we allow for entities to be left unclassified, we can get the F1-score up to 0.619, however, this comes at the cost of having a significant amount of unclassified entities.

Besides comparing the quality of the achieved results, we also make a comparison between the running times of the models with and without sampling in the tabu search feature selection. These running times are given in Table II.8.

F1-score	% Unclassified
0.507	0
0.564	16.3
0.619	43.7

TABLE II.7 – Trade-off between F1-score and % unclassified entities

Data set	Time without sampling (min)	Time with sampling (min)	Time difference
30 days	4485	1933	-56.9%
90 days	7932	2652	-66.6%
180 days	9992	2707	-72.9%
Average	7470	2431	-67.5%

TABLE II.8 – Running times with and without sampling

From the table we can see that substantial time savings are achieved. The time saving on the average running times is 67.5% and for the data set concerning 180-day readmission this even gets as high as 72.9%.

Do note that the mentioned running times include the whole training process. However, when this method would be use in a real case, the training process is done only once and thus the running time is not important. In this case only the computation time of the actual prediction is important, which can be done in at most one second. Therefore, when the quality of the prediction is considered to be the most essential characteristic, generally a method without sampling would have the preference.

II.9 Conclusion

Readmission to the hospital, or to the emergency department, is a highly undesirable event. If we can predict such an event reliably, this provides the medical staff with a tool to foresee the need for a more extensive treatment. In this way, readmission might be avoided, which would help both to increase the performance measures of the hospital as well as the well-being of the patient.

In this chapter classification and feature selection were discussed. We presented the Tabu/DAMIP framework. In this framework we make use of the discriminatory model DAMIP for classification and tabu search for feature selection. Besides, we have used a balancing technique as healthcare data is typically highly imbalanced and we have applied a sampling technique as a means to speed up the process of the complete framework.

Our proposed method was tested on two case studies. The first concerned emergency department readmission, for which a relatively small amount of features was given. We could see that the classification performance of most algorithms was poor.

The Tabu/DAMIP framework outperformed those algorithms, but only slightly. The second case study concerned all-cause hospital readmission in different time frames. This data set contained many features, mostly related to the medical acts and diagnoses of the patient. The results showed that the general classification performance was better for all algorithms. All machine learning algorithms benefited from the increased amount of information present. Besides, we could see that in the hospital readmission prediction, feature selection led to a better result compared to only DAMIP. In the case of the emergency department this difference was not as obvious. It seems that in order to make a reliable prediction we will need more information than just administrative data and that adding medical data will improve the results.

Moreover, we have shown that using a sampling technique, we can significantly decrease the computation times of the framework, even though this does cause a small decrease in performance. Finally, we looked at the possibility that DAMIP offers for leaving entities unclassified when the classification is too uncertain. When looking at the F1-score, we showed that, as expected, the performance increases if we allow for more unclassified entities.

In future research, it could be fruitful to look further into the feature selection process. In the case study concerning general hospital readmission we had a large number of medical diagnoses and medical acts, where not all of them might be of great importance. An increase in features means that more feature selection combinations are possible, implying an increase in computation time as well as less of a guarantee that a (near-)optimal combination will be found. This could potentially be solved by pre-selecting a group of features, which may be done based on several runs of the framework or on expert opinion.

Finally, in order to evaluate the performance of our methods, we have compared our results to those found in literature. However, we do not know the best possible score which can be reached given our data. It could be interesting in future research, to investigate the possibility of putting an upper-bound on the best possible solution. If we would have such measure, we also get insight into whether our results can still be significantly improved or not.

Chapter III

Autoencoding and Classification for Breast Cancer Treatment Decision Support

Contents of the chapter

Abstract of the chapter

This chapter considers the problem of post-surgery treatment decision for elderly breast cancer patients. This is generally consider a difficult decision as elderly patients are often in a worse physiological state and no general guidelines exist on which treatment is efficient for this specific group of patients. For this purpose we propose a framework combining an autoencoder with the previously presented classification model DAMIP.

First, we present existing literature both on the topic of autoencoding as well as on medical decision support for breast cancer treatment. From this overview we can see that autoencoding has been used in medical contexts before, but that it has not been combined with a large variety of classification algorithms, which we will do in this chapter. Besides, the specific problem of breast cancer treatment decision support was not widely found in literature.

After, we present the autoencoder, which is used to decrease the dimensionality of the data, while trying to preserve all the information. The result of the autoencoder is lower dimensional data containing continuous values. As we wish to combine the use of the autoencoder with the DAMIP model, it is necessary that we discretize the data. We present two discretization techniques and provide an example of how this transforms the output data from the autoencoder to input data for our classification model. Besides combining autoencoder with DAMIP, we also combine autoencoder with other classification models.

In the case study, we consider several approaches to the breast cancer treatment decision problem. First, we try to predict the decease of a patient within five years after surgery. Second, we try to predict whether a patient needs any kind of treatment or not. After, we consider the prediction of whether a patient needs chemo therapy or not. Finally, we try to predict the decease of a patient after having chemo therapy.

For this case study we have the availability over a rich data set containing a wide range of features. This includes administrative data, biological data, treatment data, and data about the cancerous tumor. This data has been gathered from different sources and gives a very detailed and complete image of the patients.

In the results we can see that autoencoding generally works rather well. We compare the current approach to the Tabu/DAMIP framework and we can see that we achiever similar results in much shorter computation times. Generally, the result of autoencoding in combination with DAMIP is not the best performing, which may be due to the fact that a discretization step is necessary. In this process some information might be lost. The best results are shown by combining autoencoder with linear discriminant analysis, this combination outperforms all other methods in each approach. By providing such reliable predictions, we can provide trustworthy decision support to practitioners.

III.1 Introduction

In this chapter we consider the problem of breast cancer treatment decision for elderly patients, where we focus on providing decision aid for the necessity of a post-surgery treatment. The decision of treatment for this specific group of patients is generally found to be complicated, as older patients are generally in a worse physiological state. As a result, no general guidelines exist on the most efficient treatment of elderly patients.

In the previous chapter we have seen that better results are achieved using DAMIP if we use lower-dimensional data. Where before we looked at features selection, in this chapter we will consider an encoding of data for dimensionality reduction. An efficient way of learning data encoding is an autoencoder. This makes use of an artificial neural network to learn a representation of a data set. Using this technique we can reduce the dimensionality of the data while keeping all the information present. Presumably, the lower-dimensional data will be useful for classification. Especially the high number of features, relative to the number of patients and the large heterogeneity of our data set, makes the use of an autoencoder seemingly interesting. Moreover, the data set used in the case study of this chapter has some missing data. An autoencoder might prove useful for this purpose, as known and unknown data are compressed together, keeping all the information which we do have.

In this chapter we present the methodology to reduce the dimensionality of a data set by means of an autoencoder. We propose several methods for classification to be applied after dimensionality reduction. One of the methods used for classification is the DAMIP model, which has shown good results in the previous chapter. To be able to combine autoencoding with DAMIP, an additional step in the process is needed. As in DAMIP different patients are compared to one another, the data should be discrete. The outcome from autoencoding, however, is continuous data. To account for this, we present two different variants of a discretization approach. We present a case study where we test the given methodology on data concerning breast cancer in older patients. This data consists of many different types of features, including administrative data, treatment data, and data concerning the tumour(s). All those different types of data were combined from different sources, making it a unique data set. The goal of the case study is to provide decision aid on the necessity of a post-surgery treatment and of chemo therapy in particular. We propose several approaches for this purpose, which are discussed later in this chapter.

The remainder of this chapter is organized as follows: an overview of relevant literature is given in Section III.2. After, the autoencoding model is described in Section III.3 and two methods for discretization in Section III.4. An overview of the full framework is given in Section III.5. The case study on breast cancer treatment decision aid is described in Section III.6, a discussion is provided in III.7, a note to practitioners with some suggestions on how this can all be used in practice is given in Section III.8 and finally a conclusion is given in Section III.9.

III.2 Literature review

In this section we present literature relevant to this chapter. A distinction is made between the theoretical part (autoencoding) and the application (medical decision support for breast cancer treatment). In the first section we present some recent applications of autoencoding with a variety of applications. In the second part we take a look at what has been done on the topic of providing decision support for breast cancer treatment.

III.2.1 Autoencoding

In [Y. Wang et al., 2016] the authors investigate the use of an autoencoder for dimensionality reduction. The theory of autoencoding is explained in this article as well as various other techniques for dimensionality reduction. Several methods are used for experimentation on synthetic data sets and real data sets. It is shown that the result of autoencoding is different from the other considered methods. The authors mention that autoencoding can, besides reduce the dimensionality, also detect repetitive structures, which is considered to be a good property for many applications.

In [Zong et al., 2018] the problem of unsupervised anomaly detection on high-dimensional data is considered. A Deep Autoencoding Gaussian Mixture Model (DAGMM) is introduced. This model first applies a deep autoencoder in order to generate a low-dimensional representation of the data. This data is fed into a Gaussian Mixture Model. The authors mention that in this model the parameters of the deep autoencoder and the mixture model are optimized simultaneously and that this helps the model to escape local optima. In the experiments done by the authors, they show that an F1-score of up to 0.927 is reached, which outperforms all the methods that are used for comparison.

In [Nousi and Tefas, 2017] a new type of autoencoder is presented. This discriminant autoencoder has as goal to increase the intra-class compactness and the inter-class separability. The proposed autoencoder is combined with nearest neighbors, nearest centroid, and multilayer perceptrons for classification. It is shown that the proposed model gives better results than denoising autoencoder on datasets concerning handwriting, facial expression recognition, and object recognition.

In [Shankar and Manikandan, 2019] the problem of 30-day hospital readmission is discussed. They state that, according to the Agency for Healthcare Reasearch and Quality (AHRQ), the United States alone has spent 41.3 billion dollars between January and November 2011 to treat patients readmitted within 30 days of discharge. The authors create a baseline using SVM and random forest. As their proposed method, they developed a deep neural network based on an optimized sequential architecture. The authors show that their method outperforms the baseline methods based on accuracy.

The authors of [Toğaçar et al., 2020] combine a convolutional neural network with an autoencoder in order to classify invasive ductal carcinoma breast cancer. The autoencoder model is used to reconstruct the data set and the discriminative

features were obtained from the convolutional neural network. Classification is done by linear discriminant analysis. The proposed method achieves accuracies of up to 98.59% and is considered successful.

One of the most common cancer types in the world, cervical cancer, is considered in [Adem et al., 2019]. The authors propose softmax classification with stacked autoencoder. In this method, first the stacked autoencoder is applied to the data set, which results in a data set of reduced dimension. After, the sofmax layer is used for classification. The proposed method is tested on a data set of 668 samples and the results are compared to those of state-of-the-art machine learning methods. The method suggested by the authors is shown to outperform the other methods, with an accuracy of 97.8%.

The authors of [Danaee et al., 2017] consider the problem of cancer detection from gene expression data, which typically concerns high dimensional data. A Stacked Denoising Autoencoder (SDAE) is proposed to extract features from the gene expression profiles. Next, supervised classification is used to verify the usefulness of the features for the purpose of cancer detection. The authors mention the usefulness of the proposed method for breast cancer detection and propose further research.

In [L. Wang et al., 2020] the authors apply an autoencoding technique in order to predict 30-day readmission. The method is tested on simulated data and compared to a logistic model with least absolute shrinkage and selection operator (LASSO) and to a random forest algorithm. The results show that generally the autoencoder outperforms the random forest algorithm and reaches similar results as the LASSO algorithm.

III.2.2 Medical decision support for breast cancer treatment

In [Nedungadi et al., 2018] an overview of different methods and directions of data-driven methods in precision oncology is given. Precision oncology concerns providing cancer treatment for each patient individually. The patient's genetic data, clinical data, environmental data, social data, and lifestyle data should be taken into account. There are many challenges in this field including large amounts of data, heterogeneous data and data coming from different sources such as electronic health records, clinical registries, medical imaging, demographics, wearables, and sensors. The authors mention that predictive models for cancer progression and survival, drug sensitivity and resistance, and identification of the most suitable combination of treatments for individual patients have been developed. As an open challenge, the authors mention the problem of precision medicine in clinical practice due to a lack of integrated systems. Moreover many of the existing clinical systems do not assist clinicians in providing precision oncology based recommendations. As a result the patients are still unable to benefit from this new knowledge for early diagnosis, prevention, or treatments.

The authors of [Stotter et al., 2015] propose a risk score to estimate 3-year survival of frail patients with early breast cancer. The data set used in this article consists of 328 patients between 43 and 98 years of age, the median being 82. The 3-year mortality rate is 29.6% of patients. Logistic regression is used to determine

the relationship between predictors and the 3-year mortality. The authors conclude that the risk score can support treatment planning and the communication of advice. This will be more and more necessary in the future, as an increasing number of older women is being diagnosed with early breast cancer.

[Hughes et al., 2004] researches early breast cancer in patients of 70 years and older. The question they pose is whether lumpectomy plus tamoxifen in combination with radiation therapy is more effective than just lumpectomy with tamoxifen. The data set analyzed consists of 636 women of 70 years of age and older, with stage 1 breast cancer. 317 of the patients were treated with lumpectomy with tamoxifen and radiation therapy and 319 of the patients had lumpectomy with only tamoxifen. In the results, the authors show that only one significant difference between the two patient groups was found, namely in the rate of local or regional recurrence at five years. One percent of the patients who had radiation therapy against four percent of the patients without radiation therapy. No significant differences were found with regard to distant metastases or five-year rates of overall survival. The authors conclude that lumpectomy plus tamoxifen without radiation therapy is a realistic choice of treatment in the case of patients aged 70 years and older with an early case of breast cancer.

In [Lo-Fo-Wong et al., 2015] an overview is given of studies on the subject of identification of predictors of health care use among women with breast cancer. Sixteen studies were included in the review and the types of health care the authors considered are hospital utilization and provider visits. It was found in the review that higher age, a more advanced cancer stage, more comorbid disorders, having a mastectomy, alymph node dissection, and breast reconstruction are all consistently associated with higher hospital utilization. The authors note that in the sixteen studies psychosocial and paramedical associations are rarely examined.

The research of [Clough-Gorr et al., 2012] focuses on older women with early stage breast cancer. The goal is to examine five- and ten-year survival based on cancer-specific geriatric assessment (C-SGA). The data set used in this study consists of 660 women who are 65 years or older and who are diagnosed with early stage breast cancer in the United States of America. The C-SGA is based on six measures: financial resources, comorbidity, obesity, physical function limitations, general mental health, and social support. The C-SGA is the sum of domain deficits, which varies from 0 to 4. It was found in this study that ten-year survival was consistently significantly lower for women with a C-SGA score of 3 or higher. Besides, survival rate decreases as the C-SGA score increases. The authors state that the death rate, both all-cause and breast-cancer-specific, was consistently approximately two times higher for women having a C-SGA score of 3 or higher. It is concluded that C-SGA may provide a means to guide treatment decision-making and to identify risk factors for intervention.

In [Handforth et al., 2014] the frailty of older cancer patients is examined by means of a literature review. From the data of 20 studies, including 2916 patients, it is found that more than half of the older cancer patients have pre-frailty or frailty. The authors mention that those patients are at an increased risk of chemotherapy intolerance, postoperative complications, and mortality.

The authors of [Ganggayah et al., 2019] explore multiple machine learning techniques to predict the survival or death of a breast cancer patient. The used techniques are: decision tree, random forest, neural networks, extreme boost, logistic regression, and support vector machine. The data to which those methods are applied consisted initially of 113 variables, of which 89 have been discarded based on expert insights. The remaining data consists of 23 independent variables and 1 outcome variable. The data consists of 8066 patients, of which approximately 70% survived and the remaining 30% died. Variable selection is done in R using the packages VSURF (variable selection using random forests) and randomForestExplainer. The highest prediction accuracy is achieved by random forest (82.7%). The variables found to be important in the prediction of survival of breast cancer patients are: cancer stage classification, tumour size, number of total axillary lymph nodes removed, number of positive lymph nodes, types of primary treatment, and methods of diagnosis.

The authors of [Mazurowski et al., 2008] study the topic of imbalanced training data in neural network classifiers for data-driven medical aid. They consider typical characteristics of medical data, small training sample size, large number of features, and correlation between features. Experiments are done using two methods of neural network training: classical backpropagation (BP) and particle swarm optimization (PSO). It is shown that even for slightly imbalanced data classification performance decreases. Furthermore, the authors show that BP is generally preferred over PSO in the case of imbalanced training data. Besides, it is shown that there is no clear evidence of better performance in case of oversampling in contrary to no compensation approach.

From this literature review we can see that some work has been done on autoencoding for medical applications and also some research was done on trying to provide data-driven approaches for cancer treatment decision aid. However, it seems that combining a large variety of classification approaches after using autoencoding for dimensionality reduction is still scarce. One aim in this chapter is to see which classification methods perform well after autoencoding. Besides, the specific problem of breast cancer treatment decision support was not widely found in literature and even less so with the use of an autoencoder. In this chapter we hope to achieve good classification results by applying said techniques to this specific problem.

III.3 Autoencoding model

An autoencoder is a type of neural network which learns how to encode data in such a way that when it is reconstructed, it is as close to the original input as possible. The autoencoder thus reduces the dimension of the data by learning to ignore the noise in the data. The main parts of an autoencoder are:

1. Encoder: compresses the input data into an encoded representation which reduces the dimension
2. Latent representation: contains the compressed representation

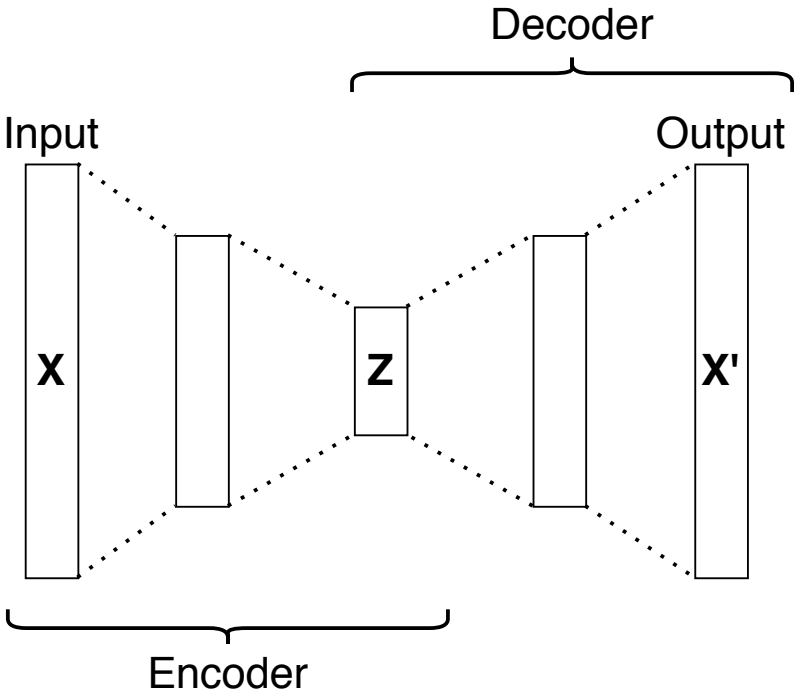


FIGURE III.1 – Representation of autoencoder

3. Decoder: reconstructs the data from the encoded representation

4. Reconstruction loss: measures the difference between the output and the original data

An overview of an autoencoder is given in Figure III.1.

In this chapter, we make use of a classical autoencoder. In this autoencoder two functions are used; the encoding and the decoding functions. The purpose of the encoding function is to transform a vector x from the input space into a new vector z in the latent space. In this new representation the dimensionality is reduced. The goal of the decoder is the opposite from the encoder function, namely to take the vector z and decode it back to the input space. The result is a new vector x'. The autoencoder is trained by minimizing the reconstruction error, that is, x' should be as close as possible to x. The goal is to keep the useful information from the input space, but with a reduced dimensionality.

In more detail, let $\mathbf{x}$ be the original input, which we assume to be n-dimensional. Similarly, let $\mathbf{z}$ be the new representation, which is m-dimensional, with $m < n$. If $\mathbf{x}'$ represents the reconstructed data, then we can define $\mathfrak{L}(\mathbf{x},\mathbf{x}')$ to be the reconstruction error. Moreover, we can define the latent variables as follows $\mathbf{z} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$, where σ is the activation function, $\mathbf{W}$ is the weight matrix, and $\mathbf{b}$ is the bias vector. Similarly, we can define the reconstructed data as $\mathbf{x}' = \sigma'(\mathbf{W}'\mathbf{z} + \mathbf{b}')$. σ', $\mathbf{W}'$, and $\mathbf{b}'$ of the decoder may be completely unrelated to σ, $\mathbf{W}$, and $\mathbf{b}$ of the encoder.

We will apply the autoencoder as is described above with a binary cross entropy loss function. The encoder and decoder functions consists of feed-forward, fully connected neural networks.

III.4 Discretization

In order to use the DAMIP model for classification, the encoded data needs to be discretized. The reason for this necessity is that the encoded data is in a continuous form. In the DAMIP classification model, patients are compared to each other, in order to make the classification. When we compare the continuous values, very few, if any, will be exactly equal. When we discretize the data, we create a specific number of possible values for each feature. In this case, patients where the continuous value were close, will now be considered equal. We consider two different versions of discretization; an equal-frequency method and an equal-width method. Below, descriptions of both variations are given. After, an example is given.

Equal-frequency discretization

In the equal-frequency method the data points are divided into bins where every bin will contain an equal amount of points. Within each bin, all data points get the same value, which is the mean of all data points in the bin.

The number of data points in each bin can be defined as follows. Let d be the total number of data points and let k be the number of bins desired. Let the result of $d\%k$ be n remainder r. Then, bins $1..r$ will have $n+1$ elements in it and bins $r+1..k$ will have n elements in it.

Equal-width discretization

In the equal-width discretization approach, the interval of all data points is divided into equal-width sub-intervals. Each bin gets the data points which lie within the sub-interval. Within each bin, all data points get the same value, which is the mean of the data points.

More formally, let k be the number of desired sub-intervals and let A and B be the minimum and maximum of the data points, respectively. Then, we can define the width w of the sub-intervals as $w = (B - A)/k$. The ith sub-interval becomes $[A + (i-1) * w, A + i * w]$ for $i = 1, 2, ..., k$.

Example of discretization

We present an example of how the two variations of discretization work. For both examples, we use the same data. A small example of the original data set is given in Table III.1. This data stems from the data set which is also used later in this chapter in our case study.

The first step is to apply the autoencoder to the original data. In the sample of data, we can see that there are three features. The autoencoder will reduce this to a specified number of features. In this example, we reduce the three features to one single encoded feature. That is, the three features are reduced into one. The result from this autoencoding step is given in Table III.2. The left table shows the encoded data, obtained from applying the autoencoder as it was described before.

Id	Breast	ATCD cancer	Polymedication
	Side of tumor 0 = left 1 = right	Antecedent cancer 0 = no 1 = yes	Use of multiple medications 0 = no 1 = yes
1	0	1	1
2	1	0	0
3	1	0	0
4	1	1	1
5	0	0	1
6	0	0	0
7	1	0	0
8	0	1	0
9	1	1	0
10	1	1	0

TABLE III.1 – Example of original data

The right table shows the same data, but ordered from small to large. Ordering of the data simplifies further steps in the discretization process.

Equal-frequency

In this example the number of data points d equals 10. Assume we want to divide those data points over 3 (k) sub-intervals. The result of $d\%k$ is $10\%3$ is 3 remainder 1 (n remainder r). So, bin 1 will have 4 elements in it and bins 2 and 3 will have 3 elements each. Within each bin, all values are set to the mean of the values. For example, in the first bin we get the values 0.013272, 0.084124, 0.122546, 0.200137. Calculating the mean and rounding it gives 0.11. Doing this for all three bins gives us the following binned data, shown in Table III.3.

Equal-width

We use the same example as for the equal-frequency method. Again, we are looking to divide the data points presented in Table III.2 in 3 (k) bins. This time we create sub-intervals of equal width. The minimum (A) of the data points is 0.013272 and the maximum (B) is 0.664676. The width (w) of the intervals thus becomes (0.664676 - 0.013272) / 3 = 0.217135. The sub-intervals then become [0.013272, 0.230407), [0.230407, 0.447542), [0.447542, 0.664677]. As a result, data points 1, 10, 8, 2, 5 are together in a bin, data points 3 and 4 form a bin and data points 7, 9, and 6 form the last bin. Like in the equal-frequency example, all values within a bin get the same value, which is their mean value. This results in the following discretized data points, shown in Table III.4.

Note that in fact, the specific value that are assigned to a bin does not matter. In the DAMIP classification model, values are compared to each other, so the important part is which data points get the same value, no matter what that value is.

Id	Feature 1
1	0.013272
2	0.200137
3	0.358410
4	0.429262
5	0.210265
6	0.664676
7	0.467684
8	0.122546
9	0.555402
10	0.084124

Id	Feature 1
1	0.013272
10	0.084124
8	0.122546
2	0.200137
5	0.210265
3	0.358410
4	0.429262
7	0.467684
9	0.555402
6	0.664676

Table III.2 – Example of encoded data

Id	Feature 1
1	0.11
2	0.11
3	0.33
4	0.33
5	0.33
6	0.56
7	0.56
8	0.11
9	0.56
10	0.11

Table III.3 – Binned data using the equal-frequency method

III.5 Framework

We have now seen all the separate parts of the autoencoder and classification framework. In this section, we give some insights in how the parts work together as a complete framework. For the classification part several methods are considered. First of all, we test with the DAMIP model, of which the full description can be found in Chapter II. Besides, we also consider other classification methods. Note that in the case of other methods for classification no discretization method is needed as those methods can handle continuous values. In the description of the framework we focus on DAMIP for classification.

The steps in the complete framework are as follows:

1. The autoencoding model is trained by minimizing the loss function (see Section III.3)
2. The data set is encoded using the trained autoencoding model
3. The encoded data is discretized using either the equal-frequency method or

Id	Feature 1
1	0.13
2	0.13
3	0.39
4	0.39
5	0.13
6	0.56
7	0.56
8	0.13
9	0.56
10	0.13

TABLE III.4 – Binned data using the equal-width method

the equal-width method (see Section III.4)

4. The DAMIP model is trained (see Chapter II)
5. Classification is made using the trained DAMIP model

III.6 Case study: breast cancer in older patients

III.6.1 Problem description

Breast cancer concerns about 54,000 new cases in France and 11,500 deaths per year (Source INCa). For the majority of cancers, the risk of being affected increases with age. Nearly 50% of the cases of breast cancer are diagnosed after age 65, of which more than 30% are older than 70 years. However, the major campaigns for breast and colon cancer screening, organized within the framework of public health, do not concern people over 74 years of age.

The process of aging is marked by progressive difficulties in the body's adaptation to stress, a gradual decline of health of various types and a higher occurrence of loss of autonomy. This process of aging is very variable among individuals and age is an insufficient criterion to evaluate the physiological state of a person. The aging process is characterized by a variable decline in organ function and the accumulation of comorbid medical conditions that can vary greatly between older people of the same age [Jolly et al., 2016]. Besides, Older cancer patients may have different and varied values, goals and preferences with respect to the trade-off between longevity and quality of life [Jolly et al., 2016]. This increases the difficulty of decision-making for the oncologist, who must integrate in his decision the complete benefit, life expectancy and tolerance to the treatment. A thorough geriatric evaluation, to identify medical problems and psychosocial and functional capacities, should provide appropriate geriatric health management, but cannot be performed on all elderly patients affected by cancer. Thus, the need for medical decision support tools to guide the choice of practitioners is important.

The goal of this study is to provide insight to the physician as to which post-surgery treatment should be given to a specific patient. The possible treatments we consider are: chemo therapy, radio therapy, hormonal therapy, and antibody therapy. Besides, combinations of the given treatments are also permitted.

III.6.2 Data

A large effort was made to create a highly complete data set containing a rich set of information about each patient. The data was collected from different sources. Firstly, we have the data from the medical record. This was combined with medico-administrative data, where the link between individual patients had to be made. Finally, information was collected and matched to each patient from doctor reports. Those reports were all hand-written and were digitalized to complete the data set.

The data set consists of 2048 breast cancer patients who are aged 70 and older and had a surgery. As we are interested in 5-year survival among patients, we only select those who have been followed for at least five years. For the patients we keep in the data set, we know either they died within five years or they survived at least five years. The patients for whom we do not have this information are deleted. 1128 patients remain in the data set. Among the patients who were followed for at least 5 years, 208 (18.4%) died within 5 years.

Treatment	**Occurrence**	**Mortality rate**
Chemo	15.6%	21.0%
Radio	77.3%	17.3%
Hormonal	75.2%	17.2%
Antibody	2.4%	14.8%
No treatment	9.0%	20.8%
Overall		18.4%

Table III.5 – Mortality rates after different treatments

In Table III.5 we can see the occurrences of the different treatment options as well as the mortality rate among patients having had this treatment. Note that the occurrences of treatments add up to more than 100% as patients can have multiple treatments. We can see that the mortality rate is the highest for chemo therapy. Probably this is due to the fact that patients who receive chemo therapy had a more severe case of cancer and thus a higher probability of dying within five years after surgery.

In Table III.6 we can see the mortality rate for the different SBR grades. The SBR grade is the Scarff-Bloom-Richardson grade and this is a measure to indicate the gravity of the cancer. A grade 1 is the least aggressive and 3 the most aggressive. The occurrences here do not add up to 100% as the SBR grade is not known for all patients. The effect we see is the one to be expected, when the cancer is more aggressive, the mortality rate increases.

In Table III.7 the occurrences and mortality rates for different health problems

SBR	Occurrence	Mortality rate
1	16.8%	9.5%
2	47.4%	15.9%
3	25.4%	33.2%
Overall		18.4%

TABLE III.6 – Mortality rates for different SBR grades

Health problem	Occurrence	Mortality rate
Previous cancer	15.2%	26.9%
Diabetes	13.0%	25.2%
Heart failure	9.8%	25.5%
Coronary artery disease	10.5%	28.8%
COPD	4.3%	22.4%
Overall		18.4%

TABLE III.7 – Mortality rates for several health problems

are shown. As we consider elderly patients, underlying health problems are rather common. It can be seen that in all cases of a health problem besides breast cancer, the mortality rate is higher than the mortality rate over the complete population, with coronary artery disease problems showing the highest mortality rate.

The data we have consists of administrative data, treatment data, disease characteristics, and biological data. All continuous variables are categorized based on expert opinion. All categorical variables with more than two categories are binarized by creating a binary column for each category. This results in 121 features in total.

III.6.3 Experimental results

In this section we look at the results stemming from applying the discussed methods to the data set on breast cancer treatment data. We consider several approaches, each with each own hypotheses. The different predictions we try to make are as follows:

— **Death within five years:** In this approach we take into account the treatment data of a patient. The question is whether, given the treatment, the patient will survive at least five years after surgery or not. The assumption made here is that if a patient survives at least five years, the treatment has been successful. For this approach we make use of the information, which is present in the data set, on the survival of death within five years of patients.

— **Treatment / no treatment:** Here the goal is to make a prediction on whether a treatment is necessary or not. In this case, this can be any treatment. Again, we make the assumption that the decision of treatment or no treatment has been correct if the patient has survived at least five years from the moment of surgery. For this prediction, we only make use of the patients

who have survived at least five years. The reason is that for the other patients, the decision might not have been the correct one. However, we cannot assume that the opposite decision would have been the correct one, as we cannot say anything about the outcome of a treatment situation which has not taken place.

— **Chemo / no chemo:** This approach is similar to the previous approach, with the difference that we look only at whether chemo therapy is necessary or not. This implies that whenever chemo therapy is deemed to not be necessary, another treatment option might still be of interest. We specifically look at chemo therapy, because the decision for or against it is often considered a difficult one.

— **Death after chemo:** In this final approach, we consider again the use of chemo therapy. However, this time we look at all patients who have had chemo therapy and we try to predict whether the patient has survived for at least five years or not. The purpose here is to see whether we can distinguish the patients who should not have had chemo therapy, i.e. the patients who have died within five years. Of course, we cannot assume that those patients would have survived for at least five years without chemo therapy, but this treatment option is a highly demanding one, so not applying this treatment if it's not predicted to be helpful has the preference.

III.6.3.1 Performance measures

For the outcomes of the experiments, we look at the performance measures accuracy, precision, recall, and F1-score, and AUC score. Those measures have been explained before in Chapter II. Like in Chapter II we will focus on the F1-score due to the imbalanced nature of our data.

All results are based on 5-fold cross-validation, where in each step 80% of the data is allocated for training and 20% for testing. The shown results are the averages over the 5 executions.

III.6.3.2 Prediction: Death within 5 years

Given all the information about the patient as well as all the treatment information, we try to predict whether a patient will die within 5 years from the date of surgery. Note that the positive class here consists of patients who have died within 5 years, as this is the minority class. The minority class covers 18.4% of the entities. The results of this experiment are given in Table III.8. This sub-group consists of 1128 patients, those are the patients of who we know whether they survived at least 5 years or not.

From the table we can see that DAMIP without feature selection performs poorly. This is most likely due to the high number of features, where many might not be relevant. Performance is substantially improved by applying Tabu for feature selection. Similar results are achieved by linear discriminant analysis, logistic regression and neural network. Moreover, combining autoencoder with DAMIP also reaches a similar achievement, specifically when using the equal-width binning technique for

Method	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0.862	0.750	0.308	0.436	0.643
Support Vector Machine	0.796	0.571	0.085	0.148	0.534
Naive Bayesian	0.249	0.204	0.977	0.337	0.525
Nearest Shrunken Centroid	0.681	0.360	0.681	0.471	0.681
Linear Discriminant Analysis	0.791	0.429	0.615	0.505	0.722
Logistic Regression	0.796	0.439	0.641	0.521	0.734
Neural Network	0.788	0.490	0.532	0.510	0.693
Decision Tree	0.751	0.385	0.455	0.417	0.639
DAMIP + PSO	0.739	0.356	0.495	0.414	0.645
DAMIP	0.824	0.950	0.049	0.092	0.524
DAMIP + Tabu	0.796	0.464	0.619	0.531	0.728
AE + DAMIP - equal freq	0.689	0.337	0.617	0.436	0.662
AE + DAMIP - equal width	0.796	0.463	0.595	0.521	0.719
AE + RF	0.813	0.507	0.419	0.454	0.660
AE + SVM	0.824	1.0	0.048	0.090	0.524
AE + NB	0.742	0.388	0.689	0.496	0.720
AE + NSC	0.728	0.372	0.690	0.483	0.712
AE + LDA	0.834	0.541	0.749	**0.625**	0.801
AE + LR	0.828	0.529	0.772	0.623	0.806
AE + NN	0.841	0.573	0.587	0.578	0.743
AE + DT	0.739	0.349	0.473	0.401	0.636

TABLE III.8 – Results for prediction of death within five years

discretization. The overall best result is achieved when autoencoder for dimensionality reduction is combined with linear discriminant analysis for classification.

III.6.3.3 Prediction: Treatment / no treatment

For the patients who have survived at least 5 years since surgery (920 patients), we assume that the decision of giving a treatment or no treatment was the correct one. Given the information about a patient, we try to predict whether they had a treatment (any treatment) or no treatment. The positive class here is the *no treatment* class as this is the minority class. This class constitutes to only 2.6% of the population. As this means that the testing data set consists of very few positive entities, we perform oversampling on the test data to make the results more reliable. Even with oversampling we do keep the same ratio of treatment to no treatment to keep the situation similar to the one in real-life.

In the considered population, there is a big class imbalance, with only 2.8% of the cases being positive. It seems that AE + DAMIP with both means of discretization have difficulty to achieve a good classification in this case. It might be caused by the discretization step between the autoencoding and the classification processes.

Method	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0.975	0.775	0.202	0.305	0.600
Support Vector Machine	0.976	0.874	0.150	0.251	0.575
Naive Bayesian	0.257	0.035	0.976	0.067	0.606
Nearest Shrunken Centroid	0.809	0.113	0.856	0.199	0.832
Linear Discriminant Analysis	0.971	0.484	0.422	0.417	0.704
Logistic Regression	0.980	0.703	0.574	0.609	0.783
Neural Network	0.968	0.441	0.602	0.488	0.790
Decision Tree	0.960	0.344	0.592	0.431	0.781
DAMIP + PSO	0.954	0.280	0.450	0.345	0.709
DAMIP	0.974	0.661	0.064	0.105	0.531
DAMIP + Tabu	0.979	0.795	0.310	0.446	0.654
AE + DAMIP - equal freq	0.969	0.368	0.210	0.268	0.600
AE + DAMIP - equal width	0.935	0.139	0.270	0.184	0.612
AE + RF	0.979	0.993	0.224	0.361	0.612
AE + SVM	0.973	0.400	0.008	0.016	0.504
AE + NB	0.939	0.210	0.398	0.271	0.676
AE + NSC	0.790	0.088	0.702	0.156	0.747
AE + LDA	0.983	0.680	0.756	**0.710**	0.873
AE + LR	0.984	0.850	0.514	0.634	0.755
AE + NN	0.980	0.659	0.604	0.625	0.797
AE + DT	0.972	0.472	0.318	0.367	0.654

Table III.9 – Results for prediction of treatment / no treatment

Potentially, the difference between the positive and the negative class is not captured by the discretization. When autoencoding is combined with other classification methods, the results are substantially better.

III.6.3.4 Prediction: Chemo / no chemo

For the patients who have survived at least 5 years since surgery, we assume that the decision of giving chemo therapy or not was the correct one. However, in this case we also assume that patients who did not survive at least 5 years and had chemo therapy, should not have had chemo therapy. Of course, it does not mean that the person would have survived at least 5 years when no chemo therapy was given, but given that chemo therapy is a very demanding treatment, not having chemo therapy has the preference above having chemo therapy. This subgroup consists of 958 patients. The positive class here is the *chemo* class, which constitutes 14.6% of the population.

From this table we can see that in general the results are rather similar as the ones in the two previous tables. However, in this case, the results of autoencoder in combination with linear discriminant analysis, and to some extent logistic regression,

Method	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0.859	0.467	0.269	0.341	0.611
Support Vector Machine	0.890	1	0.045	0.087	0.523
Naive Bayesian	0.297	0.183	0.968	0.308	0.568
Nearest Shrunken Centroid	0.750	0.369	0.774	0.500	0.760
Linear Discriminant Analysis	0.797	0.423	0.710	0.530	0.762
Logistic Regression	0.801	0.423	0.733	0.537	0.773
Neural Network	0.854	0.588	0.323	0.417	0.640
Decision Tree	0.792	0.385	0.484	0.429	0.667
DAMIP + PSO	0.835	0.400	0.469	0.432	0.680
DAMIP	0.857	0.400	0.018	0.034	0.509
DAMIP + Tabu	0.797	0.429	0.656	0.519	0.741
AE + DAMIP - equal freq	0.657	0.262	0.794	0.394	0.714
AE + DAMIP - equal width	0.785	0.392	0.690	0.500	0.746
AE + RF	0.857	0.516	0.468	0.488	0.696
AE + SVM	0.854	0	0	0	0.500
AE + NB	0.772	0.366	0.768	0.494	0.770
AE + NSC	0.736	0.326	0.768	0.457	0.749
AE + LDA	0.955	0.776	0.977	**0.864**	0.964
AE + LR	0.925	0.680	0.917	0.780	0.922
AE + NN	0.895	0.640	0.652	0.638	0.794
AE + DT	0.814	0.395	0.490	0.433	0.679

TABLE III.10 – Results for prediction of chemo / no chemo

are outstandingly good.

III.6.3.5 Prediction: Death after chemo

The population in this case consists of all patients who have had chemo therapy. The goal is to distinguish those who should not have had chemo therapy, i.e., those who died within 5 years after surgery. This group consists of 178 patients. The positive class is the group of patients who should *not* have had chemo therapy (i.e. those who had chemo therapy but died within five years) and 21.3% of the population belongs to the positive class.

The global performance of the algorithms is quite good. Again, we can see that the best result is achieved by autoencoder with linear discriminant analysis.

III.6.3.6 Chosen features

Even though the highest F1-score is not achieved by using our previously presented Tabu/DAMIP framework, this method has one important advantage. From our framework, we can see the subset of features which has been selected. This can

Method	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0.731	0.406	0.364	0.381	0.598
Support Vector Machine	0.792	0.400	0.056	0.097	0.524
Naive Bayesian	0.405	0.205	0.637	0.306	0.493
Nearest Shrunken Centroid	0.702	0.371	0.541	0.437	0.644
Linear Discriminant Analysis	0.736	0.368	0.355	0.356	0.594
Logistic Regression	0.725	0.405	0.389	0.379	0.599
Neural Network	0.731	0.396	0.450	0.414	0.627
Decision Tree	0.719	0.351	0.431	0.379	0.616
DAMIP + PSO	0.809	0.692	0.409	0.514	0.675
DAMIP	0.798	0.400	0.053	0.094	0.527
DAMIP + Tabu	0.806	0.625	0.556	0.588	0.722
AE + DAMIP - equal freq	0.843	0.636	0.412	0.500	0.678
AE + DAMIP - equal width	0.640	0.241	0.412	0.304	0.553
AE + RF	0.787	0.513	0.370	0.420	0.635
AE + SVM	0.797	0.400	0.070	0.116	0.535
AE + NB	0.742	0.403	0.582	0.472	0.682
AE + NSC	0.730	0.407	0.709	0.513	0.721
AE + LDA	0.865	0.658	0.811	**0.720**	0.848
AE + LR	0.775	0.497	0.693	0.575	0.744
AE + NN	0.764	0.487	0.575	0.519	0.694
AE + DT	0.669	0.305	0.456	0.322	0.598

Table III.11 – Results for prediction of death after chemo therapy

give us, and more importantly, the physicians insight in how a prediction has been made and which features are potentially important.

Table C.1 shows the features used by DAMIP/Tabu for each of the four described scenarios, to provide an overview. The features which are chosen seem vary quite a bit. This may indicate that for different predictions also different characteristics are relevant. However, there are also some features which seem to be important in general, they are chosen for all the four different approaches. In all cases the size of the cancerous tumor (TTUM) seems important. The same holds for estrogen receptors (RECELLUL), lymphocytes (a subtype of white blood cells), and dissection (CURAGE).

The chosen features were also presented to the involved oncologist, who has confirmed that, even though the features per case vary highly, largely make sense.

A limitation that we face in our approach, is that we do not obtain a ranking of the features. From the selection of features we obtain from our framework, we cannot see how or how strongly each one of them influences the prediction. This would provide a major benefit to physicians.

III.7 Discussion

We showed the results from four different approaches, each with its own goal of prediction. In all cases, the best result was achieved using the same method, autoencoding in combination with linear discriminant analysis. We have seen that in general for the DAMIP classification model, the results of the Tabu/DAMIP framework are better than those of autoencoding with DAMIP. On the contrary, for other classification methods, autoencoding does show an improvement. This may be due to the fact that in the case of DAMIP we perform a discretization step between the autoencoding and the classification, which may lead to a loss of information. We do not perform this step for the other methods, as they can handle continuous values. It seems that those classification methods profit from the fact that the dimensionality is reduced without loss of information.

Overall the best results were attained in the approach where we try to predict whether chemo therapy is necessary or not. The results obtained for this scenario are promising.

From the Tabu/DAMIP framework we retrieved the chosen features in each approach. The features chosen in each case varied, only a few features were found in each approach. It may imply that for each approach indeed different information about the patient is relevant. It seems that the few chosen in all four scenarios are generally important. As was mentioned before those are the size of the tumor, the estrogen receptors, lymphocytes, and dissection.

In real-life scenarios, the results we have obtained can be helpful to practitioners. First of all, the selection of features in all scenarios, and the most chosen features specifically, can give insights to the practitioners on what characteristics should be taken into account when making a treatment decision. This might give them new ideas on what to take into account or confirm their current decision-making process. Moreover, from the results we have seen before, we could see that we can make rather reliable predictions in different approaches. Especially predicting the necessity of chemo therapy seems to be quite achievable. By using those methods for prediction, the practitioners can get guidance in their decision, even when of course caution should be taken and the prediction should not be used without any reflection.

III.8 Note to practitioners

The decision on the treatment of a breast cancer patient is complex, especially for elderly patients it is found to be a difficult decision to be made. Such decisions are generally made in Multidisciplinary Consultation Meetings (MCMs). In such meetings many specialists come together to discuss the possibilities for the patient and to take a decision collectively. Many aspects have to be taken into account in the decision-making process, making it highly complex. Especially for elderly patients the decision is complicated as those might have additional health problems. A concrete example of a complex case may be a patient aged 78, who is obese and has cardiovascular problems and for whom the decision of chemotherapy has to be

taken.

In the results we have shown that reasonable predictions can be made concerning the outcome or the necessity of treatment for breast cancer in older patients. This information can aid practitioners in taking their decision. For example, in the case of predicting death within 5 years, the patient data in combination with a possible treatment option can be entered to the model and a prediction will be returned. By entering different treatment possibilities the practitioner can get insight in the expected outcomes of different treatments. Similarly, for all the other cases, the patient data can be entered and an insight in the necessity of a treatment or chemo therapy specifically will be returned. In all cases the model is trained on historical data, which can be done in advance. Only the data of the concerning needs to be provided to get a prediction instantly. However, caution should still be taken in using the model. Especially in the case of the autoencoder, which gives the best achieved result, it is important to take into account that the method is mostly a black box, from which we can hardly deduce how the prediction was made. In the case of DAMIP we do get the subset of features which was used to make the prediction, but this is still minimal information as we do not know the exact influence of a specific feature on the outcome. It is important to keep in mind that a model like presented in this chapter is meant as a tool to aid the clinician and not as a replacement.

III.9 Conclusion

In this chapter we considered autoencoding for dimensionality reduction. This approach was tested and compare to the situation in which we use feature selection. By using autoencoding, we can use the information of all variables, while still reducing the dimensionality of the data. The autoencoder was combined with different methods for classification, amongst which DAMIP. As DAMIP does not handle continuous values well, there was a need for a discretization step between the autoencoding and classification. We have considered two such methods, equal-width and equal-frequency discretization.

The different methods were tested on a real-life case study on breast cancer treatment decision aid. The goal of the case study was to see how our methods can help the physician in making a decision on post-surgery treatment. This problem is specifically relevant for elderly patients, as no general rules exist on treatment decision, since the physical state of those patients vary highly. We considered several approaches to achieve our goal. Firstly, we attempted to predict whether a patient will die or not within five years after surgery. In this case we assume that we know which treatment a patient has received. After, we tried to predict whether a patient needed any kind of treatment in order to survive at least five years. Similarly, we tried to make the same prediction, except this time specifically for chemo therapy, as this decision is considered particularly complex. Finally, we have considered all patients who have had chemo therapy and we tried to predict whether that has been a good decision. In all approaches we consider a treatment to be successful if the patient survives at least five years after surgery.

It was shown for all the mentioned approaches that generally the results of classi-

fication are improved when an autoencoder is applied before the classification step, which shows the use of dimensionality reduction. In all the different approaches the best result was achieved by using autoencoding for dimensionality reduction in combination with linear discriminant analysis for classification. The disadvantage of autoencoding is that we cannot see based on which features the classification was made exactly. This shows an advantage of the Tabu/DAMIP framework, where at least the subset of chosen features is given, even though no ranking is known. If classification performance is the only concern, autoencoding might be the better alternative.

The methods we have presented in this chapter can be used by physicians to support them in making their decision on the treatment to give a patient. Using the approach of predicting the death within five years of a patient, we can vary the considered treatment to see what the expected difference in death or survival of the patient is. Combining this with the other approaches, which take into account the necessity of any treatment or chemo therapy specifically, we can create a complete image of what will happen in different scenarios. This information may guide the physician towards a decision.

One direction that could be explored in future research is that of dealing with missing data. In the data set used in our case study, there were quite a few values missing. In our case, we binarized the data in any case, and we decided to simply put a value zero in each column. However, results may improve when applying a more intelligent technique to the missing data.

Another interesting research challenge would be to try and predict the length of survival of patients. Currently we only look at 5-year survival, but it may make an important difference to a patient whether the expected survival is 1 year or 4 years for example. Of course, also the quality of life is generally important for patients, however, this is very difficult to quantify and to use in a mathematical model.

Chapter IV

Performance Evaluation of Extended Hospital Stays for Readmission Prevention

Contents of the chapter

Abstract of the chapter

In this chapter we look into performance evaluation of extended hospital stays for readmission prevention. In previous chapters, we have looked into making accurate predictions. In this chapter we will use our Tabu/DAMIP framework for predicting hospital readmission and we will take a look what will happen if this is applied in a real-life situation, where measures can be taken when a readmission is predicted.

Firstly, we take a look at what has been done on the topic in literature. From this overview we can see that the link between the length of stay and hospital readmissions is not obvious and that evaluations made on methods to prevent readmissions are not abundant.

Next, we present the simulation model we use to simulate a real hospital environment. The first model, is the most basic model, in which no predictions are made yet. This represents the situation as it is currently. Next, we present our model in which we incorporate the use of readmission prediction. We try to avoid readmission by extending the hospital stay of a patient for whom readmission is predicted. If we predict that an extended stay will avoid readmission, the patient stays longer and the probability of readmission in the model decreases. In our third model, we add other policies which are meant to decrease the number of readmissions. Those policies include the use of a mobile application, home visits by a nurse, regular appointments with a general doctor, and regular appointments with a specialist. For those policies we have made estimates on the effectiveness.

In the results from the different models, we can see that we can indeed manage to decrease the number of readmissions if we extend the hospital stays of patients whom are expected to be readmitted. As expected, this comes at a certain costs, but it should be taken into account that a readmission is a highly undesirable event for both patients and the hospital. The inconvenience cannot be easily captured quantitatively, but is important to be taken into account.

IV.1 Introduction

In the previous chapters our goal was to develop a model which gives the highest possible F1-score on classification and we showed how those classification methods can be used to predict, for instance, readmission to the hospital. However, with this information we do not know yet what will happen if our model would actually be applied in real-life. In this chapter we will focus on performance evaluation of the classification framework we developed when it is used in a real-life case for hospital readmission. Hospital readmissions are often considered a quality measure of hospitals and are a big inconvenience to both patients and practitioners. Reducing

the number of readmissions gives an opportunity to simultaneously lower health care costs, improve quality of care, and increase patient satisfaction. We explore several possible scenarios, considering different options when we predict a patient to have a high chance of return. Our main interest is to see the effect of an extended hospital stay on readmission to the hospital. Besides this option for readmission prevention, we also consider several other options, consisting of a mobile application to follow a patient's lifestyle, visits to a medical specialist, home visits by a nurse, and visits to a general doctor.

The simulation model is applied on data covering five years of hospital admissions. We focus specifically on the digestive care unit, but this can be easily extended to other care units and other hospitals. We are interested to know the impact of different measures on readmission rates and the possibility of preventing those. Besides, we look how the suggested measures influence the total costs and readmission specific costs. The different options are also compared to the situation in which we make no prediction. This can gives us insights in the difference that the scenarios can make on hospital readmission.

The remainder of this chapter is structured as follows. First we consider relevant literature in Section IV.2. After, the different versions of the simulation model are presented in Section IV.3 and the corresponding results to a real-life case study are shown in Section IV.4. Finally, a discussion, a note to practitioners and a conclusion are given in Sections IV.5, IV.6 and IV.7, respectively.

IV.2 Literature review

In this literature review, we take a look at methods which have been tried before to avoid readmission and of which the efficiency were evaluated in literature. Besides, we consider the literature on the link between length of stay and hospital readmission.

In [Coffey et al., 2019] a systemic review of research on the avoidance of inappropriate hospital readmission is given. The articles considered by the authors show different results. Insufficient evidence was found for tele-health and long-term care interventions to be effective. The authors state that the most effective interventions include integrated systems between the hospital and the community care, multidisciplinary service provision, individualization of services, discharge planning initiated in the hospital and follow-ups by a specialist.

The authors of [Su et al., 2020] investigate the effect of two models, the LACE index and the HOSPITAL score, on hospital readmission. The results were produced using data from hospitalization data in Taiwan. It was shown that both models have the potential to decrease unplanned hospitalization, with the HOSPITAL score having the biggest potential.

In [Sun et al., 2017] a simulation model is presented which is used to evaluate healthcare facility utilization under various scenarios. In this model the patients are modeled as agents in an agent-based simulation. The time to readmission as well as the length of stay are determined using Bayesian models. The model was tested on Florida's Medicare and Medicaid claims data and the authors state that the model

was found to be effective in its evaluation.

The aim of [Alkhaldi and Alouani, 2018] is to reduce unplanned hospital readmissions. The authors propose a real-time patient-centric system, largely based on discrete-event system modeling and supervisory control theory. With this system, patients are supported via home monitoring, which implies a lower cost, while maintaining the quality of care. It is shown by means of simulation and analysis that the system is effective in reducing unplanned readmissions.

The authors of [Gupta and Fonarow, 2018] discuss the consequences of a healthcare policy in the United States. This healthcare policy was put in place in order to reduce 30-day readmission for patients with hearth failure. In this Hospital Readmissions Reduction Program (HRRP) hospitals with higher than average 30-day readmission rates were penalized. The authors state that the policy did not have the desired effect because readmissions were delayed beyond 30 days and inappropriate triage strategies were applied. Moreover, the authors mention that the HRRP was associated with an increased heart failure mortality.

The relationship between length of stay and hospital readmission is the subject in [Bjorvatn, 2013]. The authors specifically consider elderly patients in Norway. In this study it was found that in the period 1999 to 2006 the average length of stay in the hospital decreased, while the readmission rate increased. The analysis done by the authors shows that a longer length of stay is indeed associated with a lower probability of readmission. Besides, the patient's age, comorbidities, and the complexity of the treatment procedure have a positive correlation with the occurrence of readmission.

Contrarily to the previous article discussed, the authors of [Gay et al., 2019] study the link between length of stay and hospital readmission among children between the age of 0 and 18. It is mentioned that in this study no robust association between length of stay and readmission can be found and that it seems that the length of stay in children's hospitals are efficient.

From the given literature review we can see that the link between length of stay and hospital readmission is not obvious. Moreover, only few evaluations were made of the efficiency of different methods to avoid readmissions. Finally, it was shown that the number of readmissions should not be the only performance measure as undesirable consequences may be the result of certain policies with the only goal of preventing readmissions.

IV.3 Simulation model

In order to evaluate the added value of our developed methods, we create several simulation models. The different versions of the model are explained below. Firstly, we start by developing the basic model, representing the current situation. In this situation no prediction on possible return is made. After, we create the model where we incorporate readmission prediction. In this model, we can choose to extend the patient's hospital stay, based on the prediction. Finally, we try out different policies, other than an extended hospital stay, in order to try and limit the number of readmissions.

IV.3.1 Performance measures

The main goal of the simulation and of our prediction methods, is to avoid readmissions. Therefore, in assessing the performance of our method, we are mostly interested to see how many patients are readmitted. A secondary concern in measuring performance is the cost. Both of those performance measures can be divided in multiple parts, such as how a readmission was avoided and the cost from extra stays. Below, all the different performance measures are described.

IV.3.1.1 Number of readmissions

The goal of considering several scenarios in the simulation is to avoid readmission, as readmission of patients is highly undesirable, so we count how many patients return to the hospital in each scenario.

IV.3.1.2 Total cost

We keep track of the total cost in the simulation. In order to calculate the total costs, we make use of a general daily cost. This is an overall cost, taking into account the cost of surgery, use of resources, etcetera.

IV.3.1.3 Cost from readmissions

The most important disadvantage of readmission is the inconvenience of the patient and the hospital. However, there is also a financial loss, namely the cost of the patient having to stay during a certain time in the hospital.

IV.3.1.4 Cost from extended stays

We look at the costs made from a patient staying longer in the hospital than initially foreseen. This is the cost only of the extension of the stay, it does not include the initial part of the patient's stay.

IV.3.1.5 Cost from policy

For this performance measure we look at the costs stemming from applying a specific policy to prevent readmission.

IV.3.2 Basic model

An overview of our basic simulation model is given in Figure IV.1. The basis of the simulation model consists of patients arriving at the hospital (Patient arrival). They stay in the hospital for a specified amount of time (Initial stay), which is decided at their arrival. After their hospital stay, they leave the hospital (Patient departure) and they either return or not (Return?). We make the assumption that patients can only be readmitted once, so if the actual return of a patient is true the patient comes back to the hospital (Readmission) and after this stay, leaves the

simulation model. If the actual return was determined to be false, the patient leaves the system immediately.

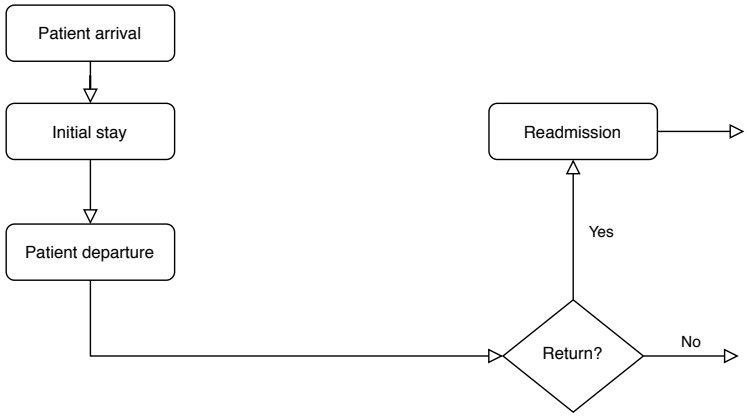


FIGURE IV.1 – Basic simulation model

The hypotheses used in this model are as follows:

— The population consists of 12088 patients. This is actual historical data, representing five years of hospital stays in the digestive care unit of a French private hospital. In each simulation run and in each version of the model, this population stays constant. A more extensive data description is given in Section IV.4.2.

— The arrival rate of patients at the hospital is set to be 7 patients per day. This roughly corresponds to 12088 in five years. We assume that a fixed number of patients is a relevant assumption as in the considered care unit the hospital stays are planned. However, do note that as we do not keep track of any resources, the time of arrival of patients is not crucial.

— Each patient has the same probability of readmission to the hospital, which is set to 23.3%. This is the percentage of patients who returned in the actual dataset.

— We assume that patients can only be readmitted once, that is, after their readmission surely leave the simulated system.

— In the basic model no predictions are made on the readmission of a patient, a patient thus stays their initially decided length of stay, which is retrieved from the actual data.

— Patients who are readmitted to the hospital after their initial stay have a fixed length of stay for their second stay. From the historical data we can see that the average length of stay of a readmission is 6.8 days. As we work only with complete days, this is rounded to 7 days.

— The cost of a patient staying in the hospital for one day is set to 1067 euros, which is based on advise from a healthcare professional. This cost is the general cost including surgery and resource occupation, such as doctors, nurses, and beds. The cost per day is set to the same value for initial stays as for readmissions.

The same model as described above is also given in detailed pseudo code in Algorithm 2. The input here are the set of patients, with their corresponding length of stay, the daily cost for a patient staying in the hospital, the probability of readmission, and the average length of stay at a readmission. After performing all steps as describes, the model gives as a result the number of readmissions, the cost related to readmissions, and the total costs.

Algorithm 2 Basic model

Input: patients P with their length of stay *los*, daily cost *cost_d*, probability of readmission *read_prob*, average length of stay of a readmission *avg_los_read*

Output: number of readmissions *nb_read*, readmission cost *read_cost*, total cost *total_cost*

```
1: nb_read := 0 ;
2: total_cost := 0 ;
3: read_cost := 0 ;
4: for p in P do
5:    total_cost += p.los * cost_d ;
6:    if p return with probability read_prob then
7:       nb_read++ ;
8:       total_cost += avg_los_read * cost_d ;
9:       read_cost += avg_los_read * cost_d ;
10:   end if
11: end for
```

IV.3.3 Model with prediction

In this version of the model we consider the same start as in the basic model, with patients arriving in the hospital and staying a specified amount of time. This time however, at the end of their stay a prediction on their return is made. If it is predicted that the patient will not return to the hospital, the patient can go home. On the other hand, if it is predicted that the patient needs to be readmitted, we make the same prediction, but with a longer hospital stay, to see if we expect readmission to be avoidable by extending the stay. If this is indeed the case, the patient stays longer in the hospital and is then sent home. If we expect that a longer stay does not make a difference, the patient is sent home without an extra long stay in the hospital. This decision is visualized in Figure IV.2, where a prediction value 1 means that readmission is expected and 0 means readmission is not expected. In the figure the process is shown for the possibility of extending the hospital stay with at most three days, however, this is of course easily extendable to any number of days. Besides making the prediction with an extended length of stay, we can in this prediction also include updates of other variables, for example on the medical condition of the patient, by re-testing.

After the patient has left the hospital, the actual return is determined. Patients who do not return, leave the simulation model. The others return to the patient

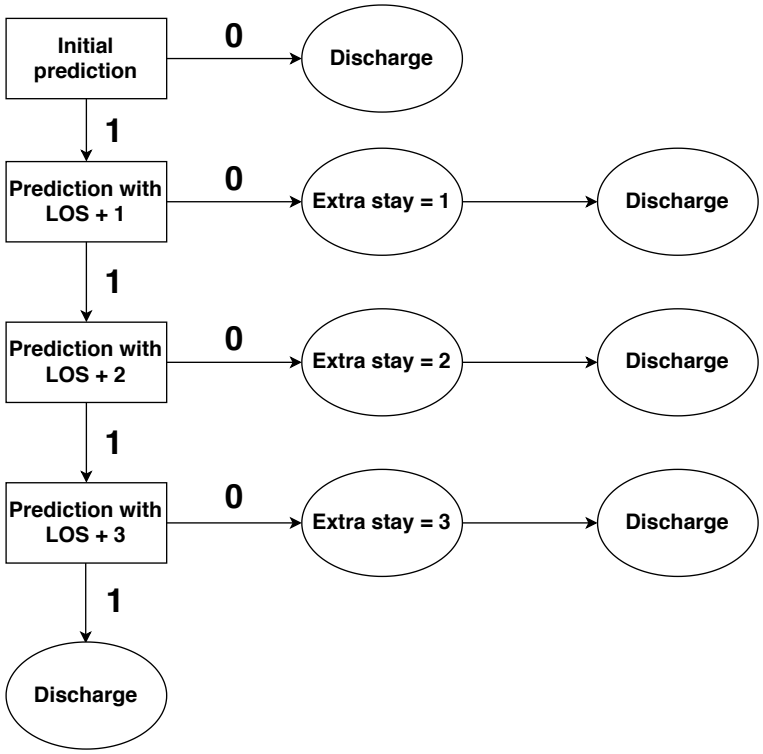


FIGURE IV.2 – The prediction procedure in the simulation model.

arrival and the same process as described is executed again. This model with prediction is shown graphically in Figure IV.3. Note that similar to the basic model, here we also assume that only a single readmission can take place, so after a readmission the patient leaves the system.

In this model the probability of the actual return of patients is dependent on the prediction that was made. The given data set is used to run our Tabu/DAMIP framework. The result of this, gives us a selection of features, which will be used in the prediction, as well as a confusion matrix. This confusion matrix will be used in the simulation model for deciding on the actual return of patients. The True Positive score (TP) will be the probability that a patient will indeed return, when we predicted he will return. Similarly, the False Negative score (FN) is the probability that a patient will actually return, while we predicted that he will not return.

As we focus on the impact of the length of stay on readmission, it is important that among the features used in the prediction is the length of stay. It is thus forced in the feature selection process that length of stay is one of the chosen features. The features which were chosen using tabu search for feature selection are the following:

In the simulation, we use the actual historical data for the incoming patients. To determine an actual return we use the confusion matrix given by 5-fold cross validation of the Tabu/DAMIP framework, as we do not have the information of actual return when we increase the length of stay of a patient. The percentages retrieved from the confusion matrix is as follows. If we predict that a patient will not be readmitted, the probability of actual return is 16.3%. When we predict that a patient will be readmitted, the probability of this happening is 35.0%.

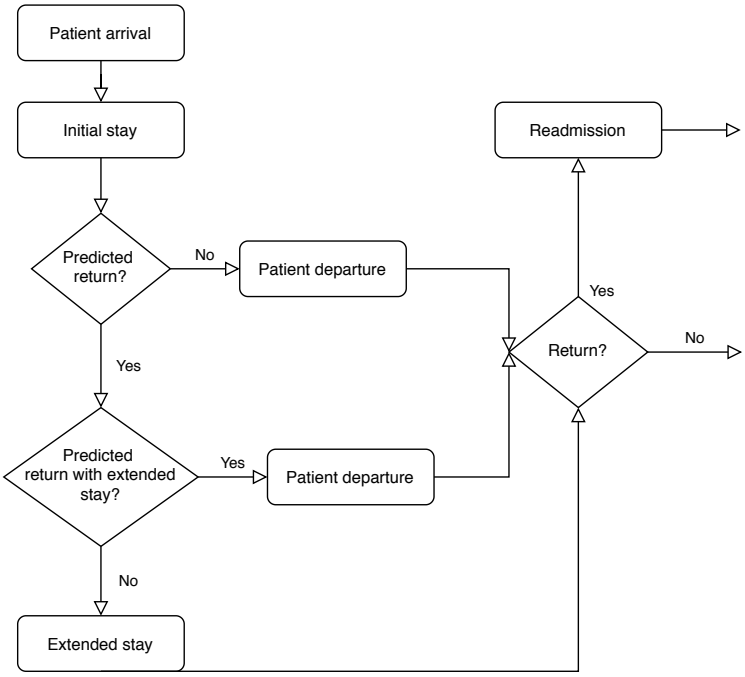


Figure IV.3 – The simulation model where the prediction of return is taken into account.

The whole simulation of the model with prediction is given in pseudo code in Algorithm 3.

IV.3.4 Policies to avoid readmission

In this section, we will look at the possibility of adding certain policies to the simulation model. The goal of those policies is to reduce the risk of return, while the patient does not have to stay longer in the hospital. This carries advantages for both the hospital system as well as for the patients themselves.

IV.3.4.1 Home visits by nurse

By having a nurse visiting a patient's home regularly the chance of the patient needing to go back to the hospital is assumed to decrease. In this scenario, we plan home visits by a nurse for patients with a predicted return. The effect on the costs is explored.

IV.3.4.2 Regular doctor appointments

Similar to the visits of a nurse at the patient's home, we can schedule regular appointments with a doctor as well to decrease the chances of readmission. Two options are possible here, appointments can be made either with a general doctor or with a specialist.

Algorithm 3 Model with prediction

Input: patients P with their length of stay *los*, daily cost *cost_d*, confusion matrix *CM*, average length of stay of a readmission *avg_los_read*

Output: number of readmissions *nb_read*, readmission cost *read_cost*, total cost *total_cost*, extended stay cost *estay_cost*, maximum number of extra days *max_edays*

```
1: nb_read := 0 ;
2: total_cost := 0 ;
3: read_cost := 0 ;
4: estay_cost := 0 ;
5: for p in P do
6:    total_cost += p.los * cost_d ;
7:    if p.pred_read(los) == true then
8:       stop := false ;
9:       for i from 1 to max_edays and stop == false do
10:         if p.pred_read(los + i) == false then
11:            total_cost += i * cost_daily ;
12:            estay_cost += i * cost_daily ;
13:            p.prob_read = CM(FN) ;
14:            stop = true ;
15:         end if
16:      end for
17:      p.prob_read = CM(TP) ;
18:   else
19:      p.prob_read = CM(FN) ;
20:   end if
21:   if p return with p.prob_read then
22:      nb_read++ ;
23:      total_cost += avg_los_read * cost_d ;
24:      read_cost += avg_los_read * cost_d ;
25:   end if
26: end for
```

Feature	Description
LOS	length of stay
age_26_50	age between 26 and 50
unitesoin_UROL	urology care unit
unitesoin_GAST	gastroenterology care unit
diagnostic_principal_firstLetter_J	respiratory problem
diagnostic_principal_firstLetter_S	injury or poisining
diagnostic_principal_firstLetter_E	endocrine, nutritional and metabolic diseases
diagnostic_principal_firstLetter_A	infectious and parasitic diseases
diagnostic_principal_firstLetter_Q	congenital malformations, deformations and chromosomal abnormalities
diagnostic_principal_K57	diverticular disease of intestine
diagnostic_principal_K29	gastritis and duodenitis
diagnostic_principal_K63	other diseases of intestine
diagnostic_principal_C20	malignant neoplasm of rectum
diagnostic_principal_L05	pilonidal cyst and sinus
diagnostic_principal_K42	umbilical hernia
diagnostic_principal_K62	other diseases of anus and rectum
diagnostic_principal_C25	malignant neoplasm of pancreas
diagnostic_principal_D37	neoplasm of uncertain behavior of oral cavity and digestive organs
diagactes_YYYY300	imaging for interventional radiology
diagactes_GLLD002	discontinuous mechanical ventilation
diagactes_HFFC018	sleeve gastrectomy
diagactes_YYYY028	ultrasound guidance
diagactes_HFCC003	gastric bypass in Y for morbid obesity
diagactes_HMQH008	intraoperative cholangiography
diagactes_YYYY145	radiological examination of the gallbladder and biliary tract
diagactes_YYYY115	CT guidance

Table IV.1 – Selected features

IV.3.4.3 Mobile application

Using a mobile application developed to track the life style of patients, a doctor can, from a distance, keep an eye on the (un)healthy behaviour of a patient. In case of worries, the doctor can contact the patient to discuss the seen behaviour and the doctor can give a suitable advice.

The cost of the different options have been suggested by a healthcare professional. However, the impact of the policies on the probability of return is not clear. In the

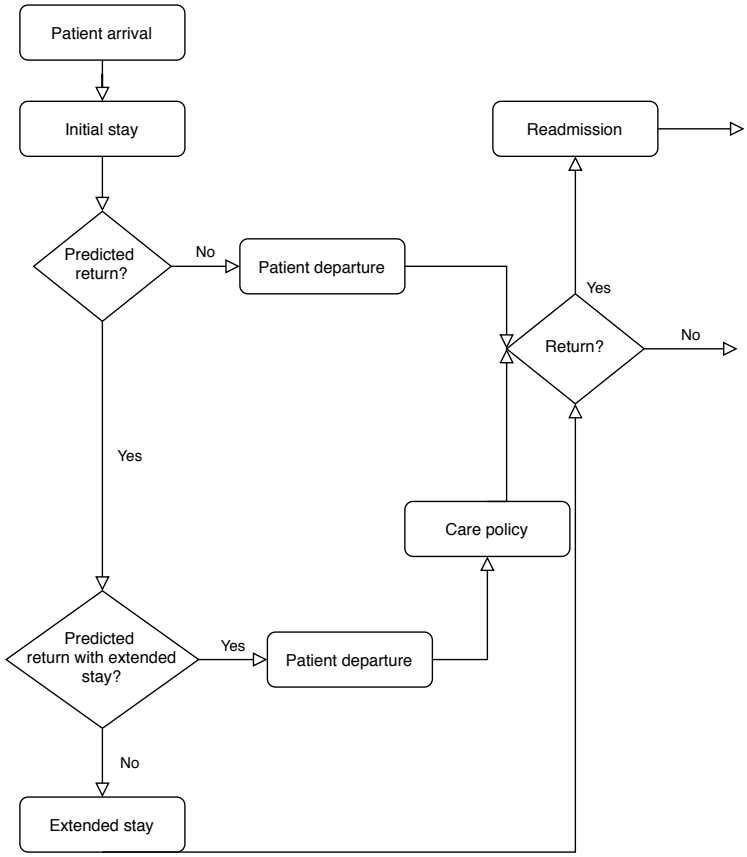


FIGURE IV.4 – A graphical overview of the simulation model extended with care policies

simulation model we make use of estimated values, where the order of impact is based the costs, the most expensive policy is assumed to have the biggest impact on the probability of return. Of course, this assumption might not hold true in real-life, but it seems a fair assumption given the lack of actual information. The different policies are given in Table IV.2, with their associated cost and estimated impact on the probability of return. The impact factor is the factor by which the probability of return is reduced. That is, an impact factor of 0.1, decreases the probability of return by 10%.

Policy	**Cost**	**Impact factor**
Mobile application	50	0.1
Specialist	60	0.15
Nurse	75	0.2
General doctor	100	0.25

TABLE IV.2 – Policies with cost and impact

In Algorithm 4, the precise procedure is described. This is similar to the previous model, with prediction but without policies. The difference between the two is that, in case we predict that a longer hospital stay will not prevent a readmission, a care policy will be put in place. That means that in the algorithm in this case the

probability of return is adjusted and the costs need to be updated.

IV.4 Case study: return to digestive care unit

IV.4.1 Problem description

For this case study we consider the digestive care unit of a French hospital. In this care unit, we can assume that all the initial hospital stays were planned and that the arrival rate of patients is fairly constant. On the other hand, we assume that none of the readmissions were planned and that they can be avoidable. We investigate the use of extended hospital stays to avoid a patient having to go back to the hospital after discharge. Besides, we consider the possibility of several policies where the patient does not have to stay in the hospital, but alternative measures are put into place. For those different scenarios we look at the number of readmissions and the cost for the hospital. Additionally, we consider a basic simulation model in which nothing is done to avoid readmission. The results of the different models are compared to get insight in the possibilities of readmission reduction for the hospital.

IV.4.2 Data

The data covers five years of stays at this unit. The raw data consisted of a lot of text and many diagnoses and medical acts in a single column, implying the need for thorough data treatment. All those diagnoses and medical acts were extracted and binarized. That is, for each existing diagnosis and medical act, a column is created where a patient gets value 1 if he had this diagnosis or medical act and 0 otherwise. This is the same as was done in Chapter 2.

From the hospital data, only the patients who stayed at the digestive care unit during their stay are kept. This results in 12088 hospital stays. In the data set we have a total of 189 columns.

In Figure D.1 we can see the diagnoses which are present in our data and their relative occurrence. The diagnoses are represented by ICD-10 codes, it contains one letter, followed by one to five numerals, and optionally a seventh character, which is a letter. To reduce the number of possible diagnoses, the codes have been reduced to in total three characters: one letter followed by two digits. The different medical procedures in the data set are represented by CCAM codes. A procedure is represented by a code consisting of four letters followed by three digits. In Figure D.2 we can see an overview of the medical acts and their occurrences.

In the following figures we explore the relationship between length of stay and readmission. In Figure IV.5, we can see the length of stay plotted against the percentage of readmissions for the digestive care unit. The different lines indicate the readmissions for 30, 90, and 180 days.

From the figure we can see that the percentage of readmission decrease slightly when going from 0 days to a single day hospital admission. However, after that, there is an increasing trend. This does not imply that a longer length of stay causes a higher chance of readmission though. There would rather be a more indirect

Algorithm 4 Model with prediction and policy

Input: patients P with their length of stay *los*, daily cost *cost_d*, confusion matrix *CM*, average length of stay of a readmission *avg_los_read*, cost of policy *cost_p*, impact factor of policy *impact_factor*, maximum number of extra days *max_edays*

Output: number of readmissions *nb_read*, readmission cost *read_cost*, total cost *total_cost*, extended stay cost *estay_cost*, policy cost *policy_cost*

```
1: nb_read := 0 ;
2: total_cost := 0 ;
3: read_cost := 0 ;
4: estay_cost := 0 ;
5: policy_cost := 0 ;
6: for p in P do
7:    total_cost += p.los * cost_d ;
8:    if p.pred_read(los) == true then
9:       stop := false ;
10:      for i from 1 to max_edays and stop == false do
11:         if p.pred_read(los + i) == false then
12:            total_cost += i * cost_daily ;
13:            estay_cost += i * cost_daily ;
14:            p.prob_read = CM(FN) ;
15:            stop = true ;
16:         end if
17:      end for
18:      total_cost += cost_p ;
19:      policy_cost += cost_p ;
20:      p.prob_read = (1 - impact_factor) * CM(TP) ;
21:   else
22:      p.prob_read = CM(FN) ;
23:   end if
24:   if p return with p.prob_read then
25:      nb_read++ ;
26:      total_cost += avg_los_read * cost_d ;
27:      read_cost += avg_los_read * cost_d ;
28:   end if
29: end for
```

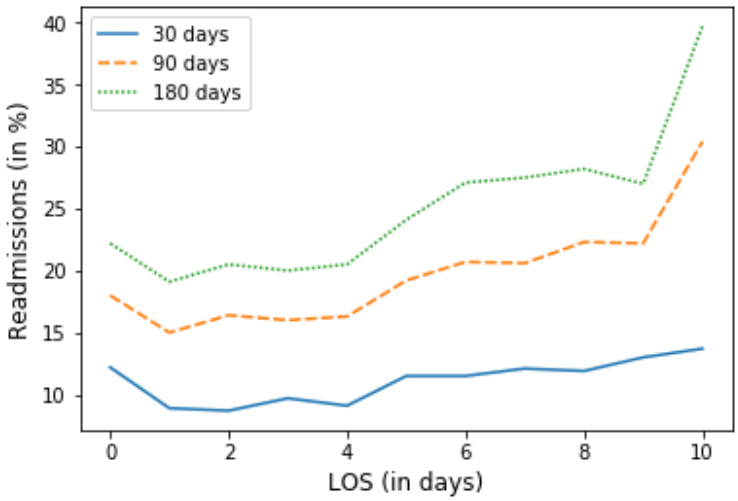
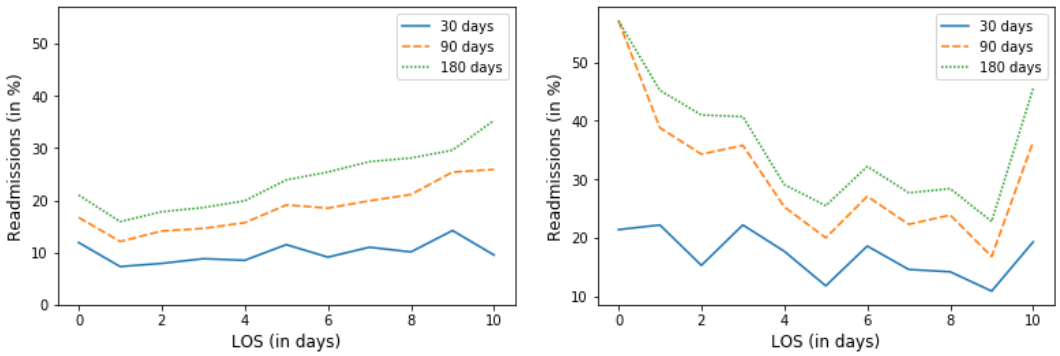


Figure IV.5 – Readmission at the digestive care unit

relationship, where patients with a worse medical condition have a longer length of stay, but also a higher probability of return. It may help if we can find more homogeneous groups of patients to see a different relationship between length of stay and readmission. In the following graphs we make several attempts.

In Figure IV.6 a distinction is made between patients who have a cancer and those who do not. For both groups we look at the percentages of readmission for different lengths of stay.

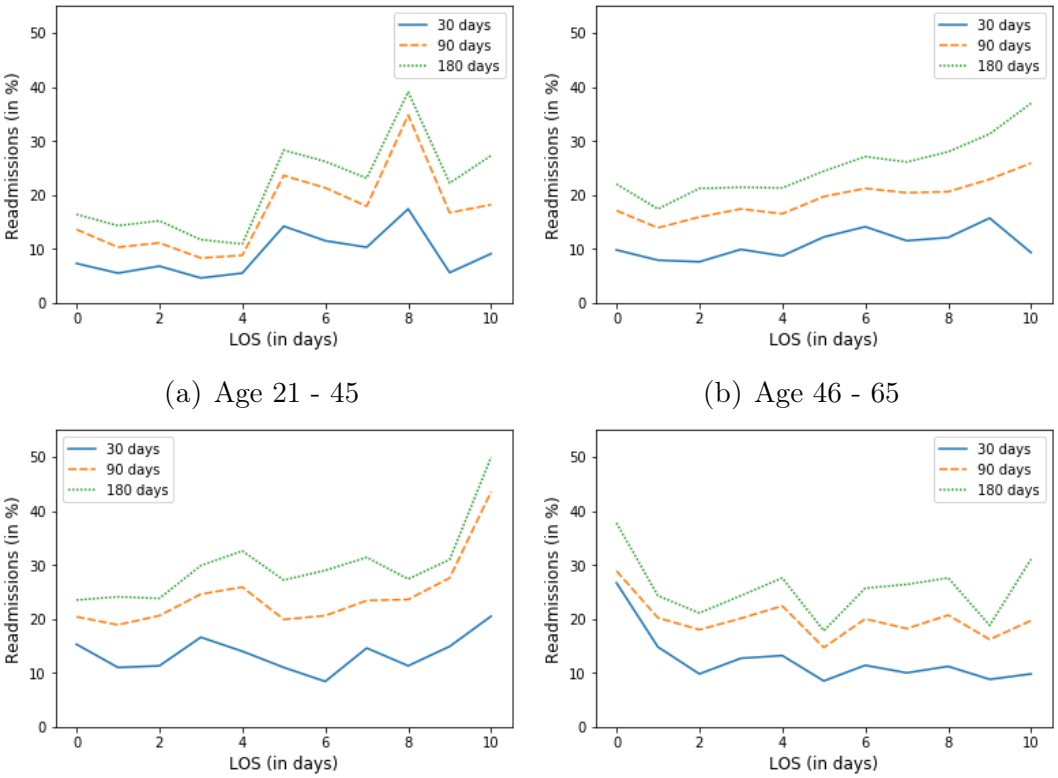


(a) Non-cancer (b) Cancer

Figure IV.6 – Graphic overview of hospital readmission data

It is remarkable to see that the two groups show rather opposite trends. In the group of patients where people do not have any cancer, we see a similar trend to the complete population, a slightly increasing trend. However, for the group of patients with a cancer, it is quite the contrary. Generally, for this group of patients, with a longer duration in the hospital the percentage of readmission decreases.

Besides a difference in whether or not a patient has a cancer or not, we also look at the different age groups of patients. The graphs of different age groups are shown in Figure IV.7. Note that the age group from 0 to 20 years is missing, in this age group there are not enough patients per length of stay to give a reliable indication of readmission rates.



(a) Age 21 - 45

(b) Age 46 - 65

(c) Age 66 - 75

(d) Age 76+

FIGURE IV.7 – Graphic overview of hospital readmission data for different age groups

From the different age groups we can see that, even though there are some differences between the graphs, they generally show a slightly upward or a stable trend.

Even though we cannot see a downward trend of readmission in most of the presented graphs, we still assume that a longer length of stay may reduce the probability of readmission. The reason for this is that in the given graphs, we compare patients with a certain length of stay to different patients with a different length of stay. An upward trend in the readmission percentage may indicate that the patients with a longer length of stay were in a worse situation or had a more complicated medical condition and therefore were readmitted more often. We thus assume that if a given patient stays longer in the hospital, the probability of readmission is reduced.

IV.4.3 Experimental results

As was mentioned before, in the simulation model we use the actual data of patients as input. We use the data of all the 12088 patients in the data set. The daily cost, which includes everything ranging from occupation of resources to surgery, is set at 1067 euros. This value is specific to this case study and thus specific to the digestive care unit of a French hospital.

IV.4.3.1 Basic model

The results are as follows. On average 2814.62 (23.3%) patients return to the hospital. The 95% confidence interval is less than 1% below or above this average. There is so little variation in the simulation runs as the only random part in the simulation is the actual return of a patient. This is true for all results mentioned in this chapter.

The costs found in this simulation are shown in Figure IV.8. In this figure we can see both the total cost and the cost caused by patients who return to the hospital. The costs are shown in millions of euros. Note that the cost from return is also included in the total cost. The cost from return is the cost that is made from the extra stay of return patients, their initial stay is not included in this cost. We can see that approximately 25% of the total cost comes from patients who return to the hospital.

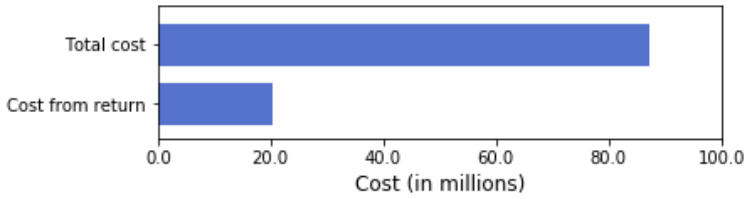


Figure IV.8 – Costs in the basic model

IV.4.3.2 Model with prediction

Now we add the prediction step to the model. As it was described before, in this step we make a prediction on whether we expect the patient to return or not. If we predict the patient not to return, the patient can leave the hospital. In the other case, we make a new prediction, but this time with a longer stay in the hospital. If we find there that we think the patient will not return, the patient stays more time in the hospital. Again, for the actual return, we make use of probabilities. Those probabilities are deduced from the confusion matrix obtained from training and testing the DAMIP model. We arrive at the following probabilities: if we predict a patient to return to the hospital, the patient has a probability of 35.0% to actually return to the hospital. If we predict the patient not to return, the probability of returning is 16.3%. This also implies that the probability of return drops when we predict that with an additional stay, the patient does not return. As the DAMIP model is pre-trained and the parameters are determined, the prediction is the same each time the model is ran, when all parameters are equal. This implies that in the results the number of patients who stay longer in the hospital remains the same for multiple simulation runs. The cost and the number of patients who actually return do vary. We vary the maximum number of days that a patient can stay extra, to see how it affects the KPI's. The results are shown below.

In Figure IV.9 we can see the number of patients who are readmitted. In creating those results the maximum number of extra days a patients can stay was varied. That is, with the maximum extra stay equal to 1, patients can stay at most 1

day extra after their initial length of stay. For the maximum extra stay equal to 2, patients can stay either 1 or 2 days after the initial length of stay, and so on. We can see that if we increase the maximum number of extra days, the number of patients who return is decreased. In the basic model, without any additional stays, approximately 2814 patients were readmitted. If we increase the maximum number of additional days to 7, we can reduce the number of readmissions by almost 200.

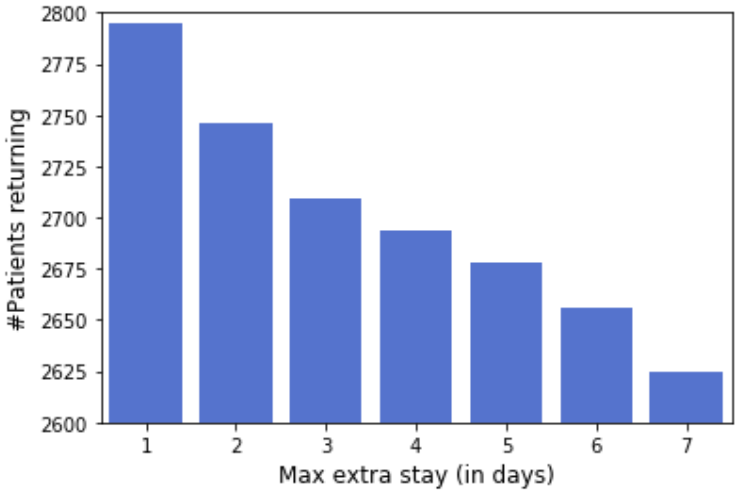


FIGURE IV.9 – The number of patients readmitted for a different maximum extra length of stay

In Figure IV.10 it is shown how many patients had their hospital stay extended for a different number of maximum extra days. For each number of maximum extra days, the total number of patients staying longer is shown. From the figure we can see that, logically, the number of patients who stay longer increases as we increase the maximum number of days that an extended stay may last.

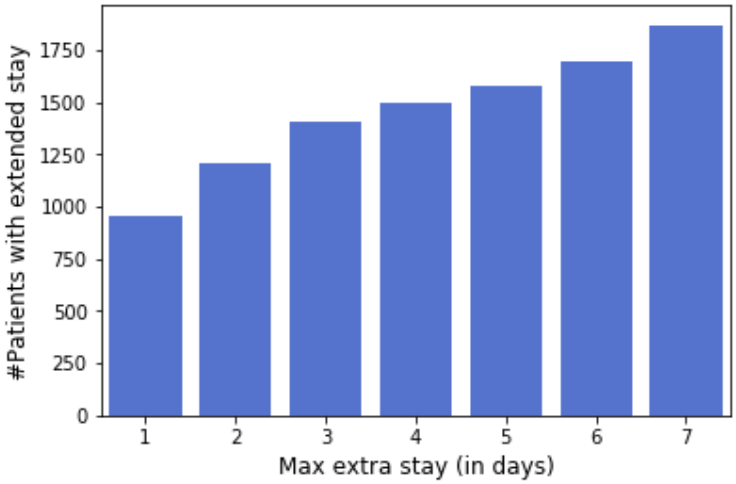


FIGURE IV.10 – The number of patients for whom their hospital stay was extended

Figure IV.11 shows the total costs in each of the different scenarios. The costs are shown in millions. We can see from the figure that, even though the number of patients who return to the hospital decreases with extended stays, the total costs

for the hospital increase. If we compare the increase of the costs to the decrease of the readmissions, we can see that the costs are increased by 3.2% for a decrease of 6.1% of readmissions.

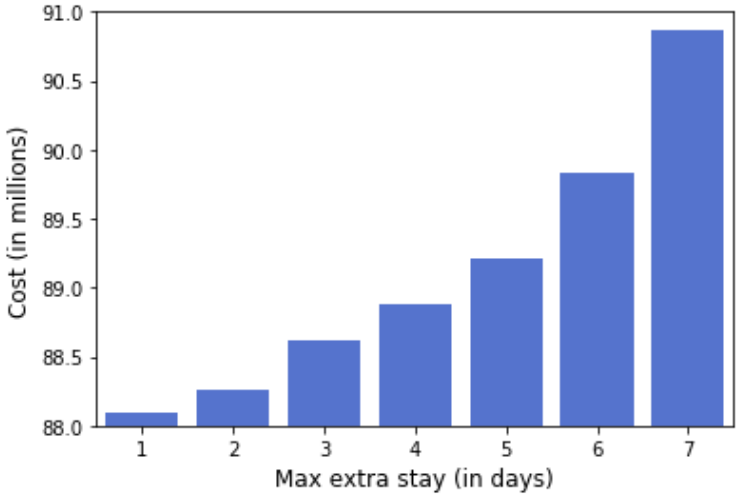


Figure IV.11 – The total costs

The costs stemming specifically from the extended hospital stays are shown in Figure IV.12. Logically, those costs increase as more people will stay longer when the maximum number of extra days is increased and the extended hospital stay is on average longer when the maximum allowed is higher.

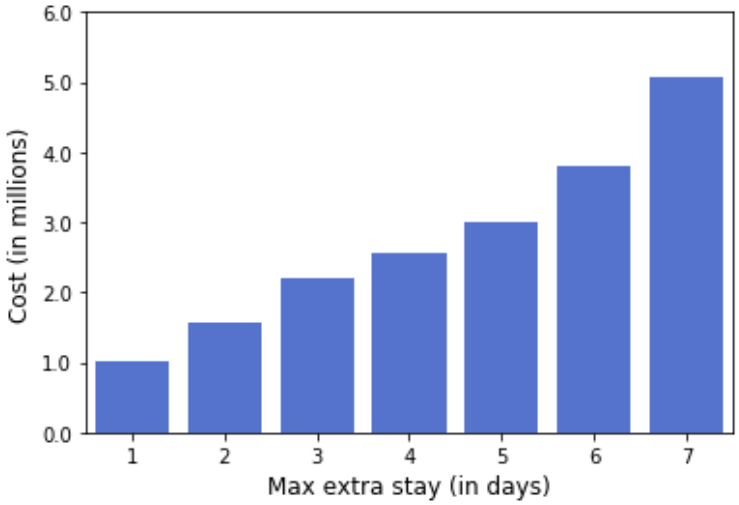


Figure IV.12 – The costs stemming from extended hospital stays

Contrarily to the previous figure, in Figure IV.13, we can see that the cost stemming from readmissions decreases if we allow patients to stay longer in the hospital.

IV.4.3.3 Policies to prevent return

In this section we look at the results of the simulation model in which certain policies are implemented with the goal of further reducing hospital readmissions.

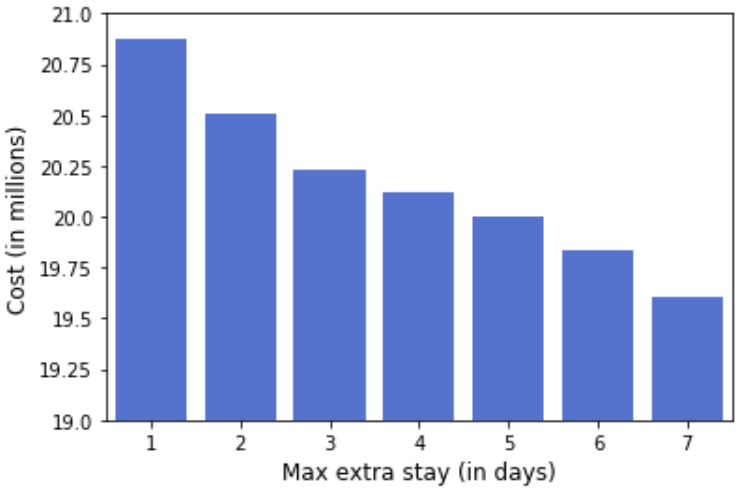


FIGURE IV.13 – The costs stemming from return patients

As was mentioned before, the costs for the different policies were indicated by a healthcare professional. However, the impact of the different policies is not clear and we had to estimate them. The different policies with their associated costs and assumed impact factors are given below in Table IV.3.

Policy	**Cost**	**Impact factor**
Mobile application	50	0.1
Specialist	60	0.15
Nurse	75	0.2
General doctor	100	0.25

TABLE IV.3 – Policies with cost and impact

Firstly, we look at the number of return patients per policy and in the situation without any policy. Note, that we keep the maximum number of extra days constant in the results in this section. This number is fixed at 7. In Figure IV.14 the number of readmissions are shown for each situation. We can see the influence of the different impact factors. Clearly, those results are highly dependent on the chosen impact factors. As those were chosen by estimation, some uncertainty is present.

In Figure IV.15 we can see the total costs when a specific policy is applied. We can see that the total costs actually decrease for more expensive policies. This is caused by the reduced number of readmissions.

In Figure IV.16, which shows the cost related to the policy for each specific policy, we can see that, logically, the cost related to the application of the policy becomes higher as the policy becomes more expensive.

Figure IV.17 shows the costs related to the readmission of patients. As we have seen before, the number of readmissions decreases when a more expensive policy is applied. Clearly, this implies that the costs related to the readmissions decrease as well.

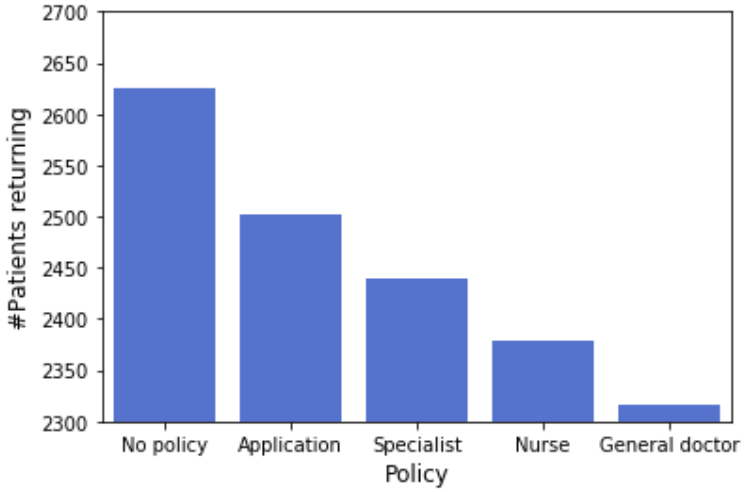


Figure IV.14 – The number of readmissions per policy

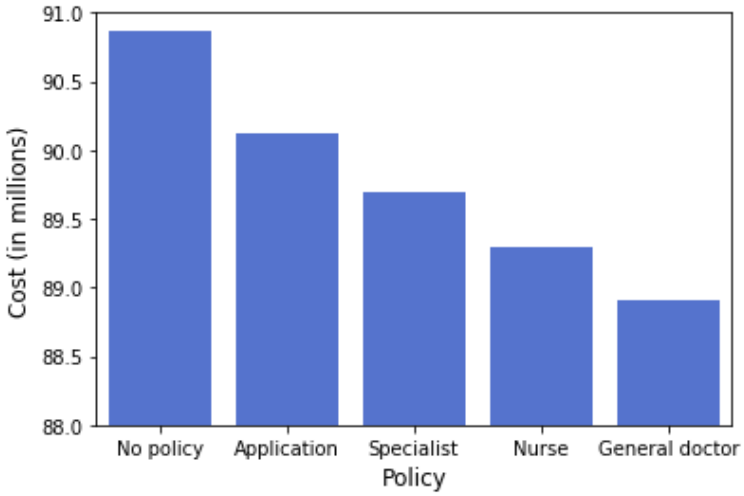


Figure IV.15 – Total cost per policy

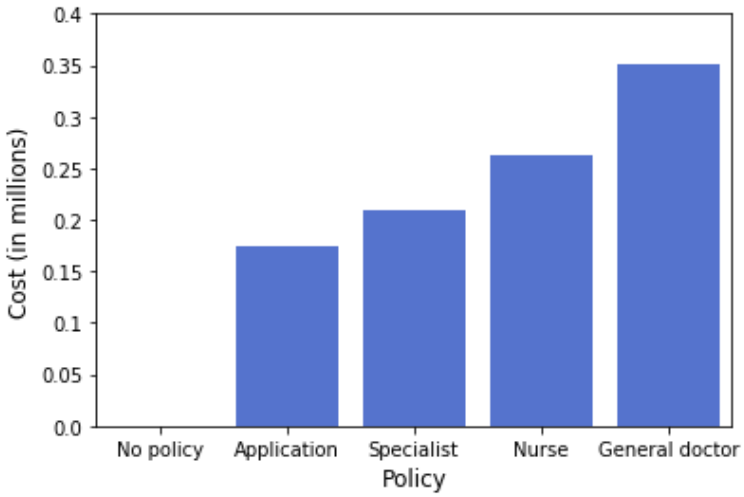


Figure IV.16 – The cost stemming from the policy

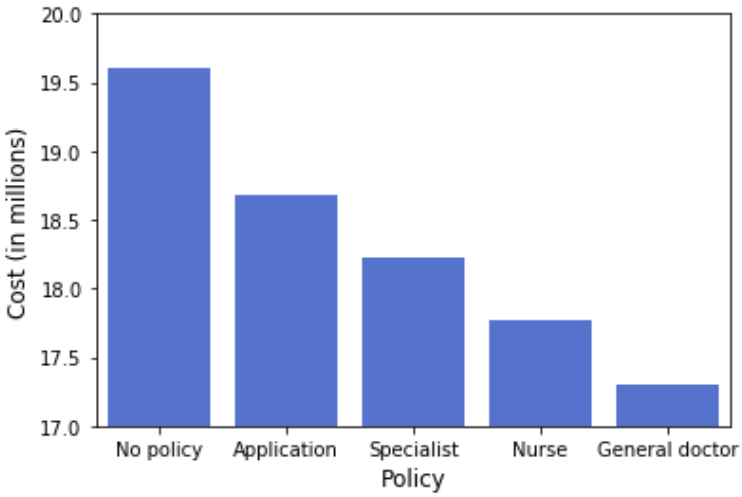


FIGURE IV.17 – The cost coming from readmissions

IV.5 Discussion

In this case study, we have looked at the influence of prevention of hospital readmission by using a prediction model. From the results we can see that if we can indeed manage to predict the patients who will return to the hospital correctly, we can manage to reduce the number of readmissions, leading to less inconveniences both for the patients as for the hospital. The extra stay of those patients does cause an increase in the total costs, even if it decreases the cost of the return of patients.

In the simulation model we have also applied several other policies. More research is necessary to know the usefulness of the different policies to prevent readmission. The costs were known in the model, but we had to make estimates on the impact of the policies. The costs are lower than an extended stay, so it can be useful to perform more research on its effectiveness.

In future work the model could be tested on a wider range of patient data, such as simulated data, or data from different hospitals or different countries. Generally, the presented model would be easily adjustable for other situations, such as different care units.

More specifications could be made in future models, such as specifying the probability of return per length of stay.

Finally, the model could be extended, to better represent reality. One possible extension might be to take into account the maximum capacity of resources of the hospital. Relevant resources may include doctors, nurses, and beds. In such an extension of the model, the flow of patients becomes of great importance. However, the flow of patients might not be directly obvious and could even depend on the number of resources available at any moment in time. Besides, the number of resources is generally not constant throughout time and depends on the policies decided by the hospital. All in all, this makes the process highly complicated.

IV.6 Note to practitioners

From the results of the simulation model we have seen that there is a clear trade-off between avoiding readmissions and the cost for the hospital. In some cases it may be useful to keep a patient in the hospital for a few extra days in order to avoid readmission. Even if the cost may increase we see that there are fewer readmissions. In the model we presented in this chapter, we can only measure quantitative performance measures. However, it should definitely be kept in mind that preventing readmissions can also avoid a lot of inconvenience for both patients and practitioners.

Besides an extended stay, it may be useful to consider other policies to avoid readmission, such as regular doctor appointments, or a mobile application. The costs of those are lower than an extra hospital stay or an extended hospital stay, however, the benefits are still unclear and should be further investigated.

IV.7 Conclusion

In this chapter we considered the problem of hospital readmissions. Such readmissions cause a big burden to both patients and the hospital. By preventing readmissions large inconveniences could be avoided, even though not all of them might be quantitatively measurable.

In Chapter II we have presented the Tabu/DAMIP framework which can be used to predict a possible readmission. In this chapter we investigated its purpose in a real-life scenario. For this goal, we created a simulation model in which we try to avoid readmissions by means of longer hospital stays for patients whom are predicted to return to the hospital. We compare this situation to the situation in which no predictions are made and thus hospital stays are not extended.

The results have shown that a trade-off exists between the number of readmissions and the total costs. However, it should be taken into account that a readmission is a highly indesirable event, of which the advantage of its avoidance cannot be measured in a clear, quantitative manner.

Besides longer hospital stays, we have also looked at the possibility of adding other care policies. Those policies are applied to patients who are predicted to return and for whom we expect that an extended hospital stay will not lead to avoidance of readmission. For these care policies we have an indication of the costs. However, the impact of the policies had to be guessed. It could be useful in future research to find out what the effect of such policies is on readmissions.

Further research may also include adjusting the model to other care units or other hospitals. In an extension of the presented model it would be interesting to include a limited amount of resources, as this represent reality. However, this is complicated as the flow of patients is unclear and may actually depend on the number of resources available and on policies decided by the hospital.

Conclusion and Future Work

In this thesis we have looked at data-driven processes and machine learning algorithms for medical decision aid. Our goal was to develop innovative machine learning techniques to support medical professionals in their decision-making process. For this goal, we have considered several case studies concerning emergency department readmission, hospital readmission, and breast cancer treatment decision aid. Multiple methods were developed and tested on several different data sets.

The data sets which were used for experiments were rather varying:

— Emergency department readmission: In this data set, we had access to very many visits to the emergency department, approximately 12 million in total. Those were so many that we could not practically use all lines. On the other hand, the number of features was limited. The data mainly contained administrative data, which was apparently not sufficient for emergency department readmission prediction as the performance on this data set was unsatisfying.

— Hospital readmission: For this data set, we had access to five years of data from one hospital, resulting in approximately 75000 hospital stays. In this data set we had a lot of information about each patient concerning diagnoses and medical acts, leading to very many features.

— Breast cancer treatment: This data set contains the least patients, approximately 2000 in total. On the other hand, this data set has a rich variety on features. It contains administrative data, as well as biological data, treatment data, and follow-up data. To arrive at such a complete data set, multiple sources had to be combined. Besides, the patients in the data set were followed for several years.

For classification our focus was on the DAMIP model. The goal of this model is to optimize the number of correctly classified entities, where an upper bound can be set on the number of misclassifications. We have seen that this classification model performs better with a smaller quantity of features. For this purpose we have experimented with both feature selection and dimensionality reduction by autoencoding. In feature selection, a subset of the known variables are chosen for the classification, whereas in dimensionality reduction the variables are used in a combination to form fewer variables while retaining the maximum amount of information possible. As an autoencoding approach results in continuous values, we presented two discretization

approaches, to make the data suitable for the DAMIP model. From the experiments we could find that the process of feature selection was found to be rather time-consuming generally, however, this is a process which only has to be done once in a real-life application. Dimensionality reduction is shown to be faster, but in this case, we have less insight in how a specific classification was reached.

Besides, we developed a simulation model to show the impact of our methods when used in a real application regarding hospital readmission. In this model, we applied the Tabu/DAMIP framework for prediction on whether a patient will return or not. If we predict that a patient is likely to be readmitted, we make a new prediction with an extended length of stay. Based on our predictions a decision on an extended stay is made. The goal of this approach is to reduce the number of readmissions. In addition, we added the use of care policies to the model. Those policies are meant for patients who are expected to return to the hospital and who we believe will not be helped by extending their hospital stay. We looked at several care policies, namely the use of a mobile application, regular visits by a nurse, visits to a general doctor and visits to a specialized doctor. In our model, we had the availability of the costs of those different policies, but the impact is highly unclear and should be further investigated. In the model we have estimated an impact factor for the different approaches. The results of all the different experiments generally show that we can indeed manage to decrease the number of readmissions, even though in this case the cost may increase. It should be taken into account that we can only quantify certain measures such as cost. However, by preventing readmissions we avoid a large inconvenience for both the hospital and for the patient, which is difficult to measure, but highly valuable.

We have shown that we can achieve good results in classification and that we can make advances in medical decision aid. Nonetheless, of course there always remains space for further research. In future work, a theoretical addition may be to look into the possibility to determine a limit on the best possible classification. In this thesis we did our performance evaluation by comparing our results to those of other methods which were applied to similar problems in literature. However, we do not know what the best possible classification score is. It would be useful to know whether a large improvement is still possible in the desired classification or whether the limits have been nearly reached.

In this thesis we have investigated multiple options to provide medical decision support and we have achieved some good results. However, of course always possibilities of improvement and further investigation exist. Some suggestions for future work are given.

In this work we have focused on binary classification, where the outcome of classification is always zero or one. It may be interesting to investigate the use of multi-class classification as decisions in healthcare are also not always one or the other. Multi-class classification is possible using the DAMIP classification model, even though it increases its complexity and computation time, it can also increase the prediction possibilities.

As has been mentioned before, interpretability is essential when applying ma-

chine learning methods in the healthcare domain. For both patients and doctors it is important for decisions to be fully transparent and explainable. By combining Tabu search for feature selection with DAMIP for classification, we achieved that the set of features used in the final classification is known. It would be interesting to see which features get chosen most often when the whole Tabu/DAMIP framework was launched 100 times. In order to do this the process should first be sped up, as now computation times are too high to get the results of 100 runs in a reasonable time frame.

In our simulation model in Chapter 4, we made some assumptions about, for example, the arrival of patients, the costs, the impact of several policies and the infinite capacity. The model could be extended, to better represent reality. Especially taking into account the maximum capacity of resources of the hospital would make the model closer to reality. Relevant resources may include doctors, nurses, and beds. In such an extension of the model, the flow of patients becomes of great importance. However, the flow of patients might not be directly obvious and could even depend on the number of resources available at any moment in time. Besides, the number of resources is generally not constant throughout time and depends on the policies decided by the hospital.

In order to evaluate the performance of our methods, we have compared our results to those in literature. One limitation we run into in the interpretation of results is that we do not know the best possible score which can be reached given our data. It could be interesting in future research, to investigate the possibility of putting an upper-bound on the best possible solution. If we would have such measure, we also get insight into whether our results can still be significantly improved or not.

Appendix A

Machine Learning Algorithms

Contents of the chapter

In this appendix an overview is given of well-known and widely used machine learning algorithms. For each method we provide a description and some insights in how it is used in machine learning applications and specifically in the field of health care.

A.1 Naive Bayesian

A Naive Bayesian classifier based on Bayes' theorem is a probabilistic statistical classifier [Yoo et al., 2012]. This classifier is based on the assumption that all features are independent of each other. This is a rather strict assumption to impose on a data set, however, even if the assumption is not exactly satisfied, the model could still give good classification performance [Bishop, 2006]. The classification is made based on Bayes' theorem, which is as follows. $P(A|B) = P(B|A) * P(A)/P(B)$. An entity is classified to the group with the highest conditional probability, that is, the group to which it most likely belongs given the values of the variables. In [Alajmani and Elazhary, 2019] a naive Bayesian classifier is used to predict the likelihood of

hospital readmission. The authors of both [Sundar et al., 2012] and [Subbalakshmi et al., 2011] use a naive Bayesian classifier to develop a tool which aids doctors in heart disease prediction.

A.2 Nearest Shrunken Centroid

In the nearest shrunken centroid algorithm for classification, shrunken centroids are used for each class and test samples are classified to the class whose shrunken centroid is nearest to it [Tibshirani et al., 2003]. This algorithm was used in [Dabney and Storey, 2007] with the purpose of clinical classification based on gene-expression microarrays. The authors of [Shen et al., 2009] apply an NSC classification algorithm to a cancer classification task. In both studies NSC is found to be successful for the considered problem.

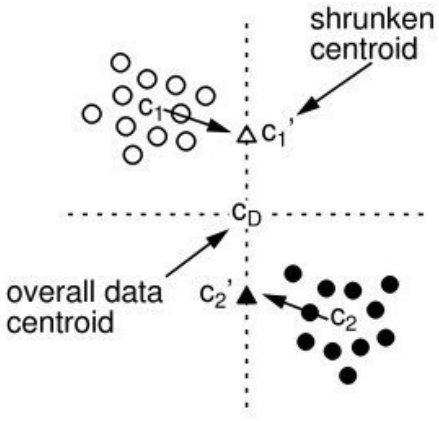


FIGURE A.1 – Nearest Shrunken Centroid.
Source: https://www.researchgate.net/publication/23459323

A.3 Linear Discriminant Analysis

Linear discriminant analysis is a commonly used technique for data classification. The method tries to maximize the ration of between-class variance to the within-class variance, guaranteeing maximal separability [Balakrishnama and Ganapathiraju, 1998]. In Figure A.2 an example of LDA is given. It can be seen that in this example, the data points are perfectly separated along the x-axis, however, they are not separated along the y-axis, making it a bad projection.

The authors of [Gao et al., 2019] show impressive results by applying linear discriminant analysis classification to EEG data in order to classify automatic epileptic seizures. In [Almeida et al., 2017] an LDA classifier is applied to georeferenced environmental factors with the purpose of forecasting hospital admissions due to asthma exacerbation. LDA was used for the classification in the diagnosis of breast cancer in [Prabhakar and Rajaguru, 2018].

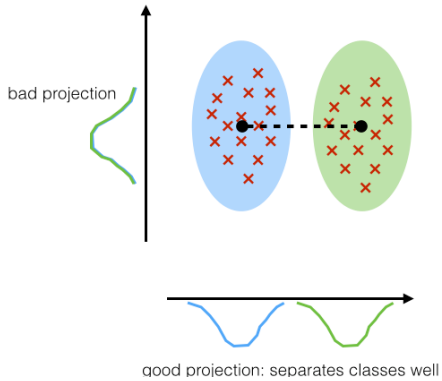


Figure A.2 – Linear Discriminant Analysis.
Source: https://sebastianraschka.com/Articles/2014_python_lda.html

A.4 Random Forest

Random forest was introduced in [Breiman, 2001]. This algorithm uses a group of classification tress, each of which is built using a bootstrap sample of the data [Diaz-Uriarte and De Andres, 2006]. A graphic overview of random forest classification is given in Figure A.3. The figure shows multiple classification trees being used to achieve a classification of an entity. By means of majority voting the final classification of random forest is determined.

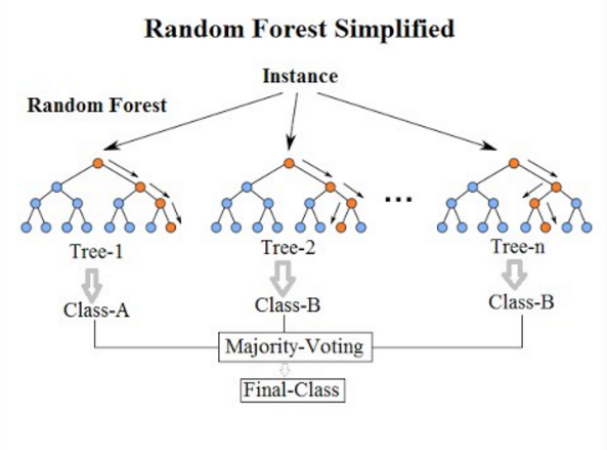


Figure A.3 – Random Forest.
Source: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

In [Khalilia et al., 2011] random forest is used to predict the risk of several diseases from the medical diagnosis history of individuals. The authors show that random forest outperforms several other classification algorithms. The authors of [Xu et al., 2017] apply random forest to determine the risk of cardiovascular problems in individual patients and in [Nguyen et al., 2013] this classification method is used for breast cancer diagnosis and prognostic.

A.5 Support Vector Machine

Fundamentally, support vector machines (SVMs) search for the optimal separating hyperplane, where the margin between two different objects is maximal. To find this maximal margin, support vectors are used [Yoo et al., 2012]. This concept can be seen in Figure A.4, where the dashed lines represent the support vectors and the solid line represents the optimal hyperplane.

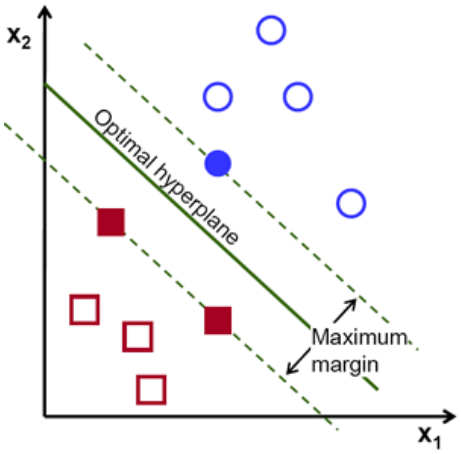


FIGURE A.4 – Support Vector Machine.
Source: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

In [Alajmani and Elazhary, 2019] an SVM classifier is used to predict the likelihood of hospital readmission. This classification technique was shown to be the best performing on this data set. Similarly, the authors of [Zheng et al., 2015] also show good results using an SVM classifier for risk prediction of hospital readmission. In [Rejani and Selvi, 2009] SVM is used for early detection of breast cancer.

A.6 Neural Network

Neural network attempts to mimic the neurological functions of the brain [Yoo et al., 2012]. The nodes are interconnected via links with adjustable weights. The weights are adjusted by learning. A visualization of a neural network is given in Figure A.5. The input layer consists of the variables of different entities. In the hidden layers the neural network constructs functions where the weights are decided in the learning process. Finally the class of each entity is given in the output layer.

The authors of [Chopra et al., 2017] apply a neural network for the prediction of hospital readmissions. In [Kuruvilla and Gunavathi, 2014] a neural network classification algorithm is used for the detection of lung cancer and in [Karabatak and Ince, 2009] this classification technique is used for breast cancer detection.

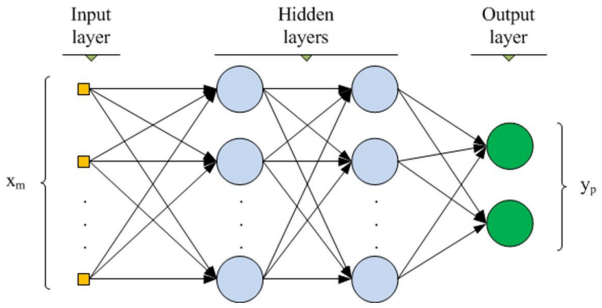


Figure A.5 – Neural Network.
Source: https://medium.com/@williamkoehrsen/deep-neural-network-classifier-32c12ff46b6c

A.7 Logistic Regression

Logistic regression is a statistical regression model, which uses a logistic function to model a binary dependent variable. This technique has as an advantage that it provides the user explicitly with probabilities and not only the class label information [Shevade and Keerthi, 2003].

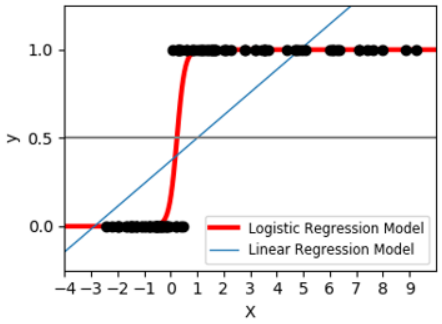


Figure A.6 – Logistic Regression.
Source: https://scikit-learn.org/stable/auto_examples/linear_model/plot_logistic.html

In [Alajmani and Elazhary, 2019] a logistic regression classifier is used to predict the likelihood of hospital readmission. The authors of [Chao et al., 2014] use logistic regression to predict the survival of breast cancer patients.

A.8 Decision Tree

Classification tree classifiers construct a tree structure, where at every step an attribute is sought whose sorting result is closest to the pure partitions by the class in terms of class values [Yoo et al., 2012]. A graphic overview of this classifier is given in Figure A.7. In each decision node a distinction of the data is made based on its variables. In each leaf node the classification is given.

In [Alajmani and Elazhary, 2019] a decision tree classifier is used to predict the likelihood of hospital readmission. Similarly, in [Sushmita et al., 2016] decision

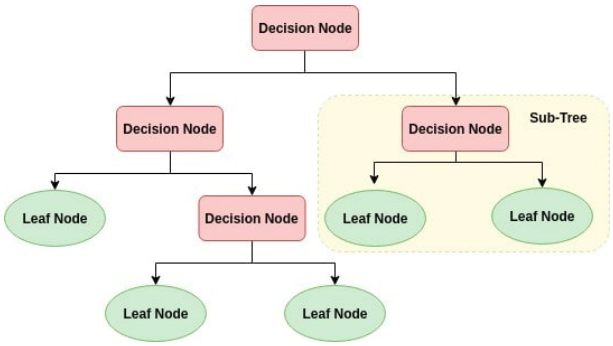


FIGURE A.7 – Decision Tree.
Source: https://www.datacamp.com/community/tutorials/decision-tree-classification-python

tree is used to predict all-cause hospital readmission. In [K. Chen et al., 2014] this classification technique is used for cancer classification using gene expression data.

Appendix B

Appendix Chapter II

Contents of the chapter

B.1 Data

Below, the two figures concerning data from the hospital readmission case study are shown. In Figure B.1 we can see the most common diagnoses by occurrence and In Figure B.2 we can see the most common medical acts in the data set, ordered by occurrence.

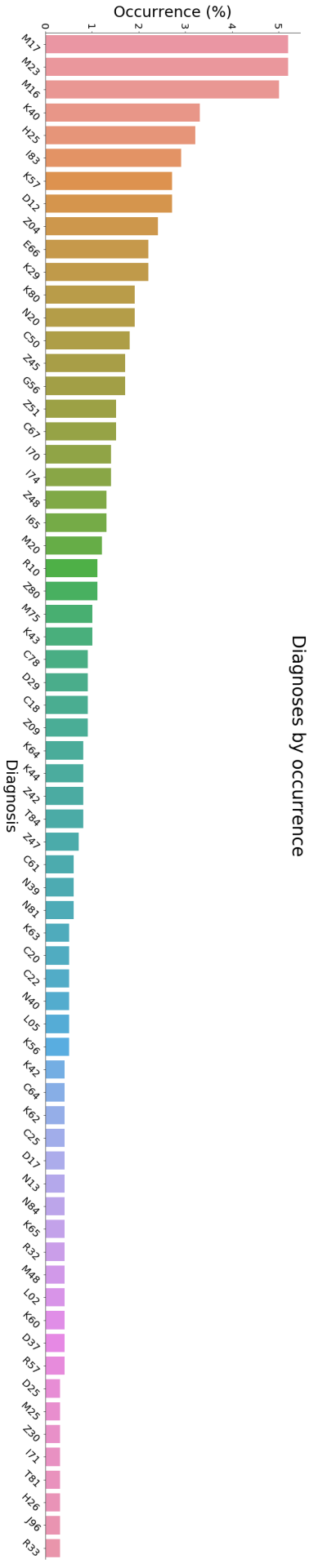


FIGURE B.1 – Occurrence (in %) of different diagnoses

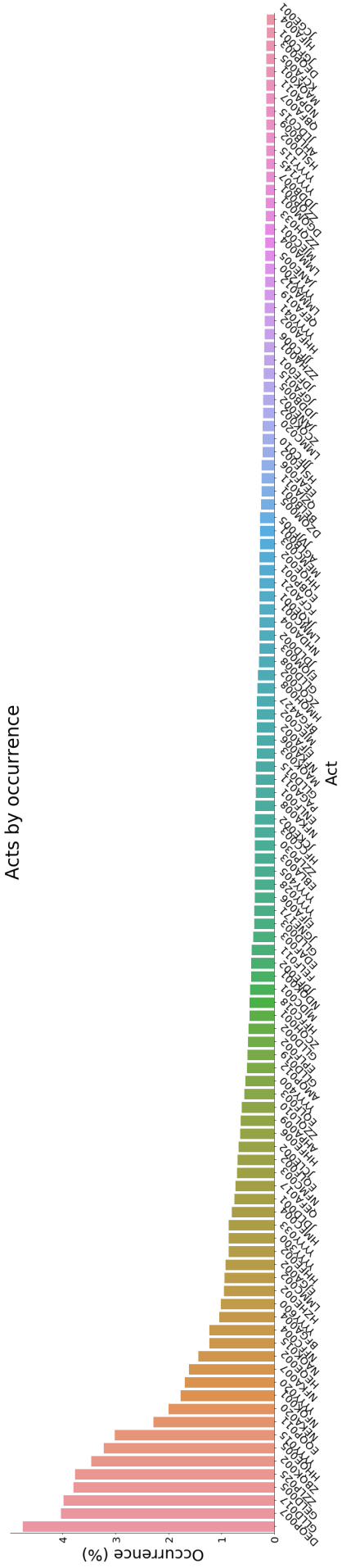


Figure B.2 – Occurrence (in %) of different medical acts

B.2 Features

Table B.1 shows the features chosen in the hospital readmission prediction case study for 180-day readmission. The features in boldface were chosen also in the 30-day and 90-day readmission predictions.

Feature name	
GHM_letter	diagactes_YYYY300
IGS2_71_80	**diagactes_JCLE002**
age_6_15	diagactes_EQLF003
age_16_25	diagactes_YYYY400
mode_sortie_last	diagactes_AMQP012
unitesoin_EFRS	diagactes_HFFC018
diagnostic_principal_firstLetter_M	diagactes_NDQK001
diagnostic_principal_firstLetter_H	diagactes_JDFE002
diagnostic_principal_firstLetter_Q	diagactes_JGNE171
diagnostic_principal_firstLetter_O	diagactes_EBLA003
diagnostic_principal_M17	diagactes_JCKE002
diagnostic_principal_M23	diagactes_PAGA011
diagnostic_principal_M16	diagactes_EJFA002
diagnostic_principal_D12	diagactes_EQBP001
diagnostic_principal_Z04	**diagactes_HSLF002**
diagnostic_principal_Z51	diagactes_JJFC010
diagnostic_principal_C78	diagactes_JDDB005
diagnostic_principal_K64	diagactes_JDFE001
diagnostic_principal_T84	diagactes_HHFA002
diagnostic_principal_Z47	diagactes_QEFA019
diagnostic_principal_N39	diagactes_YYYY200
diagnostic_principal_N40	diagactes_LMMA004
diagnostic_principal_L05	diagactes_MJEC001
diagnostic_principal_K42	diagactes_ZZQH033
diagnostic_principal_Z08	diagactes_YYYY145
diagnostic_principal_I71	diagactes_HSLD002
diagactes_HHQE005	diagactes_AHLB009
diagactes_EQQP011	diagactes_E78
diagactes_NFQK001	diagactes_B37
diagactes_NAQK015	diagactes_T81
diagactes_BFGA004	diagactes_J45
diagactes_YYYY600	diagactes_C77
diagactes_HHFE002	diagactes_L50

Table B.1 – Selected features

Appendix C

Appendix Chapter III

Contents of the chapter

C.1 Features

In Table C.1 the features chosen in the different approaches for breast cancer treatment decision support are shown.

death	**trtt/no trtt**	**chemo/no chemo**	**death after chemo**
CANALCOM	ATCDS_KC	SEIN	ATCDS_KC
MUSCLPEC	GGSENT	ATCDS_KC	CARCILOB
IMC_300_900	MUSCLPEC	GGSENT	PNONENV
TEMPS_CURAGE_4	ATCD_Diabete	CARCILOB	ENVMUSCL
TTUM_50_1000	PS_1	IMC_250_300	ATCD_insuffisance_coronarienne
RECELLUL_1	PS_2	TYPECHIR_4	ATCD_bpco
ratio_0_2	PS_3	TYPECHIR_6	PS_4
h_adj_n_4	TYPECHIR_5	CURAGE_1	CURAGE_2
anticor_2	CURAGE_3	HISTO_2	TEMPS_CURAGE_4
anticor_3	HISTO_2	TTUM_50_1000	TTUM_50_1000
lymphocytes_G_L_05_08	TTUM_0_20	SBR_3	SBR_3
lymphocytes_G_L_08_1000	SBR_3	SBR_99	SBR_99
albuminemie_30_100	N_plus_1_4	RECELLUL_4	RECELLUL_3
	N_plus_10_100	RECELLUL_5	newcopiesher2_1
	RECELLUL_1	RECELLUL_7	ratio_2_100
	RECELLUL_7	RPCELLUL_4	Score_G8_14_100
	RPCELLUL_7	her2_nb_1	Hb_g_dL_0_8
	her2_exp_2	Score_G8_0_14	Hb_g_dL_10_100
	her2_exp_4	Hb_g_dL_0_8	lymphocytes_G_L_0_02
	her2_nb_3	Hb_g_dL_8_10	albuminemie_20_30
	newcopiesher2_0	clrce_creatininemie_ml_mi_15_30	
	newcopiesher2_1	lymphocytes_G_L_02_05	
	newcopiesher2_2	lymphocytes_G_L_05_08	
	Score_G8_14_100		
	Hb_g_dL_0_8		
	Hb_g_dL_8_10		
	clrce_creatininemie_ml_mi_0_15		
	lymphocytes_G_L_0_02		
	lymphocytes_G_L_05_08		
	albuminemie_0_20		

TABLE C.1 – Features chosen for the different experiments

Appendix D

Appendix Chapter IV

Contents of the chapter

D.1 Data

Below, the two figures concerning data from the hospital readmission case study are shown. In Figure D.1 we can see the most common diagnoses by occurrence and In Figure D.2 we can see the most common medical acts in the data set, ordered by occurrence.

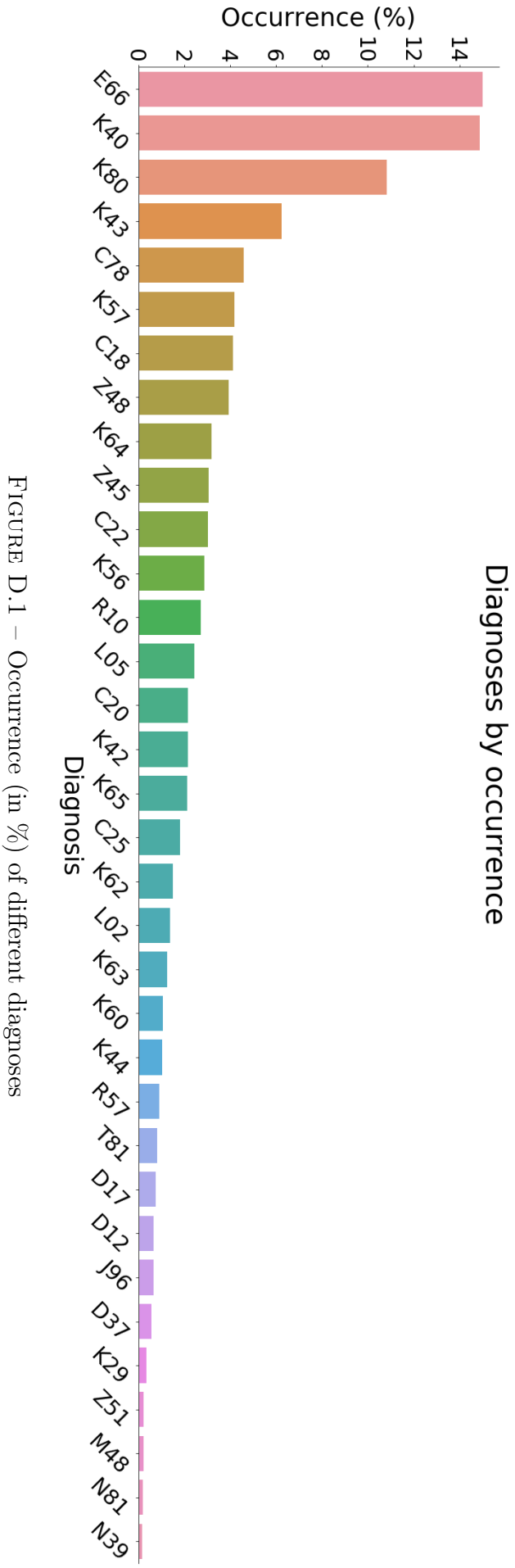


FIGURE D.1 – Occurrence (in %) of different diagnoses

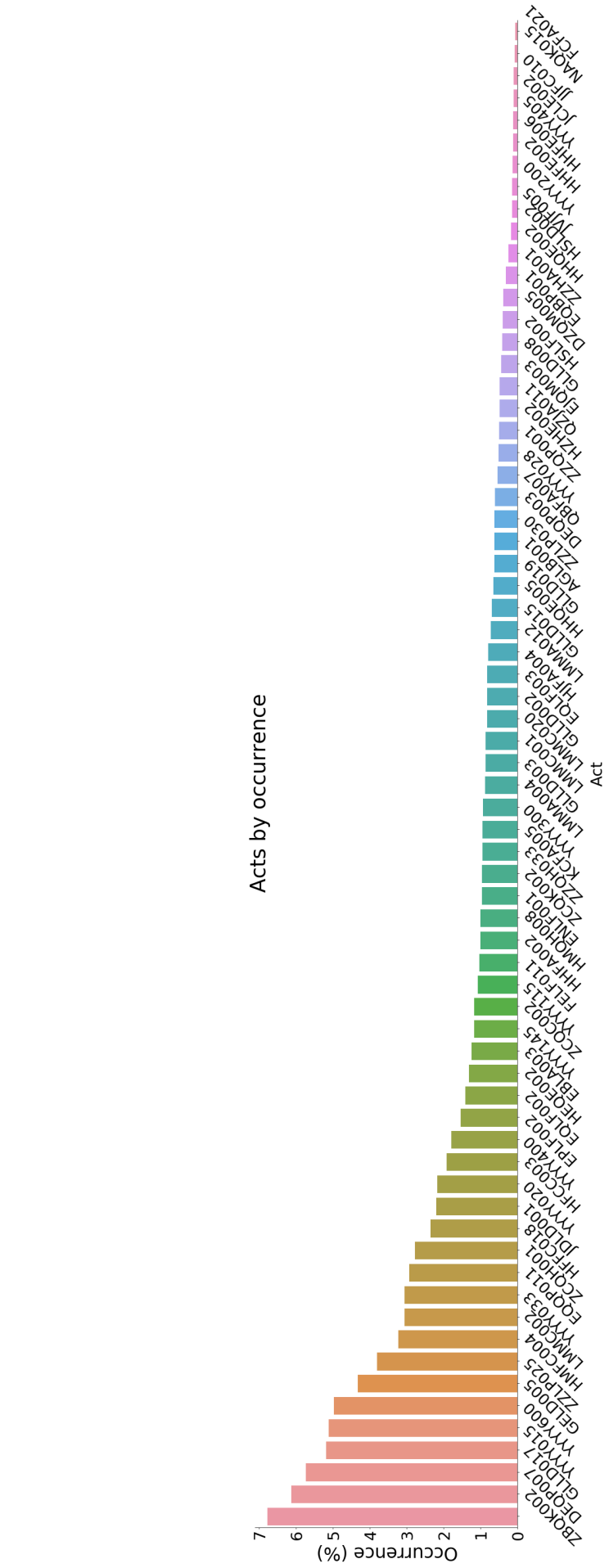


FIGURE D.2 – Occurrence (in %) of different medical acts

Bibliography

Abbas, A., Ali, M., Khan, M., & Khan, S. (2016). Personalized healthcare cloud services for disease risk assessment and wellness management using social media. *Pervasive and Mobile Computing*, *28*, 81–99.

Adem, K., Kiliçarslan, S., & Cömert, O. (2019). Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Systems with Applications*, *115*, 557–564.

Adibuzzaman, M., DeLaurentis, P., Hill, J., & Benneyworth, B. (2017). Big data in healthcare–the promises, challenges and opportunities from a research perspective: A case study with a model database, In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

Alajmani, S., & Elazhary, H. (2019). Hospital Readmission Prediction using Machine Learning Techniques. *Hospital*, *10*(4).

Ali, L., Zhu, C., Zhou, M., & Liu, Y. (2019). Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection. *Expert Systems with Applications*, *137*, 22–28.

Alkhaldi, F., & Alouani, A. (2018). Systemic design approach to a real-time healthcare monitoring system: Reducing unplanned hospital readmissions. *Sensors*, *18*(8), 2531.

Almeida, R., Teodoro, A., Gonçalves, H., Freitas, A., Sa-Sousa, A., Jácome, C., & Fonseca, J. (2017). Forecasting Asthma Hospital Admissions from Remotely Sensed Environmental Data., In *GISTAM*.

Ambale-Venkatesh, B., Yang, X., Wu, C., Liu, K., Hundley, W., McClelland, R., Gomes, A., Folsom, A., Shea, S., Guallar, E., Et al. (2017). Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circulation research*, *121*(9), 1092–1101.

Amirian, P., van Loggerenberg, F., Lang, T., Thomas, A., Peeling, R., Basiri, A., & Goodman, S. (2017). Using big data analytics to extract disease surveillance information from point of care diagnostic machines. *Pervasive and mobile computing*, *42*, 470–486.

Anderson, J. (1969). Constrained discrimination between k populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 123–139.

Artetxe, A., Beristain, A., & Grana, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer methods and programs in biomedicine*, *164*, 49–64.

Ashfaq, A., Sant'Anna, A., Lingman, M., & Nowaczyk, S. (2019). Readmission prediction using deep learning on electronic health records. *Journal of biomedical informatics*, *97*, 103256.

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064–1069.

Baig, M., Hosseini, H., & Lindén, M. (2016). Machine learning-based clinical decision support system for early diagnosis from real-time physiological data, In *2016 IEEE Region 10 Conference (TENCON)*. IEEE.

Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, *18*, 1–8.

Banu, G. (2016). Predicting thyroid disease using linear discriminant analysis (LDA) data mining technique. *Communications on Applied Electronic Journal*.

Bardhan, I., Kirksey, K., Oh, J., & Zheng, E. (2011). A predictive model for readmission of patients with congestive heart failure: a multi-hospital perspective, In *Proceedings of the thirty-second international conference on information systems*.

Benbassat, J., & Taragin, M. (2000). Hospital readmissions as a measure of quality of health care: advantages and limitations. *Archives of internal medicine*, *160*(8), 1074–1081.

Bishop, C. (2006). *Pattern recognition and machine learning*. springer.

Bjorvatn, A. (2013). Hospital readmission among elderly patients. *The European Journal of Health Economics*, *14*(5), 809–820.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brisimi, T., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I., & Shi, W. (2018). Federated learning of predictive models from federated Electronic Health Records. *International journal of medical informatics*, *112*, 59–67.

Brisimi, T., Xu, T., Wang, T., Dai, W., Adams, W., & Paschalidis, I. (2018). Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. *Proceedings of the IEEE*, *106*(4), 690–707.

Brisimi, T., Xu, T., Wang, T., Dai, W., & Paschalidis, I. (2019). Predicting diabetes-related hospitalizations based on electronic health records. *Statistical methods in medical research*, *28*(12), 3667–3682.

Buettner, F., Gulliford, S., Webb, S., & Partridge, M. (2009). Using Bayesian logistic regression with high-order interactions to model radiation-induced toxicities following radiotherapy, In *Machine Learning and Applications, 2009. ICMLA'09. International Conference on*. IEEE.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chang, V. (2018). Data analytics and visualization for inspecting cancers and genes. *Multimedia tools and applications*, *77*(14), 17693–17707.

Chao, C., Yu, Y., Cheng, B., & Kuo, Y. (2014). Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. *Journal of medical systems*, *38*(10), 106.

Chen, H., Compton, S., & Hsiao, O. (2013). DiabeticLink: a health big data system for patient empowerment and personalized healthcare, In *International Conference on Smart Health*. Springer.

Chen, K., Wang, K., Wang, K., & Angelia, M. (2014). Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. *Applied Soft Computing*, *24*, 773–780.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, *5*, 8869–8879.

Chopra, C., Sinha, S., Jaroli, S., Shukla, A., & Maheshwari, S. (2017). Recurrent neural networks with non-sequential data to predict hospital readmission of diabetic patients, In *Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics*.

Chu, A., Ahn, H., Halwan, B., Kalmin, B., Artifon, E., Barkun, A., Lagoudakis, M., & Kumar, A. (2008). A decision support system to facilitate management of patients with acute gastrointestinal bleeding. *Artificial intelligence in medicine*, *42*(3), 247–259.

Clough-Gorr, K., Thwin, S., Stuck, A., & Silliman, R. (2012). Examining five-and ten-year survival in older women with breast cancer using cancer-specific geriatric assessment. *European journal of cancer*, *48*(6), 805–812.

Coffey, A., Leahy-Warren, P., Savage, E., Hegarty, J., Cornally, N., Day, M., Sahm, L., O'Connor, K., O'Doherty, J., Liew, A., Et al. (2019). Interventions to promote early discharge and avoid inappropriate hospital (re) admission: a systematic review. *International journal of environmental research and public health*, *16*(14), 2457.

Dabney, A., & Storey, J. (2007). Optimality driven nearest centroid classification from genomic data. *PLoS One*, *2*(10).

Dai, D., & Hua, S. (2016). Random under-sampling ensemble methods for highly imbalanced rare disease classification, In *Proceedings of the International Conference on Data Mining (DMIN)*. The Steering Committee of The World Congress in Computer Science, Computer ...

Dai, W., Brisimi, T., Adams, W., Mela, T., Saligrama, V., & Paschalidis, I. (2015). Prediction of hospitalization due to heart diseases by supervised learning methods. *International journal of medical informatics*, *84*(3), 189–197.

Danaee, P., Ghaeini, R., & Hendrix, D. (2017). A deep learning approach for cancer detection and relevant gene identification, In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*. World Scientific.

Devinsky, O., Dilley, C., Ozery-Flato, M., Aharonov, R., Goldschmidt, Y., Rosen-Zvi, M., Clark, C., & Fritz, P. (2016). Changing the approach to treatment choice in epilepsy using big data. *Epilepsy & Behavior*, *56*, 32–37.

Diaz-Uriarte, R., & De Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, *7*(1), 3.

Ding, R., Jiang, F., Xie, J., & Yu, Y. (2017). Algorithmic prediction of individual diseases. *International Journal of Production Research*, *55*(3), 750–768.

Dinov, I. (2016a). Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience*, *5*(1), s13742–016.

Dinov, I. (2016b). Volume and value of big healthcare data. *Journal of medical statistics and informatics*, *4*.

Downing, N., Cloninger, A., Venkatesh, A., Hsieh, A., Drye, E., Coifman, R., & Krumholz, H. (2017). Describing the performance of US hospitals by applying big data analytics. *PloS one*, *12*(6).

El Naqa, I., Bradley, J., Blanco, A., Lindsay, P., Vicic, M., Hope, A., & Deasy, J. (2006). Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. *International Journal of Radiation Oncology* Biology* Physics*, *64*(4), 1275–1286.

Eswari, T., Sampath, P., Lavanya, S., Et al. (2015). Predictive methodology for diabetic data analysis in big data. *Procedia Computer Science*, *50*, 203–208.

Farahani, B., Firouzi, F., Chang, V., Badaroglu, M., Constant, N., & Mankodiya, K. (2018). Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare. *Future Generation Computer Systems*, *78*, 659–676.

Feng, Z., Bhat, R., Yuan, X., Freeman, D., Baslanti, T., Bihorac, A., & Li, X. (2017). Intelligent perioperative system: Towards real-time big data analytics in surgery risk assessment, In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber*

Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). IEEE.

Finkelstein, J., & Jeong, I. (2017). Machine learning approaches to personalize early prediction of asthma exacerbations. *Annals of the New York Academy of Sciences*, *1387*(1), 153.

Flaks-Manov, N., Topaz, M., Hoshen, M., Balicer, R., & Shadmi, E. (2019). Identifying patients at highest-risk: the best timing to apply a readmission predictive model. *BMC Medical Informatics and Decision Making*, *19*(1), 118.

Forkan, A., Khalil, I., & Atiquzzaman, M. (2017). ViSiBiD: A learning model for early discovery and real-time prediction of severe clinical events using vital signs as big data. *Computer Networks*, *113*, 244–257.

Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of biomedical informatics*, *56*, 229–238.

Ganggayah, M., Taib, N., Har, Y., Lio, P., & Dhillon, S. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC medical informatics and decision making*, *19*(1), 48.

Gao, Y., Zhao, Z., Chen, Y., Mahara, G., Huang, J., Lin, Z., & Zhang, J. (2019). Automatic epileptic seizure classification in multichannel EEG time series with linear discriminant analysis. *Technology and Health Care*, (Preprint), 1–11.

Garmendia, A., Graña, M., Lopez-Guede, J., & Rios, S. (2019). Neural and statistical predictors for time to readmission in emergency departments: a case study. *Neurocomputing*, *354*, 3–9.

Gay, J., Hall, M., Markham, J., Bettenhausen, J., Doupnik, S., & Berry, J. (2019). Association of extending hospital length of stay with reduced pediatric hospital readmissions. *JAMA pediatrics*, *173*(2), 186–188.

Geerts, H., Dacks, P., Devanarayan, V., Haas, M., Khachaturian, Z., Gordon, M., Maudsley, S., Romero, K., Stephenson, D., Initiative, B. H. M., Et al. (2016). Big data to smart data in Alzheimer's disease: The brain health modeling initiative to foster actionable knowledge. *Alzheimer's & Dementia*, *12*(9), 1014–1021.

Gilbank, P., Johnson-Cover, K., & Truong, T. (2019). Designing for Physician Trust: Toward a Machine Learning Decision Aid for Radiation Toxicity Risk. *Ergonomics in Design*, 1064804619896172.

Glover, F. (1989). Tabu search—part I. *ORSA Journal on computing*, *1*(3), 190–206.

Goldstein, B., Navar, A., & Carter, R. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal*, *38*(23), 1805–1814.

Gultepe, E., Green, J., Nguyen, H., Adams, J., Albertson, T., & Tagkopoulos, I. (2014). From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association*, *21*(2), 315–325.

Gupta, A., & Fonarow, G. (2018). The Hospital Readmissions Reduction Program—learning from failure of a healthcare policy. *European journal of heart failure*, *20*(8), 1169–1174.

Gusmano, M., Rodwin, V., Weisz, D., Cottenet, J., & Quantin, C. (2016). A Comparative Analysis of Hospital Readmissions in France and the US. *Journal of Comparative Policy Analysis: Research and Practice*, *18*(2), 195–209.

Hall, A., Hussain, A., & Shaikh, M. (2016). Predicting insulin resistance in children using a machine-learning-based clinical decision support system, In *International Conference on Brain Inspired Cognitive Systems*. Springer.

Handforth, C., Clegg, A., Young, C., Simpkins, S., Seymour, M., Selby, P., & Young, J. (2014). The prevalence and outcomes of frailty in older cancer patients: a systematic review. *Annals of oncology*, *26*(6), 1091–1101.

Hashi, E., Zaman, S., & Hasan, R. (2017). An expert clinical decision support system to predict disease using classification techniques, In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE.

Horng, S., Sontag, D., Halpern, Y., Jernite, Y., Shapiro, N., & Nathanson, L. (2017). Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one*, *12*(4).

Hu, C., Ju, R., Shen, Y., Zhou, P., & Li, Q. (2016). Clinical decision support for Alzheimer's disease based on deep learning and brain network, In *2016 IEEE International Conference on Communications (ICC)*. IEEE.

Hu, J., Perer, A., & Wang, F. (2016). Data driven analytics for personalized healthcare, In *Healthcare Information Management Systems*. Springer.

Hughes, K., Schnaper, L., Berry, D., Cirrincione, C., McCormick, B., Shank, B., Wheeler, J., Champion, L., Smith, T., Smith, B., Et al. (2004). Lumpectomy plus tamoxifen with or without irradiation in women 70 years of age or older with early breast cancer. *New England Journal of Medicine*, *351*(10), 971–977.

Hussain, O., & Junejo, K. (2019). Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models. *Informatics for Health and Social Care*, *44*(2), 135–151.

Jagadeeswari, V., Subramaniyaswamy, V., Logesh, R., & Vijayakumar, V. (2018). A study on medical Internet of Things and Big Data in personalized healthcare system. *Health information science and systems*, *6*(1), 14.

Jain, P., Agarwal, A., Behara, R., & Baechle, C. (2019). HPCC based framework for COPD readmission risk analysis. *Journal of Big Data*, *6*(1), 26.

Janecek, A., Gansterer, W., Demel, M., & Ecker, G. (2008). On the relationship between feature selection and classification accuracy, In *New challenges for feature selection in data mining and knowledge discovery.*

Jiang, P., Winkley, J., Zhao, C., Munnoch, R., Min, G., & Yang, L. (2014). An intelligent information forwarder for healthcare big data systems with distributed wearable sensors. *IEEE systems journal*, *10*(3), 1147–1159.

Johnson, A., Ghassemi, M., Nemati, S., Niehaus, K., Clifton, D., & Clifford, G. (2016). Machine learning and decision support in critical care. *Proceedings of the IEEE*, *104*(2), 444–466.

Jolly, T., Williams, G., Bushan, S., Pergolotti, M., Nyrop, K., Jones, E., & Muss, H. (2016). Adjuvant treatment for older women with invasive breast cancer. *Women's Health*, *12*(1), 129–146.

Kamesh, D., Neelima, V., & Priya, R. (2015). A review of data mining using big-data in health informatics. *International Journal of Scientific and Research Publications*, *5*(3), 1–7.

Karabatak, M., & Ince, M. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert systems with Applications*, *36*(2), 3465–3469.

Khalifa, A., & Meystre, S. (2015). Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of biomedical informatics*, *58*, S128–S132.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, *11*(1), 51.

Khoshhali, M., Moslemi, A., Saidijam, M., Poorolajal, J., & Mahjub, H. (2015). Predicting the categories of colon cancer using microarray data and nearest shrunken centroid. *Journal of Biostatistics and Epidemiology*, *1*(1/2), 16–21.

Koppad, S., & Kumar, A. (2016). Application of big data analytics in healthcare system to predict COPD, In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. IEEE.

Kudyba, S. (2010). *Healthcare informatics: improving efficiency and productivity.* CRC Press.

Kuruvilla, J., & Gunavathi, K. (2014). Lung cancer classification using neural networks for CT images. *Computer methods and programs in biomedicine*, *113*(1), 202–209.

Lai, H., Chan, P., Lin, H., Chen, Y., Lin, C., & Hsu, J. (2018). A Web-Based Decision Support System for Predicting Readmission of Pneumonia Patients

after Discharge, In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE.

Lambin, P., Van Stiphout, R., Starmans, M., Rios-Velazquez, E., Nalbantov, G., Aerts, H., Roelofs, E., Van Elmpt, W., Boutros, P., Granone, P., Et al. (2013). Predicting outcomes in radiation oncology-multifactorial decision support systems. *Nature reviews Clinical oncology*, *10*(1), 27.

Lee, E., & Yang, H. (2016). Predictive Analytics: Classification in Medicine and Biology. *Healthcare Analytics: From Data to Knowledge to Healthcare Improvement*, 159–187.

Lee, E., Yuan, F., Hirsh, D., Mallory, M., & Simon, H. (2012). A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department, In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

Lo'ai, A., Mehmood, R., Benkhlifa, E., & Song, H. (2016). Mobile cloud computing model and big data analysis for healthcare applications. *IEEE Access*, *4*, 6171–6180.

Lo-Fo-Wong, D., Sitnikova, K., Sprangers, M., & de Haes, H. (2015). Predictors of health care use of women with breast cancer: A systematic review. *The breast journal*, *21*(5), 508–513.

Lusa, L. Et al. (2013). Improved shrunken centroid classifiers for high-dimensional class-imbalanced data. *BMC bioinformatics*, *14*(1), 64.

Lynch, C., Abdollahi, B., Fuqua, J., Alexandra, R., Bartholomai, J., Balgemann, R., van Berkel, V., & Frieboes, H. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International journal of medical informatics*, *108*, 1–8.

Mazurowski, M., Habas, P., Zurada, J., Lo, J., Baker, J., & Tourassi, G. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, *21*(2-3), 427–436.

Munley, M., Lo, J., Sibley, G., Bentel, G., Anscher, M., & Marks, L. (1999). A neural network to predict symptomatic lung injury. *Physics in Medicine & Biology*, *44*(9), 2241.

Mylona, E., Lebreton, C., Fontaine, P., Supiot, S., Magne, N., Crehange, G., de Crevoisier, R., & Acosta, O. (2019). Comparison of machine learning algorithms and oversampling techniques for urinary toxicity prediction after prostate cancer radiotherapy, In *19th International Conference on Bioinformatics and Bioengineering, BIBE 2019*. Institute of Electrical and Electronics Engineers Inc.

Nedungadi, P., Iyer, A., Gutjahr, G., Bhaskar, J., & Pillai, A. (2018). Data-Driven Methods for Advancing Precision Oncology. *Current Pharmacology Reports*, *4*(2), 145–156.

Nguyen, C., Wang, Y., & Nguyen, H. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.

Nousi, P., & Tefas, A. (2017). Deep learning algorithms for discriminant autoencoding. *Neurocomputing*, *266*, 325–335.

Ow, G., & Kuznetsov, V. (2016). Big genomics and clinical data analytics strategies for precision cancer prognosis. *Scientific reports*, *6*, 36493.

Parikh, R., Manz, C., Chivers, C., Regli, S., Braun, J., Draugelis, M., Schuchter, L., Shulman, L., Navathe, A., Patel, M., Et al. (2019). Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. *JAMA network open*, *2*(10), e1915997–e1915997.

Peterson, U., Demerouti, E., Bergstrom, G., Samuelsson, M., Aasberg, M., & Nygren, A. (2008). Burnout and physical and mental health among Swedish healthcare workers. *Journal of advanced nursing*, *62*(1), 84–95.

Pike, M., Mustafa, N., Towey, D., & Brusic, V. (2019). Sensor Networks and Data Management in Healthcare: Emerging Technologies and New Challenges, In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*. IEEE.

Prabhakar, S., & Rajaguru, H. (2018). Performance analysis of breast cancer classification with softmax discriminant classifier and linear discriminant analysis, In *Precision Medicine Powered by pHealth and Connected Health*. Springer.

Ramirez, J., & Herrera, D. (2019). Prediction of diabetic patient readmission using machine learning, In *2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)*. IEEE.

Razzaghi, T., Safro, I., Ewing, J., Sadrfaridpour, E., & Scott, J. (2019). Predictive models for bariatric surgery risks with imbalanced medical datasets. *Annals of Operations Research*, *280*(1-2), 1–18.

Rejani, Y., & Selvi, S. (2009). Early detection of breast cancer using SVM classifier technique. *arXiv preprint arXiv:0912.2314*.

Ross, J., Mulvey, G., Stauffer, B., Patlolla, V., Bernheim, S., Keenan, P., & Krumholz, H. (2008). Statistical models and patient predictors of readmission for heart failure: a systematic review. *Archives of internal medicine*, *168*(13), 1371–1386.

Salzman, B., Knuth, R., Cunningham, A., & LaNoue, M. (2019). Identifying Older Patients at High Risk for Emergency Department Visits and Hospitalization. *Population health management*, *22*(5), 394–398.

Schwab, C., Hindlet, P., Sabatier, B., Fernandez, C., & Korb-Savoldelli, V. (2019). Risk scores identifying elderly inpatients at risk of 30-day unplanned readmission and accident and emergency department visit: a systematic review. *BMJ open*, *9*(7), e028302.

Senders, J., Staples, P., Karhade, A., Zaki, M., Gormley, W., Broekman, M., Smith, T., & Arnaout, O. (2018). Machine learning and neurosurgical outcome prediction: a systematic review. *World neurosurgery*, *109*, 476–486.

Shankar, G., & Manikandan, K. (2019). Predicting the Risk of Readmission of Diabetic Patients Using Deep Neural Networks, In *Innovations in Computer Science and Engineering*. Springer.

Sharma, A., Agrawal, P., Madaan, V., & Goyal, S. (2019). Prediction on diabetes patient's hospital readmission rates, In *Proceedings of the Third International Conference on Advanced Informatics for Computing Research*. ACM.

Shen, Q., Shi, W., & Kong, W. (2009). New gene selection method for multiclass tumor classification by class centroid. *Journal of Biomedical Informatics*, *42*(1), 59–65.

Shevade, S., & Keerthi, S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, *19*(17), 2246–2253.

Shilo, S., Rossman, H., & Segal, E. (2020). Axes of a revolution: challenges and promises of big data in healthcare. *Nature Medicine*, *26*(1), 29–38.

Shin, H., & Markey, M. (2006). A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. *Journal of Biomedical Informatics*, *39*(2), 227–248.

Stotter, A., Reed, M., Gray, L., Moore, N., & Robinson, T. (2015). Comprehensive Geriatric Assessment and predicted 3-year survival in treatment planning for frail patients with early breast cancer. *British Journal of Surgery*, *102*(5), 525–533.

Su, M., Miften, M., Whiddon, C., Sun, X., Light, K., & Marks, L. (2005). An artificial neural network for predicting the incidence of radiation pneumonitis. *Medical physics*, *32*(2), 318–325.

Su, M., Wang, Y., Chen, T., Chiu, S., Chang, H., Huang, M., Hu, L., Li, C., Yang, S., Wu, J., Et al. (2020). Assess the Performance and Cost-Effectiveness of LACE and HOSPITAL Re-Admission Prediction Models as a Risk Management Tool for Home Care Patients: An Evaluation Study of a Medical Center Affiliated Home Care Unit in Taiwan. *International Journal of Environmental Research and Public Health*, *17*(3), 927.

Subbalakshmi, G., Ramesh, K., & Rao, M. (2011). Decision support in heart disease prediction system using naive bayes. *Indian Journal of Computer Science and Engineering (IJCSE)*, *2*(2), 170–176.

Sun, X., Li, M., Meng, C., Kong, N., Meng, H., & Hyer, K. (2017). Data-driven simulation for healthcare facility utilization modeling and evaluation, In *2017 Winter Simulation Conference (WSC)*. IEEE.

Sundar, N., Latha, P., & Chandra, M. (2012). Performance analysis of classification data mining techniques over heart disease database. *International journal of engineering science & advanced technology*, *2*(3), 470–478.

Sushmita, S., Khulbe, G., Hasan, A., Newman, S., Ravindra, P., Roy, S., De Cock, M., & Teredesai, A. (2016). Predicting 30-day risk and cost of" All-Cause" hospital readmissions, In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Swain, A. (2016). Mining big data to support decision making in healthcare. *Journal of Information Technology Case and Application Research*, *18*(3), 141–154.

Szlosek, D., & Ferrett, J. (2016). Using machine learning and natural language processing algorithms to automate the evaluation of clinical decision support in electronic medical record systems. *eGEMs*, *4*(3).

Thesmar, D., Sraer, D., Pinheiro, L., Dadson, N., Veliche, R., & Greenberg, P. (2019). Combining the power of artificial intelligence with the richness of healthcare claims data: Opportunities and challenges. *PharmacoEconomics*, *37*(6), 745–752.

Thottakkara, P., Ozrazgat-Baslanti, T., Hupf, B., Rashidi, P., Pardalos, P., Momcilovic, P., & Bihorac, A. (2016). Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PloS one*, *11*(5), e0155705.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 104–117.

Toğaçar, M., Ergen, B., & Cömert, Z. (2020). Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders. *Medical hypotheses*, *135*, 109503.

Valdes, G., Simone II, C., Chen, J., Lin, A., Yom, S., Pattison, A., Carpenter, C., & Solberg, T. (2017). Clinical decision support of radiotherapy treatment planning: a data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiotherapy and Oncology*, *125*(3), 392–397.

Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 1–15.

Vergara, J., & Estévez, P. (2014). A review of feature selection methods based on mutual information. *Neural computing and applications*, *24*(1), 175–186.

Waljee, A., Lipson, R., Wiitala, W., Zhang, Y., Liu, B., Zhu, J., Wallace, B., Govani, S., Stidham, R., Hayward, R., Et al. (2018). Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. *Inflammatory bowel diseases*, *24*(1), 45–53.

Wang, F., Casalino, L., & Khullar, D. (2019). Deep learning in medicine—promise, progress, and challenges. *JAMA internal medicine*, *179*(3), 293–294.

Wang, L., Chen, D., Ranjan, R., Khan, S., KolOdziej, J., & Wang, J. (2012). Parallel processing of massive EEG data with MapReduce, In *2012 IEEE 18th International Conference on Parallel and Distributed Systems*. Ieee.

Wang, L., Tong, L., Davis, D., Arnold, T., & Esposito, T. (2020). The application of unsupervised deep learning in predictive models using electronic health records. *BMC medical research methodology*, *20*(1), 1–9.

Wang, L., Wang, Y., & Chang, Q. (2016). Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods*, *111*, 21–31.

Wang, T., & Paschalidis, I. (2019). Prescriptive Cluster-Dependent Support Vector Machines with an Application to Reducing Hospital Readmissions. *arXiv preprint arXiv:1903.09056*.

Wang, Y., Kung, L., & Byrd, T. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3–13.

Wang, Y., Li, L., Ni, J., & Huang, S. (2009). Feature selection using tabu search with long-term memories and probabilistic neural networks. *Pattern Recognition Letters*, *30*(7), 661–670.

Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing*, *184*, 232–242.

Weng, S., Reps, J., Kai, J., Garibaldi, J., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, *12*(4).

Wiens, J., & Shenoy, E. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, *66*(1), 149–153.

Xu, S., Zhang, Z., Wang, D., Hu, J., Duan, X., & Zhu, T. (2017). Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework, In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*. IEEE.

Yang, Y., Fasching, P., & Tresp, V. (2017). Predictive modeling of therapy decisions in metastatic breast cancer with recurrent neural network encoder and multinomial hierarchical regression decoder, In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, *36*(4), 2431–2448.

Yuan, F. (2015). *Modeling and computational strategies for medical decision making* (Doctoral dissertation). Georgia Institute of Technology.

Zhang, H., & Sun, G. (2002). Feature selection using tabu search method. *Pattern recognition*, *35*(3), 701–711.

Zhao, Y., Wong, Z., & Tsui, K. (2018). A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *Journal of healthcare engineering*, *2018*.

Zheng, B., Zhang, J., Yoon, S., Lam, S., Khasawneh, M., & Poranki, S. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, *42*(20), 7110–7120.

Zong, B., Song, Q., Min, M., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT: Communiqué le jour de la soutenance

Author: Daniëlle HOOIJENGA

Title: Medical Decision Support Using Machine Learning

Speciality: Industrial Engineering

Keywords: Machine Learning, Decision Support, Healthcare, Hospital Readmission, Classification, Feature Selection, Autoencoder, Performance Evaluation

Abstract

The main objective of this thesis is to develop innovative machine learning techniques to aid in medical decision making using available health databases such as PMSI, and local databases of the CLB (Center Léon Bérard, Lyon, France). In the thesis, several case studies were examined, concerning readmission to the emergency room, readmission to the hospital and decision support for the treatment of breast cancer.

Firstly, we consider a case study concerning emergency department readmission prediction. Readmission to the emergency department may be a sign of insufficient treatment at the first visit of the patient. For this goal, we combine a classification model (DAMIP) with tabu search for feature selection. Moreover, we combine this with a sampling method for speeding up the computations. This method has been shown to give slightly better results than conventional machine learning methods found in literature when applied to the emergency department readmission data set. However, the results for all the tested methods were not very satisfying, which was probably due to a lack of information in the data. The data consisted mainly of administrative data about the patient. Because of this, we considered another case study concerning hospital readmission. This data set contains a large number of features, including diagnoses and medical acts. On this data set the results were significantly better and we again showed that our Tabu/DAMIP framework outperforms other methods.

After, we developed a method that combines an autoencoder for dimensionality reduction with DAMIP for classification. In this method, we make use of a discretization method in order to be able to combine the two parts. This method was tested on CLB breast cancer treatment data. The goal of the case study is to be able to aid the clinician in making a decision on the treatment of an elderly breast cancer patient. We aim to do this by several approaches. Either we try to predict 5-year survival given the treatment the patient receives, or we try to predict whether a patient needs a treatment or not, given that the patient survives at least five years. The results are similar to those of the Tabu/DAMIP framework, but they are obtained much faster by using an autoencoder. Besides, we also combined the autoencoder with other classification algorithms, where the best result was obtained by the autoencoder with linear discriminant analysis.

Finally, we developed a simulation model to show the impact of our methods when used in a real application regarding hospital readmission. In this model, we apply the Tabu/DAMIP framework for prediction on whether a patient will return or not. If we predict that a patient is likely to be

readmitted, we make a new prediction with an extended length of stay. Based on our predictions a decision on an extended stay is made. The goal of this approach is to reduce the number of readmissions. The results show that we can indeed manage to decrease the number of readmissions, even though in this case the cost may increase.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT: Communiqué le jour de la soutenance

Auteur: Daniëlle Hooijenga

Titre: Aide à la décision médicale grâce à la classification automatique

Spécialité: Génie Industriel

Mots-Clefs: Machine Learning, L'aide à la décision, Santé, Réadmission à l'Hôpital, Classification, Sélection des Variables, Auto-encodeur, Évaluation des Performances

Résumé

L'objectif principal de cette thèse consiste à développer des techniques innovantes d'apprentissage automatique pour aider à la décision médicale à l'aide des bases de données de santé disponibles telles que PMSI, et des bases de données locales du CLB (Centre Léon Bérard, Lyon, France). Dans la thèse, plusieurs études de cas ont été examinées, concernant la réadmission aux urgences, la réadmission à l'hôpital et l'aide à la décision pour le traitement du cancer du sein.

Tout d'abord, une combinaison de DAMIP avec la recherche tabou a été développée, avec DAMIP pour la classification et la recherche tabou pour la sélection des variables. Il a été démontré que cette méthode donne de meilleurs résultats que les méthodes classiques d'apprentissage automatique. Ces résultats ont été obtenus sur les données de réadmission à l'hôpital. Ensuite, nous avons développé une méthode qui combine un autoencodeur pour la réduction de dimensionnalité avec DAMIP pour la classification. Cette méthode a été testée sur les données de traitement du cancer du sein au CLB. Les résultats sont similaires à la méthode précédente, mais sont obtenus beaucoup plus rapidement. Nous avons aussi combiné l'autoencodeur avec d'autres algorithmes de classification : le meilleur résultat a été obtenu par l'autoencodeur avec une analyse discriminante linéaire.

Enfin, nous avons développé un modèle de simulation afin d'évaluer l'impact de nos méthodes lorsqu'elles sont utilisées dans une application réelle de réadmission à l'hôpital. Dans ce modèle, nous appliquons la méthode Tabu / DAMIP pour prédire si un patient reviendra ou non. Si nous prédisons qu'un patient est susceptible d'être réadmis, nous faisons une nouvelle prédiction avec une durée de séjour prolongée. Sur la base de nos prévisions, une décision sur un séjour prolongé est prise. Le but de cette approche est de réduire le nombre de réadmissions. Les résultats montrent que l'on parvient effectivement à diminuer le nombre de réadmissions, même si dans ce cas le coût peut augmenter.

ÉCOLE DES MINES DE SAINT-ÉTIENNE
158, cours Fauriel – CS 62362
42023 Saint-Étienne Cedex 2 – France
Tél. +33(0)4 77 42 01 23
www.mines-stetienne.fr