Institute of Formal Methods in Computer Science

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Hospital Emergency Room Workload Prediction using Artifical Neural Networks

Sebastian Patrick Paule

Course of Study:	Softwaretechnik
Examiner:	Prof. Dr. Stefan Funke
Supervisor:	Anna Rosa, M.Sc. (iteratec GmbH) Dr. Michael Gebhart (iteratec GmbH)
Commenced:	15.12.2021
Completed:	15.6.2022

Abstract

In hospital emergency rooms, where workloads are inherently inhomogeneous, predicting those workloads as accurately as possible plays an essential role in employee shift planning. These workloads are not random but the result of the complicated interaction of environmental influences on human health. Artificial neural networks can use data to identify correlations between those environmental influences and hospital workloads, predicting future workloads based on new data.

While the total workload of a hospital emergency room is interesting and already researched, no work tries to predict multiple different categories of diagnosis to identify the staff that has to be ready for any given time. This thesis aims to show that artificial neural networks can predict the workloads of different diagnosis categories. We use historical data on multiple environmental influences and the emergency room of the Universitätsklinikum in Freiburg, Germany, to train four different artificial neural networks.

While the metrics show a promising result, the networks have problems accurately predicting outliers, like extremely high and low workloads. This makes the networks a reasonable basis for further research but, in their current state, irrelevant for a real-life application.

Kurzfassung

In Notaufnahmen, in denen die Arbeitsbelastung von Natur aus inhomogen ist, spielt eine möglichst genaue Vorhersage dieser Belastung eine wesentliche Rolle bei der Schichtplanung. Diese Arbeitsbelastung ist nicht zufällig, sondern das Ergebnis einer komplizierten Interaktion von Umwelteinflüssen auf die menschliche Gesundheit. Künstliche Neuronale Netze können Daten nutzen, um Korrelationen zwischen diesen Umwelteinflüssen und der Arbeitsbelastung in Krankenhäusern zu ermitteln und die künftige Arbeitsbelastung auf der Grundlage neuer Daten vorherzusagen.

Während die Gesamtauslastung der Notaufnahme eines Krankenhauses interessant und bereits teileise erforscht ist, gibt es keine Arbeiten, die versuchen, mehrere verschiedene Diagnosekategorien vorherzusagen, um das Personal zu ermitteln, das zu einem bestimmten Zeitpunkt bereitstehen muss.

In dieser Arbeit wird gezeigt, dass künstliche neuronale Netze die Arbeitsbelastung für verschiedene Diagnosekategorien vorhersagen können. Wir verwenden historische Daten über verschiedene Umwelteinflüsse und die Notaufnahme des Universitätsklinikums in Freiburg, Deutschland, um vier verschiedene künstliche neuronale Netze zu trainieren. Während die Metriken ein vielversprechendes Ergebnis zeigen, haben die Netze Probleme, Ausreißer, wie extrem hohe und niedrige Arbeitsbelastungen, genau vorherzusagen. Dies macht die Netze zu einer vernünftigen Grundlage für weitere Forschung, aber in ihrem derzeitigen Zustand wenig hilfreich in einer Feldanwendung.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

FF The artifical neural network that predicts all diagnosis categories using a feed forward architecture. See Section 4.5.1. ix

FOCUS The artificial neural network that predicts the number of patients with a single diagnosis category. See Section 4.5.3. ix

ICD-10 an international coding for medical diagnosises from the world health organisation. See ICD-10 Codes. ix

LSTM The artifical neural network, that predicts all categories using an LSTM cell. See Section 4.5.4. ix

MAE Mean average error. One metric used to measure the quality of artifical neural networks. See Metric. ix

$\mathbf{MAE_{avg}}$ The MAE of the baseline reference. 44

MAPE Mean average percentage error. One metric used to measure the quality of artifical neural networks. See Metric. xv

$\mathbf{MAPE_{avg}}$ The MAPE of the baseline reference. 44

MLP A multi-layer perceptron is an artifical neural network with an feed forward design and at least three layers. See Section 2.4.1. 8

$\mathbf{PM_{10}}$ Very small particles with a maximum aerodynamic diameter of 10 microns. 27

$\mathbf{PM_{2.5}}$ Very small particles with a maximum aerodynamic diameter of 25 microns. 27

TNOP The artifical neural network that predicts the total number of patients. See Section 4.5.2. ix

Terms

Enviromental influences factors from the enviroment that have an influence on a system. For example temperature, time or air pollution. 29

Search Space The range in which a value can be.. 43

timelag If one of the models is trained with a timelag of x timeframes, the data used for training is x timeframes older than the timeframe the model predicts for. Sometimes the data used for the prediction spans over all x timeframes. 5

1 Introduction

One of the most known constants in the health care industry is the chronically understaffed personnel. Compared to similar nations, German hospitals are critically understaffed. In 2019 Germany had 4.53 full-time nurses per 1000 citizens. Significantly less than similar countries like the USA (6.46), Switzerland (6.67), or Norway (8.14) [Sim19]. This problem, which often increases as wages decreases, can have brutal consequences. The health care industry exists to take care of patients, who rarely have an actual choice if they want to use health care. Often, it is not a luxury but a necessity to go to an emergency room, as it can significantly impact one's quality of life, not to receive help from a medical professional. Additionally, the need for emergency response, in its nature, arrives unexpectedly, making the workload of emergency rooms inherently dynamic.

In summary, emergency rooms have an unknown, uncontrollable demand while unable to secure the fulfillment of this demand by overstaffing the emergency room. The diversity of injuries and illnesses that a hospital emergency room encounters, as it can expect the full range of medical emergencies, creates a demand for specially trained personnel.

To increase the ability of a hospital to prepare for the incoming workloads, the amount and problems of future patients could be estimated. This estimation creates information that can improve shift planning, which is a predictive task in itself. The provided resources must be planned before the demand is known, leaving a margin of error. If the demand is known during planning, this margin of error can be reduced significantly.

This thesis evaluates if this demand is predictable using artificial neural networks, a variant of machine learning.

Machine learning is a collection of approaches that use large quantities of data to create a model that can solve tasks using similar information. Those tasks can range from classification, like object recognition of pictures and videos, to regression, identifying patterns in data. Patterns that can, given new information, estimate future developments.

We developed four proof of concept artificial neural networks to evaluate different aspects of the prediction. Three were designed using a feed-forward architecture to predict the total amount of patients that will be treated stationary in the emergency room (TNOP), the number of patients that predicts the number of patients with *Injuries, poisonings, and certain other consequences of external causes* (FOCUS)and the number of patients for multiple different kinds of injuries and illnesses, categorized after the ICD-10 system (FF). The last artificial neural network implements a recurrent network architecture using long-short term memory (LSTM). Similar to the FF network, it predicts the number of patients for multiple kinds of injuries and illnesses.

All four networks are trained to predict the emergency room workload daily. The prediction could then be used to improve staff planning. So medical emergencies that have a high probability of happening that day can be attended to by a qualified person immediately. In contrast, those with a low probability of happening have to wait for the correct personnel for a bit. Alternatively, they can be attended by staff who are not as well-trained for it as possible. Effectively reducing the number of times, the second case should happen, as it is not entirely preventable.

The prediction should not be viewed as a magic gaze into the future but as a rough estimation based on historical data. While a lot can happen without anybody expecting or predicting it, the world is filled with factors influencing each other. Some of them seem obvious to us: For example, it seems reasonable that heatstrokes are more common in summer than winter.

However, the system can never predict all possible injuries or illnesses entering the emergency room at any given timeframe. It is limited by the data the system is trained on for once. Therefore, the system will not recognize a wholly unknown or a sufficiently rare problem. For example, when an exceptional situation occurs, like a bridge collapse or pandemic, the system cannot predict the resulting emergency room visits.

The artificial neural networks will be trained on multiple environmental influences, like weather, air pollution, or vacation days. All data is collected in the greater area of Freiburg, Germany, to match the hospital data from the Universitätsklinikum Freiburg, which provides the historic emergency room workloads.

This thesis will not try to identify the reason for illness and injury to happen. Many papers from medical professionals have more insight into identifying those patterns [LPS+19] [NLBW71] [CSSA95] [HLY+08]. It will only try to create a system that can find relations between environmental changes and the workload in the emergency room.

Also, the prototype will, at most, be able to identify patterns in the data, not the reasons behind them or even if those patterns are of any significance. Nevertheless, the results could be used for staff planning or as the basis for further research.

Thesis Structure

The following chapter provides a quick overview of the structure of this thesis.

Chapter 2 – Foundations: At first, this thesis will look at research that has already taken place in this area. It will briefly overview machine learning, neural networks, and environmental influences on human health. Especially influences on hospital or emergency workload or connections between arbitrary influences to injuries and illnesses are the focus of this chapter.

Chapter 3 – Related Work: This chapter defines the target of this thesis using requirements engineering. It will use those requirements to look at other publications that tried to achieve similar targets. It will also highlight similarities and differences between the aim of this thesis and the related works.

Chapter 4 – Planning: Next is the preperation of the system. In this chapter, the thesis evaluates possible environmental influences, like weather or air pollution. It will discuss different parts of the CRISP-DM process, especially the data understanding and preparation steps. This includes data acquisition and data cleaning.

Additionally, it will talk about the architecture of the neural networks that this thesis uses and the output format for the prediction.

Chapter 5 – Implementation This chapter reviews the implementation of the proof of work prototype itself. It will explain the prototype's technology and highlight the different problems that appear during the development and how they are approached.

Chapter 6 – Evaluation This chapter presents the prototype results and sets them in context with the test data. We are evaluating the quality of the prototype and trying to identify if it could be helpful in a real-world application. We will analyze how accurate the predictions are when predicting based on unknown data and compare them to a baseline estimation.

Chapter 7 – Conclusion The conclusion of the thesis revisits the most important results and provides an outlook on possible further works.

2 Foundations

2.1 Enviromental Influences

The world is an interactive system in which nearly every part influences each other. The human body is one of many parts of this system. We interact with our surroundings in many different ways through our senses and actions. There are more, less noticeable interactions than those alone: Our body can react significantly when it comes in contact with an alien object like a virus or pollen. The size of the environmental influence does not always correlate to the size of impact it has on our body. For example, a virus can damage a body significantly more than a feather.

In this thesis, we will look at several environmental influences and try to use them to predict the workload of hospital emergency rooms.

The weather is among the most extensively studied environmental influences on our health. It has been shown that the temperature influences the risk of children's hospital admission [LPS+19], the risk of pneumonia [SCK+19], and other diseases [RBH02]. The weather can also influence humans' psychological well-being, which was shown in a study by Brandl E. et al. [BLB+18].

Closely connected to the temperature, the time of year can have an influence too. The amount and kind of complaints that appear in the ER changes with the season and day of the week [NLBW71].

Another area that has been heavily researched in the last years has been air pollution. It can be measured through the concentration of harmful gas like NO_2 (nitrogen dioxide) or O_3 (ozone). Another pollutant is microscopic particles, often collectively referred to as $\mathrm{PM_X}$, which defines particles with a maximum aerodynamic diameter of X microns.

Air pollution can be connected to respiratory diseases in Barcelona [CSSA95], Boston [Zan06], Santa Clara [LHO97], and Seattle [SSL+93]. Additionally, Halonen J. et al. [HLY+08] showed that air pollution affects all age groups with a different timelag. While elderly people had immediate effects, asthma-related ER visits from children occurred five days later.

It is a widespread belief that the moon, more specifically the lunar cycle, significantly influences human psychology. Several studies have looked at this with different results. Nevertheless, the consensus tends to see no connection between human behavior and the lunar cycle [KASJ11][OK83][MDH06].

2.2 ICD-10 Codes

Making medical data internationally comparable and exchangeable requires a standard structure to collect it. The world health organization (WHO) provides such a structure in the form of the ICD system. ”Clinical terms coded with ICD are the main basis for health recording and statistics on disease in primary, secondary, and tertiary care, as well as on cause of death certificates”[Org].

The newest iteration of the ICD system is version 11, released in January 2022. The version used during data collection was the previous one, ICD-10.

The ICD-10 system provides hierarchically sorted diagnoses. Multiple similar diagnoses are grouped in 22 top-level categories called chapters. Each top-level category is a semantic collection of diagnoses. For example, chapter XI is made of the ICD-10 codes K00 to K93 and represents diseases of the digestive system [Kro].

The categories are all listed in the Appendix A. Further information can be retrieved from the website [Kro]

2.3 Machine Learning

Machine learning provides a broad range of tools to work with vast data. As the name suggests, those tools can learn from this data. The aim is to generalize the problems so that the solution can be successfully applied to examples that were not part of the training data. Such a self-learned generalization is more straightforward to create than hard-coding every possible problem variation and how to solve it. This idea ”is as old as computers themselves, perhaps older still” [Kub17]. Only 1983, the first book collecting research papers in this field was published: ”Machine Learning: The AI Approach”, which makes machine learning a relatively young field [CMM83]. Since then, the field has developed dozens of machine learning algorithms that use different ideas to extract concepts from data. Generally, there are three different categories of machine learning that all have advantages.

Unsupervised learning does not require example solutions or extensive control of the output. It ”is interested in discovering useful properties of available data” [Kub17]. A typical use would be the Cluster analysis, in which ”the input is a set of examples, each described by a vector of attribute values—but no class labels. The output is a set of two or more clusters of examples” [Kub17]. These clusters will not be based on any human-given label but on the data alone, which has a higher probability of containing unexpected results[Nas17]. ”Thus, cluster analysis is a very promising tool for the exploration of relationships between many papers” [Nas17].

Another category is supervised learning, in which ”the various algorithms generate a function that maps inputs to desired outputs. One standard formulation of the supervised learning task is the classification problem: the learner is required to learn (to approximate the behavior of) a function that maps a vector into one of several classes by looking at several input-output examples of the function” [Nas17]. This provides a system that is more predefined by human design but is more likely to miss unexpected correlations.[Nas17]

Semi-supervised learning combines both those ideas by providing labeled and unlabeled examples together. It provides the learning process with hints on which labels should map the data, which is advantageous if the generation of labeled example data is especially expensive. [ZG09]

Alternatively, there is reinforcement learning, which uses environmental feedback to identify the system's quality. This feedback is often called the fitness function. ”One of the challenges that arise in reinforcement learning, and not in other kinds of learning, is the trade-off between exploration and exploitation” [SB98].

As this thesis classifies input data into predefined classes, supervised learning is the most relevant category of algorithms. This category is separated into classification, which sorts the input into predefined classes, and regression map the input space in a real value domain.

”Along with regression and probability estimation, classification is one of the most studied models, possibly one with the greatest practical relevance. The potential benefits of progress in classification are immense since the technique has a great impact on other areas, both within Data Mining and in its applications” [Nas17].

While machine learning uses collected data and examples to learn the relevant patterns instead of expert heuristics, it is important not to believe that this removes bias from the system. Data and examples can be as biased as an expert and must be treated that way.

As early as 1994, machine learning was proposed as an alternative to statistical analysis in the medical field. In their paper, Harry B. Burke et al. propose to replace the traditional pTNM staging system for cancer prognosis with neural networks, as they ”are able to significantly improve breast cancer outcome prediction accuracy when compared to the TNM stage system” [BRG94]. The accuracy of 10 different systems was compared with data from 5169 training cases and 3102 test cases. The most significant improvement they mentioned was the ability of machine learning to include more factors in the prediction to increase the accuracy[BRG94].

2.4 Artifical Neural Networks

Artificial neural networks are a common implementation that can be used in every kind of machine learning problem listed above. It is inspired by the biological brain, which uses many connected neurons and is represented by a directed graph. The nodes are the artificial neurons, and the edges are the connections of the neurons.

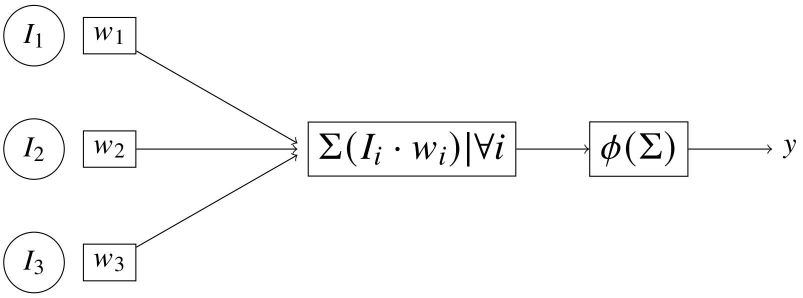
Each neuron weights each edge pointing into it, and an activation function ϕ decides when the neuron fires. A firing neuron provides a signal on each outgoing edge for the subsequent nodes to work with.

Overall a neuron with n inputs is represented by the function

$$y = \phi(\Sigma(I_i \cdot w_i))|i \in [0, n] \tag{2.1}$$

In this equation, I_i represents the input provided by the node i from the last layer. w_i is the corresponding weight, and y is the output of the current node.

Figure 2.1: The abstract structure of an artifical neuron (picture based on [Joa20])



Additionally, a learning function exists that can change each neuron's bias to fit the intended solution. Depending on the design and problem, the learning and activation functions can vary from neural network to neural network.

How drastically the values change is defined by the learning rate. If the rate is high, the model may jump over the optimal configuration and only stays in its proximity without ever reaching it. If it is too low, it can take significantly longer to train. The most common solution is to start with a high learning rate and reduce it over time to allow more precise corrections.

The other main design point of an artificial neural network is the architecture of the neuron graph itself. There are many ideas on how to structure the network. Each approach has advantages and disadvantages, like the different machine learning approaches.

2.4.1 Feed Forward networks

Feed Forward is a simple design principle wherein each layer feeds its outputs only to the next layer. The resulting graph structure is directed and acyclic. Each piece of information only passes once through the network.

One implementation of the feed-forward design is the so-called multi-layer perceptron (MLP) It is the most common design pattern for artificial neural networks and can be found in classification, regression, and forecasting systems[Fin99]. A multi-layer perceptron contains at least three layers:

1. **Input Layer:** This layer is the interaction point to insert data into the artificial neural network.
2. **Output Layer:** This layer contains all the solutions the system can have. For example, the classes it tries to differentiate between.
3. **Hidden Layer(s):** A MLP can contain one or more hidden layers, which are all fully connected and do the main part of the calculation.

Layers in such networks are often named fully connected, a simple term to describe two layers in which every note of the first layer has an edge to every node of the second layer. It was already used in the multi-layer perceptron in Figure 2.2. A connected input can still be trained to be ignored, while a missing connection provides an irremovable bias.

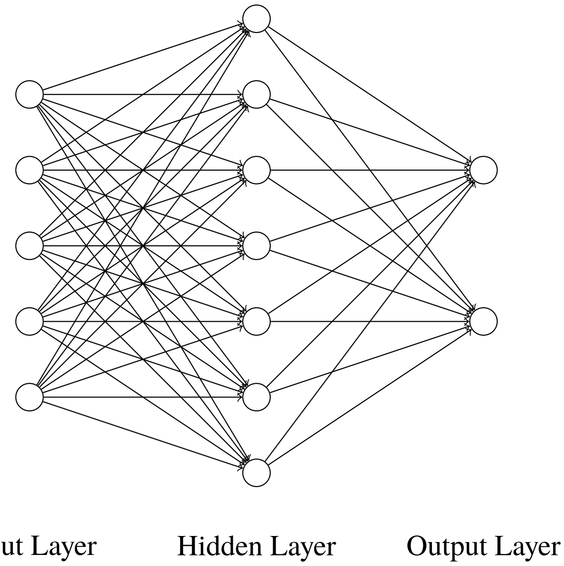


Figure 2.2: An example for a feed forwards artifical neural network: Multi Layer Perceptron (designed after [KBK11])

2.4.2 Recurrent networks

While a directed graph still represents recurrent networks, they contain loops designated to keep states over multiple executions of the artificial neural network. Those short-term memory structures are called context layers [ST05] and provide a signal back to a previous layer.

An example of recurrent artificial network architecture is the long-short term memory (LSTM) architecture. They introduce a so-called LSTM cell, which provides the artificial neural network with the ability to keep an internal state over multiple inputs, making LSTM networks especially good when the input is a sequence, like text, audio, and video [YSHZ19].

3 Related Work

3.1 Requirements

This chapter explains in detail what the proof of concept prototype is intended to be. Requirements are a core concept in software development, as they allow a precise definition of the intended system. This happens in a process called requirements-engineering. It intends to identify the problem the system should solve and the so-called stakeholders interested in a solution [Par10].

Stakeholders are persons that directly or indirectly influence the requirements for the system [Kla21]. Usually, they would be interviewed to ensure the correct requirements depending on the importance of their influence. However, because this thesis only tries to create a proof of concept prototype, this is not needed as extensively as for an actual product.

Two doctors of the Universitätsklinikum Freiburg are interviewed as part of the requirements engineering process. They represent the stakeholders in the context of this thesis.

3.1.1 Stakeholders

Patients

Patients are interested in the quality of the received care, especially as this care is vital for their well-being and even their survival. When patients arrive at the emergency room, they want to be taken care of as soon as possible by staff that is as qualified as possible for their situation.

A systematic literature review about patient experiences in the emergency department analyzed 107 publications. It identified ”staff-patient communication” as the most frequent theme, with ”wait times” as the second most frequent one [SAL+17].

Medical Staff

The medical staff is an essential part of every emergency room. Without doctors, nurses, and other employees, a hospital would be unable to operate. Nevertheless, this essential resource is unable to work indefinitely. People cannot work arbitrarily long without problems; in most countries, they are protected by labor laws. This gives every shift an upper limit for length. Additionally, people do not want to idle without work, making overstaffing a problem.

Hospital Management

The hospital management is the operational head of a hospital. It decides how the hospital acts each day. For the hospital to operate as well as possible, the management wants to increase effectiveness, maximize predictability and reduce the idle time of essential resources like equipment and staff.

Independent Ethics Committee

The independent ethics committee is a part of universities and similar institutions that reviews and monitors research efforts to ensure ethical standards. It can be crucial if the research works with people or personal information, like medical patient data.

They intend to protect the rights and welfare of people who are part of the research or any study.

The independent ethics committee assesses research and procedures with relevant recognized scientific procedures and criteria and in accordance with authoritative international ethical norms and standards. If needed, additional professional associations, like the professional, ethical guidelines of the German Psychologists' Association and the German Psychological Society are consulted [Kri21].

Government

As a representative of the people, the government imposes restrictions on the actions of all other stakeholders, especially the hospital management, independent ethics committee, and medical staff. The intention here is to protect the patients and employees further. Patients are protected in multiple ways.

For example, the ”Patientendaten-Schutz-Gesetz” (Patient Data Protection Act) protects patients' personal data and ensures the right to a digital medical record. For that, it regulates how the medical sector has to work with said data [PDSG].

Another group of laws that is interesting for this thesis is labor laws. For example, the ”Arbeitszeitgesetz” limits the time an employee can work per day or how long breaks have to be [ArbZG].

3.1.2 Requirements

Workload Prediction

functional requirement

One way to improve the provided service in a hospital emergency room is to optimize the available staff to the patient demand. Optimization means, in this context, that the staff should neither be idle nor should it exceed legal or sensible shift lengths, while the workforce is always able to take care of the patients with appropriate wait times. The hospital needs to know what kind of personnel will be needed in advance to achieve this.

Therefore, the core feature of this thesis is the prediction of how many patients will arrive at the hospital and what kind of treatment they will need.

Prediction of Multiple Kinds of Diagnoses

functional requirement

When a patient arrives at the emergency room, his problem can be one of many, and it would be impossible for a single medical professional to be able to treat all possible problems. Overall, the ICD-10 system lists 22 categories, each containing dozens of individual groups, which can also be separated into different diagnoses [Org][Kro]. For example, a neurologist is not always qualified to take care of patients that have an essential (primary) hypertension (I10), and a Hematologist is not always able to take care of the consequences of inflammatory diseases of the central nervous system (G09). It makes the availability of specialized staff important.

Similarly, the prediction itself should not focus on one kind of illness or one diagnosis alone. Instead, it should give a prediction for multiple ICD-10 categories, which should be helpful in staff planning.

Sensible Time Frame

non-functional requirement

The workload prediction must predict the number of patients for a specific time frame to support staff planning. This time frame must be matched to work times in an emergency room and the environment data on which the prediction will be based. Several options could be chosen: hourly, daily, weekly, and monthly. While a more granular timeframe like hourly would be the most useful, it is also more probable that the data is unavailable in this granularity or that the information is too sparse for the machine learning algorithm to use.

We will use daily and weekly as timeframes. That should provide a reasonable tradeoff.

Sensible Granularity

non-functional requirement

The prediction categories in which we separate the diagnoses must be chosen well. They should not be too broad, as that would prevent the advantages for staff planning that this system should provide. On the other hand, neither should they be too granular. Increasing the number of categories has a diminishing return, as the same staff will be responsible for multiple categories. Meanwhile, the quality of the machine learning model deteriorates with the increase of output categories, and an unnecessary increase should be prevented.

This thesis will use less than 20 categories. Those reflect the medical categorization of the ICD-10 system, with some less frequent cases packed together to improve the machine learning output.

Quality of Analysis

non-functional requirement

As reality is naturally chaotic, the predictions can not be perfectly accurate. A machine learning system can only find patterns in the provided training data and apply those to new data. Especially extreme, random events like a bridge collapse or a major crash create high numbers of patients that the system cannot predict. A perfect prediction is also unnecessary, as a minor variance in the number of patients a single nurse or doctor takes care of does not drastically impact their ability to care for those patients.

The expected quality of analysis is measured by taking the deviation of the predicted value from the expected value. This deviation has to be consistently better than the deviation from the average expected value for that timeframe.

Secret Data

specification

The machine learning system has to work with medical data of real patients. This data has to be especially protected under the german law [1] [2] [3]. The system is developed without data access to protect the patient's personal data. Instead, the system has to be developed as a Docker container that will be manually executed on the hospital server. A doctor from the Universitätsklinikum Freiburg takes care of the deployment and execution of the container. Additionally, he checks each output for protected data before sending it back to the developer.

[1] DSGVO (https://eur-lex.europa.eu/eli/reg/2016/679/oj?locale=de)[DSGVO]

[2] PDSG (https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/P/PDSG_bgbl.pdf)[PDSG]

[3] Verschwiegenheitspflicht (https://www.gesetze-im-internet.de/stgb/__203.html)[MBO-Ä]

In the City of Freiburg

specification

The system should use correlations between environmental data and emergency room workload to predict future workload, which is impossible if those datasets are from entirely different locations. The emergency room that provides the medical data is located in Freiburg, Germany. As a result, the prediction is also limited to Freiburg (Baden-Württemberg), Germany.

3.2 Evaluation of Existing Work

There are already papers that try to identify the impact of the environment on the medical industry or use machine learning to predict workloads and performance in other areas. In this chapter, the thesis will give a brief overview and evaluate them against the requirements described above.

3.2.1 Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases [LBX+21]

Lu et al. analyzes different artificial neural networks to predict hospital emergency room visits of patients with respiratory diseases. The intention is to solve the problem that traditional statistical approaches have ”difficulties in dealing with situations with multi-factor effects. When multiple factors fluctuate, the prediction accuracy reduces significantly”.

They collected data from 10 hospitals in the greater area of Bejing from January 1 to December 31, 2013. The data was used together with ”daily mean concentration of PM2.5 (μ g/m3), daily mean temperature (°C), daily mean relative humidity (%), and daily mean wind speed (MSPD) (m/s)”[LBX+21].

The paper trained three different neural networks to evaluate them against each other.

- **ARIMA** or Autoregressive Integrated Moving Average is ”one of the most classic models for the prediction of time series, which has been widely applied in many areas, such as electricity price prediction, energy consumption forecast, and so on. The basic idea of the ARIMA model is to use a certain mathematical model to describe the random time series of the data, then predict the future values based on the past, and present values, so-called autoregression.” [LBX+21]

- **MLP** or Multi-Layer Perceptron is one of the standard architectures for artificial neural networks. It uses at least three fully-connected neuron layers. The input layer, the output layer, and at least one hidden (calculation layer). ”MLP model is one of the most effective artificial neural network technologies for modeling and forecasting, so it has been used as a benchmark model by many studies”[source (use the same source as Lu2021)].

- **LSTM** is a model is proposed by Hochreiter and Jürgen Schmidhuber in 1997 to handle long time task lag. ”In an LTSM cell, the forget gate decides what and how information to be discarded from the calculation, using nonlinear functions and weight matrixes. The input gate determines what information to be added to the calculation (sigmoid layer) and gets the new candidate information (tanh layer). The cell updates the information. Finally, the output gate determines what and how information to be output, also including a sigmoid layer and a tanh layer” [LBX+21].

The paper identifies the LSTM model as the best and the ARIMA model as the worst at predicting patient models. Each model’s best performance happened at another timelag. For example, the ARIMA model decreased in quality with each day of timelag, making the first day the best performing, while the LSTM model’s maximum prediction quality was with a timelag of three days.

This work differs from the requirements as it only looks at respiratory diseases and air-related input factors to identify patterns. It fulfills the requirement *Workload Prediction* but not the requirement *Prediction of Multiple Kinds of Diagnoses*. Also similarly to 3.2.5, Lu et al. meet the non-functional requirements *Sensible Time Frame*, *Sensible Granularity* and *Quality of Analysis*, but not the specification *Secret Data* and *In the City of Freiburg*.

3.2.2 Association between weather conditions and the number of patients at the emergency room in an Argentine hospital [RBH02]

Rusticucci et al. found correlations between emergency room workload and the weather. They looked at the summer of 1996-1997 and the winter of 1996 in Argentina and evaluated the appearance of patients in the emergency room of a hospital in Buenos Aires city. The intention was to find relations between hospital emergencies and the weather using linear correlation analysis.

The cases were seperated into 7 categories:

(1) respiratory, cardiovascular and chest-pain complaints

(2) digestive, genitourinary and abdominal complaints

(3) neurological and psychopathological disorders

(4) infections

(5) contusion and crushing, bone and muscle complaints

(6) skin and allergies

(7) miscellaneous complaints

They compared those with temperature, dew-point temperature, dew-point depression, sea-level pressure, visibility, wind speed, daily calm frequencies, and wind-direction frequencies. Significant correlations were found between groups 2,3 and 6: Skin and allergies (group 6) are positively correlated with temperature and dew-point temperature while negatively correlated with sea-level pressure. Group 2 and the frequency of westerlies showed a negative correlation. Another negative correlation was found for group 3 compared to windspeed and pressure.

The paper does look at the correlation between environmental influences and the emergency room workload. However, it does this without using machine learning, which restricts the number of influences evaluated and provides reduced prognostic ability (*Workload Prediction* not met). On the other hand, those correlations are analyzed with multiple kinds of illnesses in mind (*Prediction of Multiple Kinds of Diagnoses*).

Again, this paper meets the non-functional requirements *Sensible Time Frame*, *Sensible Granularity* and *Quality of Analysis*. The specification *Secret Data* and *In the City of Freiburg* are not fulfilled.

3.2.3 Unorganized Machines to Estimate the Number of HospitalAdmissions Due to Respiratory Diseases Caused by PM10 Concentration [TBC+21]

In this paper, Tadano et al. explore deep neural networks to estimate hospital admissions' complex, nonlinear behavior based on PM_{10} concentrations. PM_{10} is the group of particles with 10 μm diameter or less. The authors propose that deep neural networks improve estimations compared to Generalized Linear Models (GLM) and Generalized Additive Models (GAM). The critical downside is that regression models need more data than GLM and GAM.

For their proposal, Tadano et al. use specialized versions of artificial neural networks: so-called unorganized machines, which are themselves separated into echo state networks (ESNs) and extreme learning machines (ELMs).

- **ELM** Extreme learning machines are feed-forward networks with a single hidden layer that uses a Linear Combiner as the output layer. This structure is quite similar to a multi-layer perceptron. In this paper, the linear combiner uses a Moore-Penrose generalized inverse operator.
- **ESN** Echo state networks are a variant of recurrent neural networks without iterative adjustment. Nodes of the hidden layer can influence each other in each iteration. Similar to the ELM, the ESN also uses a linear combiner.

To evaluate the networks, the authors used three cities with different characteristics:

- **São Paulo**

  " [...] São Paulo City, the largest city in Brazil, has almost 12 million people (data of 2010) in 1500 km2, 7398.26 inhabitants per km2. The average climate is tropical, about 28∘C in summer and 12∘C in winter. "[TBC+21] The data is from January 2014 to December 2016 and contains 159,683 cases.

- **Campinas**

  " Campinas City is the third most populous city in São Paulo State, with approximately 1,1 million people (data of 2010) spread over 795.7 km2, a demographic density of 1359.6 inh/km2[52]. The climate is tropical with dry winter and rainy summer with an average of 37∘C during summertime. "[TBC+21] The data is from January 2017 to December 2019 and contains 15,46 cases.

- **Cubatão**

  " Cubatão has an estimated 118,720 inhabitants with 142.8 km2and 831 inh/km2. "[TBC+21] The data is from January 2017 to December 2019 and contains 802 cases. The case number is so tiny because the city is comparably small. It is still fascinating, as the city is "one of the most global polluted cities "[TBC+21]

Apart from the PM_{10} concentration, the authors also included the maximum temperature, day of the week, relative humidity, and holiday information in their training dataset. Additionally, the data had a 7-day timelag to simulate the time the pollution needs to affect the human body.

The paper finds that ELM performs the best out of all tested models, having a mean average error in all cities. While the ESN model provided worse results in most cases, statistical analyses showed that, in Campinas, the prediction quality of the ESN model was similar to the prediction quality of the ELM model, especially regarding the root mean square error metric metric (RMSE).

This paper matches the requirements of this thesis the most. The intention of Tadano et al. is a *Workload Prediction* with a *Prediction of Multiple Kinds of Diagnoses*. Besides the functional requirements, the paper also meets all non-functional requirements *Sensible Time Frame*, *Sensible Granularity* and *Quality of Analysis*. Only the specifications *Secret Data* and *In the City of Freiburg* remain unfulfilled.

3.2.4 Machine Learning Based Workload Prediction in Cloud Computing [GWS20]

”It is important for cloud service providers (CSPs) to provide cloud service resources with high elasticity and cost-effectiveness and then achieve a good quality of service (QoS) for their clients. However, meeting QoS with a cost-effective resource is a challenging problem for CSPs because the workloads of Virtual Machines (VMs) experience variation over time” [GWS20]. This problem is the main factor that decides the earnings of a cloud service provider. If the quality of service agreement is not met, the cloud provider often has to pay a fine. Furthermore, over-providing resources is a waste of money on its own.

Gao et al. show that it is possible to predict the performance needs of a virtual machine using a combined approach of different machine learning algorithms. They used Support Vector Machines (SVM), Bayesian Ridge Regression (BRR), ARIMA, and an LSTM to predict the future workload of virtual machines and, therefore, support cloud providers’ resource allocation. To train the system, they used the Google cluster trace. It is a data collection of 12.5 thousand machines in the Google cluster, spanning over 29 days and containing resource utilization of CPU and memory usage of each task on those machines.

The machine learning algorithms were used to predict the workload in the next timestep (called 0-gap prediction) and the workload m steps in the future (called m-gap prediction). The results show that the Bayesian Ridge Regression provides the best results for both predictions, closely followed by the LSTM and ARIMA approaches. Both perform equally well. The worst performing approach in the 0-gap prediction was the Support Vector Machine. In the m-gap prediction, the SVM was not evaluated at all.

Similar to the aim of this thesis, the paper tries to identify the future workload of a system. The system in this paper is a cloud computing system, not a medical environment. Nonetheless, parallels still exist. For example, both areas have resources that need to be allocated to different system parts, and this demand reacts to external influences.

Also similar is using an LSTM artificial neural network as a prediction tool. For Gao et al., the LSTM was only one of the evaluated approaches. The paper has similar intentions in predicting the demand for resources in a system.

However, it has also differences from the aim of this thesis. The most apparent one is that the tool predicts cloud systems’ workload and not a hospital emergency room *Workload Prediction*. Also, Gao et al. only try to predict the demand for a single resource, rather than multiple resources as specified in the Requirement *Prediction of Multiple Kinds of Diagnoses*. Less important, but still relevant, the paper neither works with *Secret Data* nor does the data originates *In the City of Freiburg*

3.2.5 Peak Outpatient and Emergency Department Visit Forecastingfor Patients With Chronic Respiratory Diseases Using MachineLearning Methods:Retrospective Cohort Study [PCZ+20]

In this paper, Peng et al. experimented with different machine learning algorithms to predict the peak arrival of chronic respiratory disease patients in the emergency room. They identified the sudden inflow of outpatient and emergency patients as one of the fundamental issues in hospital management. They intended to reduce the adverse effects of such crowds by predicting them. This would allow health staff to prepare for the increased demand.

Examples of such a workload peak in an emergency room would be an influenza season.

Namely, they used bagging, adaptive boosting, and random forests for their experiments and compared those results to a general linear model (GLM) and polynomial nuclear Support Vector Machine (SVM).

- **GLM** General Linear Models are a generalized version of nonlinear models introduced by John Nelder and Robert Wedderburn in 1972. It was intended to unify different statistical regressions, like linear, Poisson, and logistic regression. [NW72]
- **SVM** Support Vector Machines are a supervised learning model developed by Vladimir Vapnik with colleagues. It is intended to classify data into two categories based on a dataset with labeled data. [Vap98]
- **Bagging** or Bootstrap aggregating is a process that uses multiple regression or classification models and uses the average of the results to reduce the variance of the total output. In this paper, the aggregated machine learning approaches are N different runs of the same tree algorithms [Bre96].
- **Adaptive Boosting** This method assigns a weight for each entry in the training dataset. These weights will then be changed for each iteration of a classifier applied to the data. If the classifier pays more attention to a data point, its weight will increase, otherwise decrease. Yoav Freund and Robert Schapire initially formulated this method to boost the performance of other classifiers. [FS96]
- **Random Forests** The algorithm creates *N* random decision trees in this machine learning approach. It combines the result with a special ënsemble”method, creating a highly efficient algorithm, as the different trees can be executed in parallel [Ho95].

To train those models, they collected data based on related work. ”Namely, wind speed, atmospheric pressure, outdoor temperature, relative humidity, carbon monoxide, ozone, sulphur dioxide, nitrogen dioxide, and PM25” [PCZ+20]. They removed entries with less than ten people and introduced a 3-day lag to clean the data, resulting in a dataset with 559 entries.

This paper is fulfilling a lot of the requirements of this thesis: The machine learning approach does a *Workload Prediction* for medical workloads. This prediction has a *Sensible Time Frame*, *Sensible Granularity* and a sensible *Quality of Analysis*. However, it only looks at a single kind of illness instead of a range of multiple different ones. Therefore it does not fulfill the requirement *Prediction of Multiple Kinds of Diagnoses*. Also, Peng et al. neither work with *Secret Data* or have data origin *In the City of Freiburg*.

3.3 Overview of Related Work

Lu et al. evaluate MLPs and LSTMs as possible improvements for workload predictions in emergency rooms. Using the collected data from ten hospitals, they predicted the daily number of patients. Peng et al. test different machine learning algorithms to predict larger crowds arriving in emergency rooms. Tadano et al. use deep learning algorithms to predict the number of respiratory diseases in a hospital based on air pollution. Rusticucci et al. identify correlations between different environmental influences and seven types of hospital emergencies using linear correlation analysis. Gao et al. use machine learning to predict the workload of cloud computing systems to improve scaling.

Multiple papers show that it is possible to use machine learning and especially artificial neural networks to predict the workload of a system that is heavily influenced by the environment. However, only one of the papers tries to predict multiple kinds of diagnoses at once, and it is not using a machine learning approach, which is the use case this thesis intends to cover. Additionally, none of the papers use data from Freiburg, Germany, or have to work with confidential data.

An overview of the papers and requirements they fulfill can be seen in Table 3.1. The checks indicate fulfilled requirements. They are in brackets if the requirement is only partly fulfilled.

Table 3.1: Overview of related work

	Workload Prediction	**Prediction of multiple kinds of diagnoses**	**Sensible Time Frame**	**Sensible Granularity**	**Quality of Analysis**	**Secret Data**	**In the City of Freiburg**
Machine Learning Based Workload Prediction in Cloud Computing [GWS20]	(✓)		✓	✓	✓		
Peak Outpatient and Emergency Department Visit Forecastingfor Patients With Chronic Respiratory Diseases Using MachineLearning Methods:Retrospective Cohort Study [PCZ+20]	✓		✓	✓	✓		
Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases [LBX+21]	✓		✓	✓	✓		
Association between weather conditions and the number of patients at the emergency room in an Argentine hospital [RBH02]		✓	✓	✓	✓		
Unorganized Machines to Estimate the Number of HospitalAdmissions Due to Respiratory Diseases Caused by PM10 Concentration [TBC+21]	✓		✓	✓	✓		

4 Planning

In order to create any meaningful machine learning model, the underlying data must be adequately prepared, understood, and used. The whole process called Data Mining, ”is a creative process which requires a number of different skills and knowledge” [WH00]. The quality of the resulting dataset can depend heavily on the person or team preparing it.

A standardized process can be followed to increase the quality of the resulting model. Such a model helps in the data mining work and prevents important parts from being overlooked or lost.

This thesis uses CRISP-DM described by Wirth et al. [WH00].

4.1 CRISP-DM

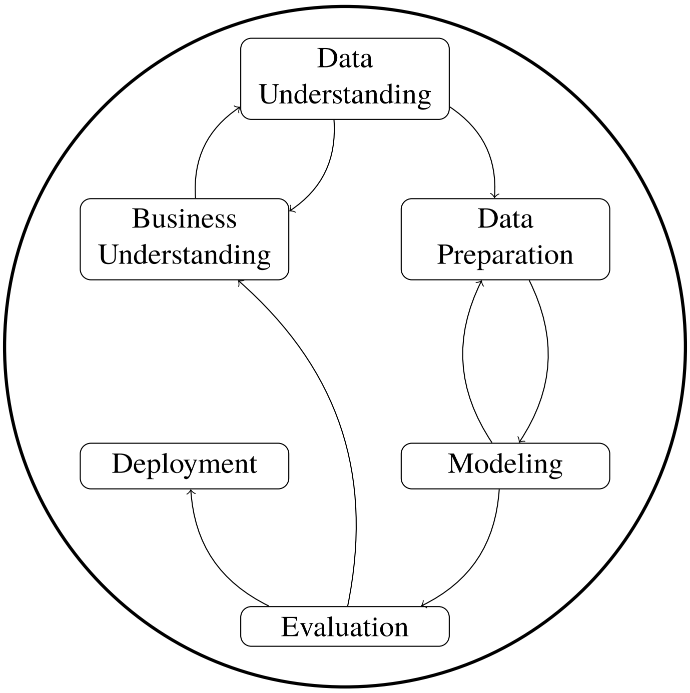


Figure 4.1: The phases of CRISP-DM [WH00]

CRISP-DM (Cross Industry Standard Process for Data Mining) is an industry-standard for data mining. Rüdiger Wirth and Jochen Hipp originally proposed it in 2000 to provide a structured approach to data mining. They intended to reduce variance in the data mining process. The CRISP-DM process includes ways to decide which data is relevant and how to retrieve and prepare data before using it in a data model.

The methodology of CRISP-DM separates the data mining process into several steps. Those are ordered in four levels of abstraction: phases, generic tasks, specialized tasks, and process instances.

Here is a short overview of the highest level of abstraction, the phases. An important aspect is that the analyst is encouraged to jump back to an earlier phase if needed.

- **Business Understanding**

  The first step is understanding the project objectives and requirements from the business perspective. In the case of this paper, this is done in Chapter 3 through requirements engineering.

- **Data Understanding**

  Data understanding means data collection and initial work with the data. It means understanding the semantics behind the data. This step is closely connected to the business understanding step, as the data must be set in context with the use case. Chapter Section 4.2 will cover this step.

- **Data Preparation**

  In this phase, the collected data is converted into the dataset that the model can use. This contains "table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools." [WH00]

  This step is described in the Planning and the Implementation chapter.

- **Modeling**

  Next, the dataset is used in the intended algorithm. This thesis uses artificial neural networks as a machine learning algorithm, so the dataset is used to train the artificial neural networks. It is partly reflected in the Implementation chapter.

- **Evaluation**

  Before deployment, the quality of the model needs to be evaluated. For this thesis, the networks are compared against each other in the Evaluation chapter.

- **Deployment**

  As the last step, the resulting model needs to be deployed to be used. This does not necessarily mean that it will be embedded in an application or accessible to users. It can also mean writing a report, paper, or, like in this case, a thesis about it.

4.2 Data

In order to create a machine learning model, the first planning step is to understand and prepare the data the system is trained on. An important rule here is bad in equals bad out, meaning that the quality of the training data is strongly influenced by the quality of the resulting machine learning model: "Machine Learning is, after all, Data-Driven AI, and your model will be only as good or as bad as the data you have"[Cha17]. Therefore, it should be analyzed which data fits the intended use of the system the best, how the data is pre-processed, and where it could be compromised.

Weather

Weather is one of the most researched environmental influences on human health and greatly impacts human health because it is a whole group of influences collected under a single term. Nevertheless, also because, at a young age, we learn to expect to get ill in lousy weather. This is visible in illnesses whose name is based on the weather condition they seem to be connected to: A cold and heatstroke.

To represent the weather, the machine learning algorithm analyzes nine different features for each day:

- maximum temperature
- average temperature
- total precipitation
- snow height
- wind direction
- average wind speed
- maximum wind speed
- air pressure
- humidity
- hours of sunlight

The data is from meteostat[1], a website that collects meteorological data from multiple national weather services. It is provided by weather station 10803, located in Freiburg, Germany. All data is provided as a floating-point number with a single decimal place. The unit is the one provided by meteostat, which is shown in Overview

As those variables are closely connected, they influence each other as well. For example, the minimum, average, and maximum temperatures depend on each other, as seen in Figure 4.2. Still, they all introduce their own information that none of the others can consistently provide. The average cannot show extremes, while minimum and maximum only provide a single outlier. Therefore, all of them are used.

Another problem could be using only a single weather station, as temporary errors in their measurements would reduce the data quality. However, the number of weather stations in Freiburg is limited. Furthermore, the learning algorithm uses only daily averages, which counteracts short-lived errors.

[1] https://meteostat.net/de/station/10803 accessed on 10.01.2022

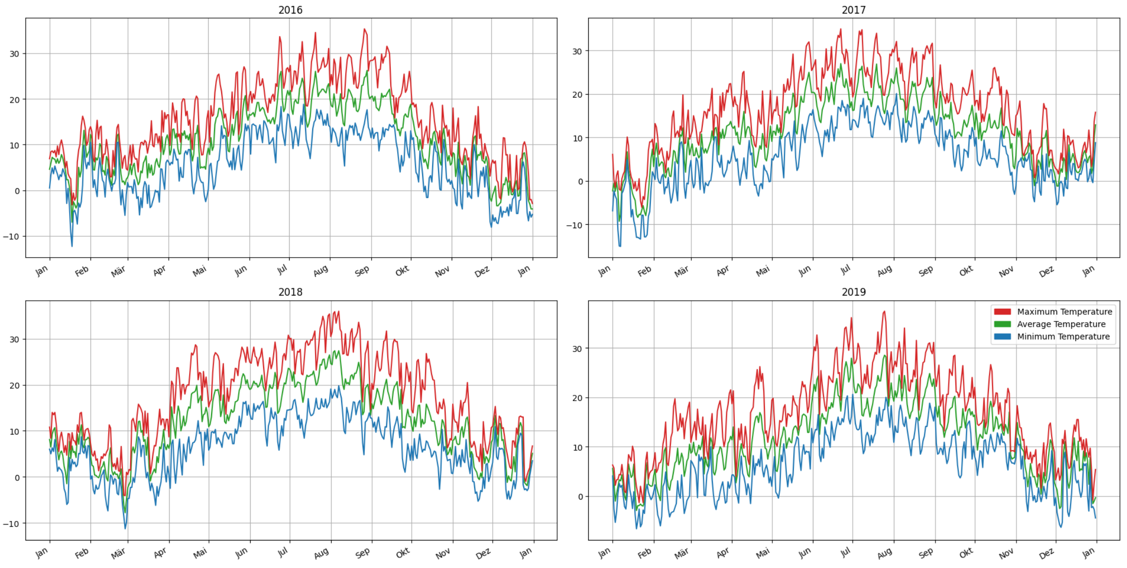


Figure 4.2: The minimal, average and maximum temperature in Freiburg, Germany from 2016 to 2019

Time

The date will be an essential data point for the system, as it is the key to joining the data from the hospital, weather, air pollution, lunar cycle, and holiday.

Other important data points are closely related to the date: The season, the day of the week, and the day of the month. This data needs no individual data source, as it can be calculated from the date itself. The day of the year can be represented by a number from 1 to 365. It also includes information about the season, as those are more or less arbitrary splits of the year. The exact dates on which the season switches do not affect day-to-day life but are yearly repeating times for similar diseases, like, the cold season.

The day of the month is probably less impactful, but in day-to-day life, many changes happen on a particular day in the month. For example, most people receive their paycheck or pay their rent on the same day of the month, which may impact their behavior.

The day of the week is essential. Most lives are scheduled around the seven-day rhythm of the week. Work and regular recreational activities happen on fixed days. Primarily the weekend provides a substantial shift in personal activity, which can impact the risk of injury and the ability to recognize illnesses.

All three pieces of information, day of the week, day of the month, and day of the year can easily be retrieved from the python time library[2]. It provides the information so that no special effort is needed to separate the length of different months or leap years.

[2] https://docs.python.org/3/library/time.html last accessed on 12.01.2022

It is possible to view the season as redundant to the weather data, as the cold of the winter is already reflected in the temperature and snowfall. Nevertheless, it is also valid to expect that people behave differently depending on the temperature they expect in the current time, ignoring the actual temperature. The day of the week and month, on the other hand, can have substantial impacts. People will encounter different dangers and, therefore, different reasons to go to an emergency room in their work and free time. Therefore, all three pieces of information are used to train the artificial neural networks.

Holidays

There can be quite a difference between injuries during work and free time. The holidays and school holidays can indicate work vs. free time for at least a big part of the citizens. As some holidays are based on religious or historical context, such as easter, they are not directly linked to a fixed date of the year. Therefore we use an individual Boolean for each day, indicating if it is a holiday.

This data is mainly provided by the website Feiertage API[3] and filtered to only contain the holidays of Baden-Württemberg. Additionally, Christmas Eve (24.12.) and New Year's Eve (31.12.) are marked as holidays too. Feiertage API does not mark them as holidays themselves. Probably because, in Germany, they are each only a half-holiday. Still, in the context of this thesis, they should be evaluated like any other holiday.

This still does not contain all holidays for that time, as the moveable holidays can be set differently by different schools and pre-schools. For the same reason, those holidays should not significantly impact the data, and as we cannot know about those holidays, they are not in the scope of this thesis.

Air Pollution

Air pollution has received increasing interest over the last few years. Especially in big cities, the air can be a health hazard. [GKG+98], [Zan06] and [CSSA95], like many others, showed the significant impact that air pollution can have.

The data is provided by the Landesanstalt für Umwelt Baden-Württemberg [4] and downloaded from the World Air Quality Historical Database[5]. It consists of the amount of ozone ($\mathrm{O_3}$) and nitrogen ($\mathrm{NO_3}$), as well as the concetration of fine ($\mathrm{PM_{2.5}}$) and inhalable ($\mathrm{PM_{10}}$) particles.

Each of them is in relation to a cubic meter of air. The air pollution data is visualized in Figure 4.3.

None of the enviromental influences mentioned before have a direct connection to the air quality and multiple studies connected air quality to decreasing health and accute medical conditions for many years [Zan06], [SSL+93], [CSSA95], [HLY+08], [LHO97], [SSB+81], [GKG+98]. Therefore, air quality is an essential factor to consider. The different kinds of air quality seem to have similar patterns, especially $\mathrm{PM_{2.5}}$, $\mathrm{PM_{10}}$, and O^2. Nevertheless, extreme values of individual air pollutants can be impactful, and all those pollutants were used in most referenced papers.

[3] https://feiertage-api.de last accessed on 12.01.2022

[4] https://www.lubw.baden-wuerttemberg.de/luft last accessed on 11.01.2022

[5] https://aqicn.org/historical last accessed on 11.01.2022

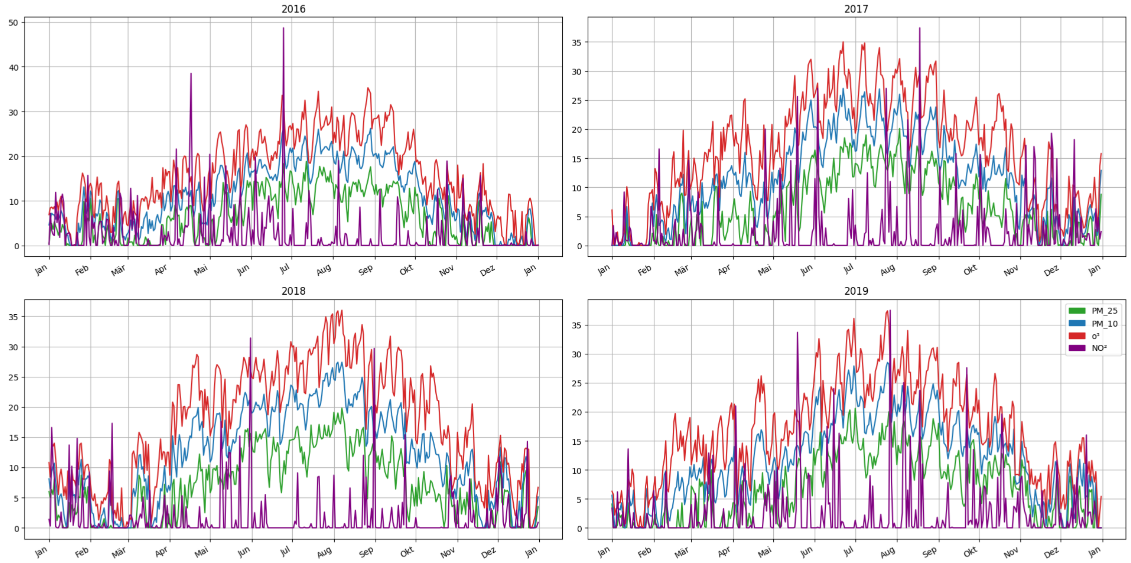


Figure 4.3: The air pollution in Freiburg, Germany from 2016 to 2019

Lunar cycle

While there is little to no scientific evidence that supports lunar effects on human health, the topic appeared in enough different discussions that it is interesting if the expectation alone can impact the emergency room. There is a widely established belief that the moon can influence daily life [Scu11]. If this belief influences the decision of people to do something that could harm them or go to the emergency room, it could influence the workload.

To include the lunar cycle in the dataset, it can be encoded as numbers between one and four:

1. New moon
2. First Quarter
3. Full moon
4. Last Quarter

The timeframe for the dataset starts on January 1, 2016. The first two days of 2016 were part of a complete moon phase, making the third of January the start of the last quarter phase. Each phase has an approximate length of 7 days. Therefore, numbers one to four are added in groups of seven to the dataset following January the third.

Twitter & Google Trends

The local Twitter and Google trends provide an overview of what happens in the local area. Mainly twitter can spread news faster than any traditional information channel. However, those fast and diverse pieces of information are difficult to monitor and analyze automatically.

The obvious way would be to generate categories, like fire, earthquake, or sports event, and map the trends into them. Nevertheless, that would cost diversity, as it cannot cover all information provided by Twitter. It would be more beneficial to monitor such factors over traditional ways, as this system is intended to run daily. The traditional channels, like TV, paper, or radio, provide more accurate information.

Additionally, exceptional events often receive their names on the internet, which generates a need for active moderation of the monitoring tool.

There are ways to analyze such trends automatically, but those are outside the scope of this thesis. Therefore, Twitter and google trends are not part of the feature set.

Overview

Table 4.1 provides an overview over all 20 Enviromental influences that we will use as input features for the artificial neural networks. It contains the identifier in the dataset, the name, and the unit in which the data is provided.

Short Name	**Name**	**Datatype**	**Unit**
minT	minimum temperature	Float	Celcius
maxT	maximum temperature	Float	Celcius
avgT	average temperature	Float	Celcius
pert	total precipitation	Float	mm per m^2
snow	snow height	Float	mm
winD	wind direction	Float	Degree
avgW	average wind speed	Float	km/h
maxW	maximum wind speed	Float	km/h
pres	air pressure	Float	hPa
humi	humidity	Float	Percentage
hsun	hours of sunlight	Float	Minutes
year	day of year	Integer	-
mont	day of month	Integer	-
week	day of week	Integer	-
hday	holiday	Boolean	-
$PM_{2.5}$	fine particles	Integer	Amount
PM_{10}	inhaleable particles	Integer	Amount
O_3	ozone	Integer	Amount
NO_3	nitrogen	Integer	Amount
cycl	lunar cycle	{1,2,3,4}	-

Table 4.1: Environmental influences used for training

4.2.1 Hospital Emergency Room Data

The supervised machine learning algorithm also needs sample data of what it should predict for supervised learning. Those samples are provided by the Universitätsklinikum Freiburg[6] in Germany. The provided data contains everyone who entered the emergency room from 01.01.2016 to 31.12.2019. That is 179515 cases over a span of 4 years or 1461 days.

However, the main diagnosis is only provided when the patient is taken care of stationary. Stationary cases are only 63070 of the total number of cases. This renders 179515 cases, or ≈ 64.87%, unusable.

There was no consent for study participation, as only anonymized data was required to answer the research question. On the one hand, informed consent in an emergency recording situation is complicated. In the case of retroactive consent, there could be an unacceptable bias due to deceased patients and patients without contact data.

On the other hand, it is needed to protect the patients' personal data (DSGVO Art 89 and BDSG §22 (1)) and honor medical confidentiality (§203 StGB). However, since only structured and anonymized data is analyzed, it is not to be assumed that a disclosure of secrets contrary to §203 (1) StGB takes place.

DSGVO Art 89 and BDSG §22 (1) provide, in deviation from DSGVO Art 9 (1), a justification for the processing of data for scientific purposes, insofar as this is necessary for the research or to answer the research question [DSGVO].

The analysis period is explicitly in the period before the Corona pandemic to assess the distribution of diagnoses without pandemic effects.

The data is provided in form of the following features

- Encounter_num
- admission_datetime
- esi_score
- inout_cd
- main_diagnosis
- age_at_admission
- sex_cd
- ZIP_cd

The age, sex, and ZIP code are metadata about the patient that does not need further explanation. The Encounter_num is a simple index. The admission datetime and the encounter number are only needed to match the medical data to the environmental influences from the same day. The ESI score, also called the ”Emergency Severity Index (ESI) Score”, represents how urgent it was to attend to this patient. inout_cd represents if a patient was treated stationary or if they were able to leave after a short inspection in the emergency room, which is called outpatient care.

[6] https://www.uniklinik-freiburg.de/de.html last accessed on 17.01.22

Most important for the machine learning system is the main diagnosis. It represents the injury/illness the patients were diagnosed with. It is classified by the ICE-10 system[7], which maps a wide area of diagnosis to nested identification codes.

The ICD-10 system contains over 1000 individual codes [Kro]. Most of them do not appear in this dataset. Therefore, a model trained on this data would be unable to predict those diagnoses. It simply does not know that they exist. Additionally, such a precise diagnosis is unrealistic and unnecessary. Not every ICD code needs its own professional. Some of them even describe the same injury, just with different intensities.

The ICD-10 codes are hierarchically categorized, beginning with 22 top-level categories.

Not all of them appear on the dataset. The distribution of cases into these categories can be seen in Figure 4.4.

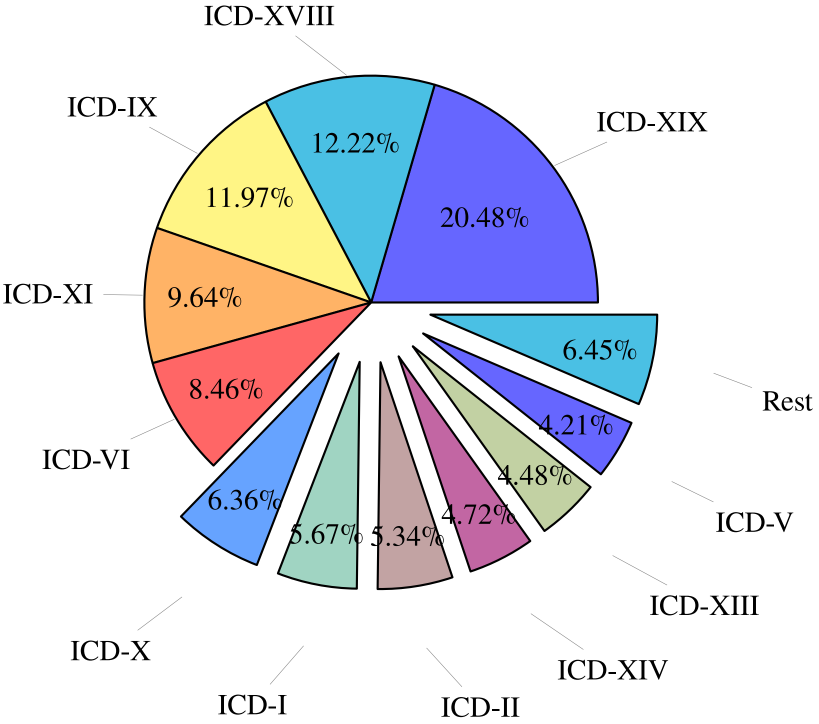


Figure 4.4: The distribution of cases in the dataset between the different ICD-10 top level categories by percent

The *Other* label contains the top-level categories III, IV, VII, VIII, XII, XV, XVII, and XXI. Category XXII was not listed as the main diagnosis for a single case in the data, including outpatient care.

An optimized categorization was created with the doctors from the Universitätsklinikum Freiburg. It focuses on two factors: Firstly, categories should be as equal as possible regarding the number of cases they represent. Categories that are too big can reduce the quality of the model. Secondly, the categories should split the cases to be useful for staff planning, meaning that each category should contain as few different expert fields as possible.

[7] https://www.icd-code.de/icd/code/ICD-10-GM.html last accessed on 17.02.22

The official top-level categories are already a good separation of the cases, which fulfills the second requirement. To make the predicted classes more equal in size, we extract diagnoses with more than 1500 individual cases and create a rest class, which contains all categories with less than 1000 cases on its own. This process creates the distribution seen in Figure 4.5.

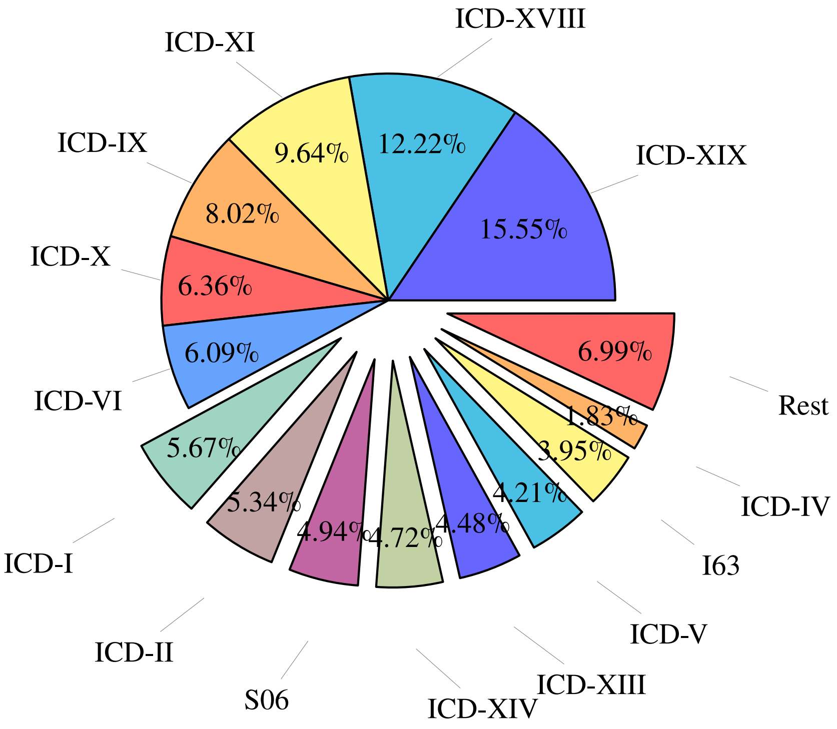


Figure 4.5: The custom distribution, including all cases that appear more than 1500 times

This introduced three diagnoses as their categories:

S06 Intracranial injury, better known as concussion or brain injury, is the most common injury in the data that had to be taken care of for stationary. It is part of the official category XIX and provided a significant part of its cases [Kro].

I63 Cerebral infarction or brain infarction originates in the original category 9 [Kro].

G45 Cerebral transient ischemia and related syndromes occur when the blood supply to part of the brain is briefly interrupted. Originally it was part of category VI [Kro].

4.3 Normalization

As the input data (environmental influences) and output data (hospital emergency room data) have quite a range, a normalization can improve the model quality. The environmental influences do not influence each other. Therefore they are individually normalized. Each influence gets divided by its absolute maximum, resulting in values between -1 and 1:

$$\vec{in}_{norm} = \vec{in} \cdot \frac{1}{max(|\{xin\vec{in}\}|)}$$

The hospital emergency data is normalized by a single value, as the values are closer together and closely connected. Therefore, each category is divided by the maximum of all categories, resulting in values between 0 and 1, as there are no negative workloads.

4.4 Timeframe

To predict the workload of an emergency room's inpatient numbers, we first have to decide which timesteps we want to predict.

When the timeframe is too small, some disadvantages occur. While a prediction every second could be helpful, most environmental influences are not measured that frequent, making it impossible to generate such a precise dataset. Additionally, no ICD-category has a patient every second. This would leave the hospital dataset with an overwhelming amount of zeros, which would make the learning process extremely difficult for the artificial neural network. It would be easy for the model to reach a good score by only predicting zeros all the time, as it would only be wrong in a fraction of cases.

On the other hand, making the timeframe too big, for example, yearly predictions would make the information useless for the use-case of this thesis. The training would also be pointless, as we only have four years of data. With only four data points, the network could not find significant patterns.

Therefore we need a middle ground between those two extremes. In cooperation with the Universitäts-Klinikum Freiburg, we identified daily predictions as our best options. All environmental influences are provided on at least a daily basis, and the information would be useful in staff planning. For some categories, the daily number of cases is relatively low. Therefore, we also train all models on weekly predictions to see if more data would significantly increase prediction quality. In the weekly dataset, we use the average of daily values, except for minT, maxT, and maxW, for which we calculate the maximum or minimum value of that week.

4.5 Artificial Neural Network Architecture

4.5.1 Feed Forward Network (FF)

The first model is a simple feed-forward design, a multi-layer perceptron (MLP). It has an input layer with 20 nodes, one for each environmental influence, and an output layer with 15 nodes for each predicted category. Between are three hidden layers. The number of nodes in those layers is chosen through tuning later (see Section 5.5). Each layer is fully-connected with its predecessor and follower. A simplified version of the resulting architecture can be seen in Figure 4.6.

We choose this architecture for its simple design. It makes development easy and allows problem-solving with fewer variables than in a more complex architecture, meaning less data is needed for successful training. Despite its simple design, there is no reason to believe that a multi-layer perceptron cannot predict the emergency room workload, as it was used by Lu et al. in [LBX+21] successfully.

This thesis identifies this model as FF or feed-forward model.

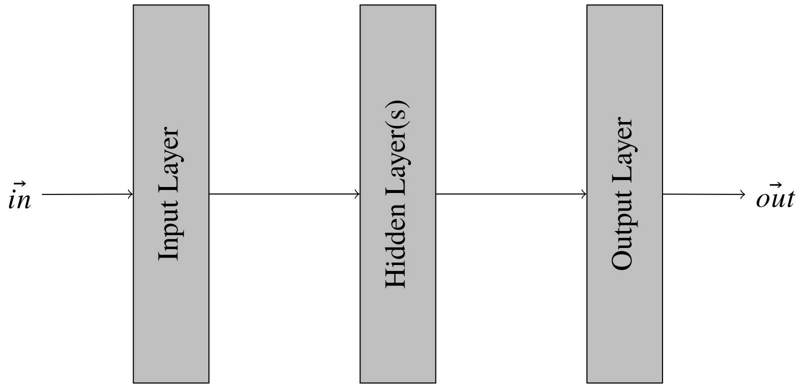


Figure 4.6: Simplified architecture of the feed forward network, as implemented for this thesis

This network is not expected to perform spectacularly without a timelag, as multiple papers found the strongest correlations between environmental influences and human health with a small time lag of three to four days [PCZ+20] [HLY+08]. Still, this network provides an interesting base to compare the other networks.

4.5.2 Total number of Patient Prediction Network (TNOP)

Additionally, to the workload prediction of the different diagnosis categories, it is of interest to predict the total workload of the emergency room per day. We designed a multi-layer perceptron based on the FF model, with the difference that it has only a single output neuron. It receives a modified dataset in which the sum of all cases replaces the prediction categories. It is a less helpful prediction for the hospital, but it allows us to compare our results to Lu et al. in [LBX+21]. We call this model the TNOP model.

4.5.3 Focus Prediction Network (FOCUS)

Another simplified version of the feed-forward model is the FOCUS model. Like the TNOP model, it is a multi-layer perceptron with only a single output neuron. This time, the dataset is modified to only contain the ICD-10 category ICD-XIX as output. Therefore it only predicts the number of patients with ”Injuries, poisonings, and certain other consequences of external cause ”[Kro, translated][8]

This network should show if the prediction quality increases if only a single ICD-10 category is predicted by an individual model instead of one that predicts all of them. We choose ICD-XIX, as it is the biggest category and has neither any entry with zero cases in the daily nor the weekly timeframe, making the learning process easier as zeros can have problematic interactions with some metrics.

We call it the FOCUS model.

4.5.4 Recurrent Network (LSTM)

To improve the prediction quality, another artificial neural network is developed. The FF, TNOP, and FOCUS model used a feed-forward architecture with only data from the same day as the prediction. But like [PCZ+20] and [HLY+08] proposed, a 3 to 5-day timelag could improve the prediction quality. We could modify the dataset so that a multi-layer perceptron is trained with multiple data from multiple days, but some architectures are especially good with sequenced data: The so-called Recurrent Neural Networks (RNNs) analyze their data in multiple steps while ”remembering” parts of it by using, for example, a ”Long Short Term Memory”(LSTM) layer. This gives it a distinct advantage over feed-forward networks in sequence-related tasks. Its architecture can be seen in Figure 4.7

We choose a three-day timelag as each day of the timelag reduces the dataset by one. This happens because, for each additional day in the timelag, we need an additional day of data before the first day that appears in the resulting dataset.

The data needs to be reformatted to include the three-day timelag in each data point. The new format has an additional dimension. While the dataset for the multi-layer perceptrons only had two dimensions (features x timesteps), this dataset has an additional dimension for the timelag entries. This is further explained in Data Structure

[8] Original: ”Verletzungen, Vergiftungen und bestimmte andere Folgen äußerer Ursachen”

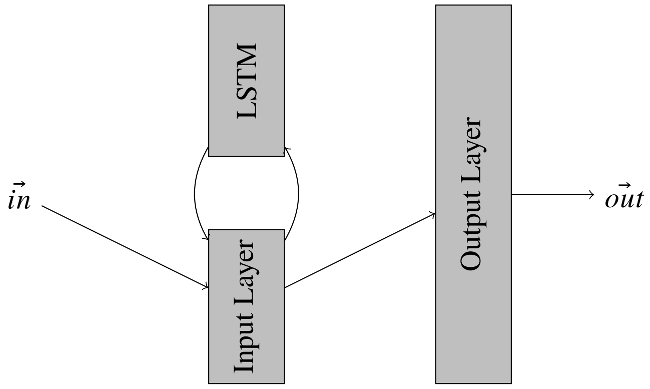


Figure 4.7: Simplified architecture of the recurrent network, as implemented for this thesis

5 Implementation

This chapter covers notable parts of the implementation. It presents the used technology and the most important problems.

5.1 Technology

To implement the neural networks presented in Chapter 4, we use Keras[1].

Keras is a high-level library for TensorFlow 2[2] in python that supports the rapid development of deep learning neural networks. It makes the development of neural networks easier without losing functionality. TensorFlow 2 is one of the most well-known machine learning libraries. It contains a lot of different tools to develop state-of-the-art machine learning models.

To identify the best hyper parameters, we will use Keras tuners[3]. Tuners are a Keras module that allows training models multiple times in quick succession to identify the best possible hyperparameters.

5.2 Code Example

The Algorithm 5.1 shows the Keras implementation of the feed-forward network. It is contained in a function that receives a hyperparameter object *hp* to create a model with hyperparameters dictated by the tuner. In this case, the hidden layer's size and the learning rate are tuned.

The variable *model* contains the three layers that make up the neural network. Each entry of the list contains a layer object. Those themselves are initialized with their amount of nodes, activation function, and name. The first layer receives an additional parameter specifying the input shape of the data.

Before returning the model, it gets compiled, sets an optimizer for use during training, and sets the metrics it will be measured.

For further information on how to use Keras, we recommend the Keras documentation[1].

[1] https://keras.io/

[2] https://www.tensorflow.org/

[3] https://keras.io/keras_tuner/

Algorithm 5.1 The implementation of the feed forward network

```
    def build_model(hp):
    amount = hp.Int('hidden layer', min_value=1, max_value=80, step=1)
    lr = hp.Float('learning rate', min_value=0.15, max_value=0.35, step=0.001)
    opt = hp.Choice('optimizer', ["Ada", "SGD"])
    layers_func = [hp.Choice('layer_1', ["tanh", "relu"]),
                   hp.Choice('layer_2', ["tanh", "relu"]),
                   hp.Choice('layer_3', ["tanh", "relu"])]

    return build_model_instance(amount, opt, lr, layers_func)


def build_model_instance(amount, opt, lr, layer_funcs):
    act_funcs = []
    for i in layer_funcs:
        act_funcs.append(activations.relu if i == 'relu' else None)
        act_funcs.append(activations.tanh if i == 'tanh' else None)

    if opt == 'Ada': opt = tf.keras.optimizers.Adadelta(learning_rate=lr)
    if opt == 'SGD': opt = tf.keras.optimizers.SGD(learning_rate=lr)
    model = keras.Sequential(
        [
            layers.core.Dense(19, activation=act_funcs[0], input_shape=(19,), name="input"),
            layers.core.Dense(amount, activation=act_funcs[1], name="hidden-1"),
            layers.core.Dense(amount, activation=act_funcs[1], name="hidden-2"),
            layers.core.Dense(amount, activation=act_funcs[1], name="hidden-3"),
            layers.core.Dense(output_size, activation=act_funcs[2], name="output")
        ]
    )
    model.compile(optimizer=opt,
                  loss=tf.keras.losses.MeanAbsoluteError(),
                  metrics=['mean_absolute_percentage_error', 'mean_absolute_error']
                  )
```

5.3 Data Structure

The data was formatted in two ways. First as a two-dimensional data structure for the multi-layer perceptrons (FF, TNOP, FOCUS) and then as three-dimensional for the LSTM model.

The two dimensional dataset, as seen in Table 5.1, contains 1461 × 20 entries for daily predictions and 209 × 20 for weekly predictions

The three-dimensional dataset can be imagined similarly, except that each line is its small table containing the entries of three successive days/weeks. Therefore it contains 1459 × 20 × 3 entries (207 ×20 ×3 for weekly predictions) to represent the timelag. The size was reduced by two, as two days do not have their predecessors in the data.

index	minT	maxT	...	NO_3	cycl
0	4.6	6.9	...	0.2	3
1	5.9	8.4	...	7.2	3
2	7.2	8.6	...	8.6	4
...	...	...	...	...	...
1460	-1.1	2.8	...	0	4
1461	-0.4	5.4	...	5.4	4

Table 5.1: The 2D datastructure for the enviromental influences

Before training, the dataset is further split into a training set (90%) and a validation set (10%). The validation set is used to test the results individually. It is removed manually from the dataset before Keras receives it so that the network never trains with it. Most of the evaluation will take place with tests on this validation set.

The training set is used to train the artificial neural network. It contains the data that the model sees in each epoch. Keras splits it in another train/validation split of 90% / 10% so that each epoch can be validated individually. Therefore, the actual training happens on 81% of the entire dataset, 167 entries on weekly predictions, and 1183 entries on daily predictions (166 and 1181 in the 3D dataset).

5.4 Implementation Problems

5.4.1 Secret Data

Working with patients' personal data is inherently difficult. Under German law, they need special protection. As required in Section 3.1.2, to protect the personal data, the authors of this thesis never gains access to the data itself. The data is kept on the computers of the Universitätsklinikum Freiburg, and only the doctors can access it. This creates a somewhat complicated pipeline to deploy the machine learning model.

First, the code is packaged in a docker image. This image contains all the machine learning system parts, including the input dataset and code to reformat the patient data. The docker image is then provided and sent to the hospital, where a doctor starts it on a computer with the patient data. The data is reformatted and matched with the input data automatically. Then the machine learning models are trained with the generated dataset multiple times to identify the optimal hyperparameters. The results are sent back to the developer. The data understanding used a similar process.

As this process contains multiple manual steps, it can be relatively slow. At maximum, the code could only be run 2-3 times a day. Additionally, debugging was further complicated as it was impossible to see data lines that created problems or run a debugger over the code while working with real data.

Pandas[4], the tool we use to handle the data, does not provide three-dimensional data structures. It is two-dimensional at maximum. Therefore, the data needs to be converted into another structure. Numpy ndarrays[5] to be precise. Ndarrays provide a ”multidimensional, homogeneous array of fixed-size items” [Dev22]

5.4.2 Extrem Overgeneralisation / Underfitting

In early experiments, the models consistently predicted an average instead of individual predictions. This happens because the problem that the network tries to learn is quite complex, and the average provides decent results. This was solved by increasing the degrees of freedom the model has. We increased the number of hidden layers from one to three and explored more nodes and epochs. Initially, we tested only up to 32 hidden nodes and 64 epochs. With both significantly increased, the models started to improve.

5.5 Tuning

An artificial neural network contains several hyper-parameters. A hyperparameter is a value that decides how the network behaves in the learning process [TBST21]. The networks in this thesis have the following hyper-parameters:

- **Learning rate** The learning rate defines the size in which a single learning step can change the weights of the nodes. It is the most important hyper-parameter in the context of exploration versus optimization. A learning rate that is too high could miss the optimal configuration, while a learning rate too low is unable to reach an optimum at all [Kub17].

  The tuner evaluates training rates between 0.001 and 0.5.

- **Optimizer** The optimizer is the algorithm that applies changes during training. It uses the loss at the end of a batch to change the weights in each layers [Seb21].

  For this thesis, the optimizers adaDelta and SGD are possible candidates. SGD, or stochastical gradient descent, is a variant of gradient descent that is more viable with small datasets[Rud16]. adaDelta is an extension of AdaGrad, which allows the optimizer to choose the learning rate itself, but is computationally expensive based on the current loss [Rud16].

- **Activation Function** The activation function decides if the inputs are enough for the neuron to *fire*. [HDR19] recommends a rectified linear unit (or relu) function for hidden nodes. Additionally, we use hyperbolic tangent (or tanh) for input and output layers, as it was used by [HDR19] together with the relu function.

- **Size of the hidden network** The amount of nodes in the input and output layer of the network is defined by the size of the input and output of the network. The hidden layer is under no such constraint. Therefore, the hidden layer’s size can be changed to increase the quality of the model [Seb21].

[4] https://pandas.pydata.org/

[5] https://numpy.org/doc/stable/reference/generated/numpy.ndarray.html

For this thesis, a size between 1 and 150 nodes is tested.

- **Epochs** An epoch is a training run with the whole dataset. Doing multiple epochs can increase the quality of the model as it is trained more. However, it also increases the risk of overfitting, as the same data is used multiple times for training [Kub17].
- **Batch size** To increase the speed of the training process, the data can be applied in batches. While this increases the speed of the training, it can reduce the quality of the model [Seb21]. The dataset this thesis uses for training is tiny, so the batch size can remain at one, avoiding this trade-off altogether.

For hyper-parameter that cannot be set logically, like, in this thesis, the batch size, it is common practice to test through the different combinations. To do this, Keras provides a library called Keras tuner [OBL+19]. With Keras tuner, a model can be trained and evaluated automatically. The developer only has to provide a method that creates a model, where the hyperparameters can be defined during runtime, as seen in Algorithm 5.1.

To identify the optimal hyperparameters, the network tests 300 combinations of the learning rate, hidden layer size, and optimizer. Each combination is tested five times to average the outcomes.

Epochs

The used Keras turner cannot tune the number of epochs in which the model is trained. Therefore, the epochs are tuned manually after all other hyper-parameters are fixed. To identify the optimal number of epochs for each model, each artificial neural network is trained to 640 epochs. Keras uses a small test split to test the progress of each epoch. The results of these tests can be analyzed to identify the optimal number of epochs for the model. Figure 5.1 shows such a run for the LSTM model. The x-axis represents the number of epochs it has trained, while the y-axis shows the average error in the small validation set. The aim here is to reduce that error as far as possible.

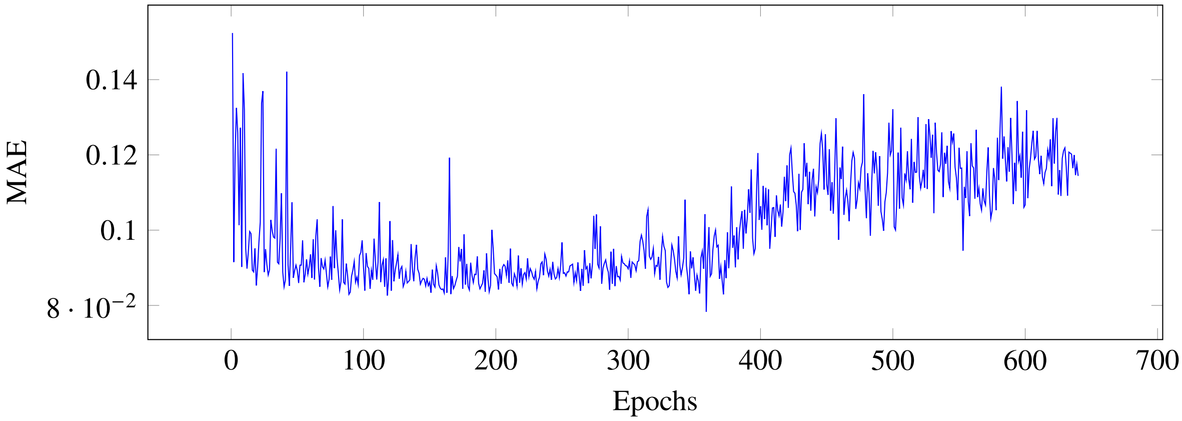


Figure 5.1: The validation MAE of the LSTM network using daily data over 640 epochs

It is visible that the error gets close to the minimum between 50 and 60 epochs. It stays there until it rises again between 300 and 400 epochs, indicating the start of overfitting.

We now use this area between 50 and 350 epochs to further test the optimal epoch number by spot testing different sizes. This way was chosen because it provides a wide range of tested epochs. The added advantage is that the test can be repeated to ensure consistency without testing every possibility individually.

Optimal hyper parameter

Table 5.2 provides an overview of all hyperparamters for all models after tuning.

model	learning rate	optimizer	activation function	hidden network	epochs
FF(daily)	0.4	SGD	[tanh, relu, tanh]	60	300
FF(weekly)	(0.2)[6]	Ada	[tanh, relu, tanh]	60	300
FOCUS(daily)	0.4	SGD	[tanh, relu, tanh]	60	300
FOCUS(weekly)	(0.2)[6]	Ada	[tanh, relu, tanh]	60	300
TNOP(daily)	0.1	SGD	[tanh, relu, tanh]	120	64
TNOP(weekly)	0.45	SGD	[tanh, relu, tanh]	140	64
lstm(daily)	(0.1)[6]	Ada	[tanh, relu, tanh]	140	64
lstm(weekly)	(0.45)[6]	Ada	[tanh, relu, tanh]	120	64

Table 5.2: Overview of the optimal hyper parameters for all artificial neural networks

[6] Selected by the tuner but ignored by AdaDelta

6 Evaluation

In this chapter, we evaluate the results of our experiments with the four artificial neural networks.

6.1 Metric

The precision of the models is measured with two different metrics that complement each other. The Keras API directly provides them to evaluate artificial neural networks. The first metric is mean_absolute_percentage_error[1]. It provides the difference between the expected value and the predicted value in comparison to the expected value[Swa00]:

$$MAPE(y_{expected}, y_{predicted}) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_{expected,i} - y_{predicted,i}}{y_{expected,i}} \right|$$

This can show how precise the model is, but without knowing the expected value, it is not very helpful alone because percentages of tiny numbers can increase quickly. Therefore, the model is also measured with the mean_absolute_error [2]. It provides the absolute difference between the expected value and the predicted value.

$$MAE(y_{expected}, y_{predicted}) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| y_{expected,i} - y_{predicted,i} \right|$$

Lu et al. [LBX+21] also used those metrics.

6.2 Baseline

To identify the quality of the results, we need a baseline to compare them to. We compared the models to a heuristic that always guesses the average out of the given Search Space.

$$MAPE_{avg} = \sum_{p=min}^{max} \left(\frac{|p - avg|}{p} \right) \cdot \frac{1}{p}$$

[1] https://keras.io/api/metrics/regression_metrics/#meanabsolutepercentageerror-class

[2] https://keras.io/api/metrics/regression_metrics/#meanabsoluteerror-class

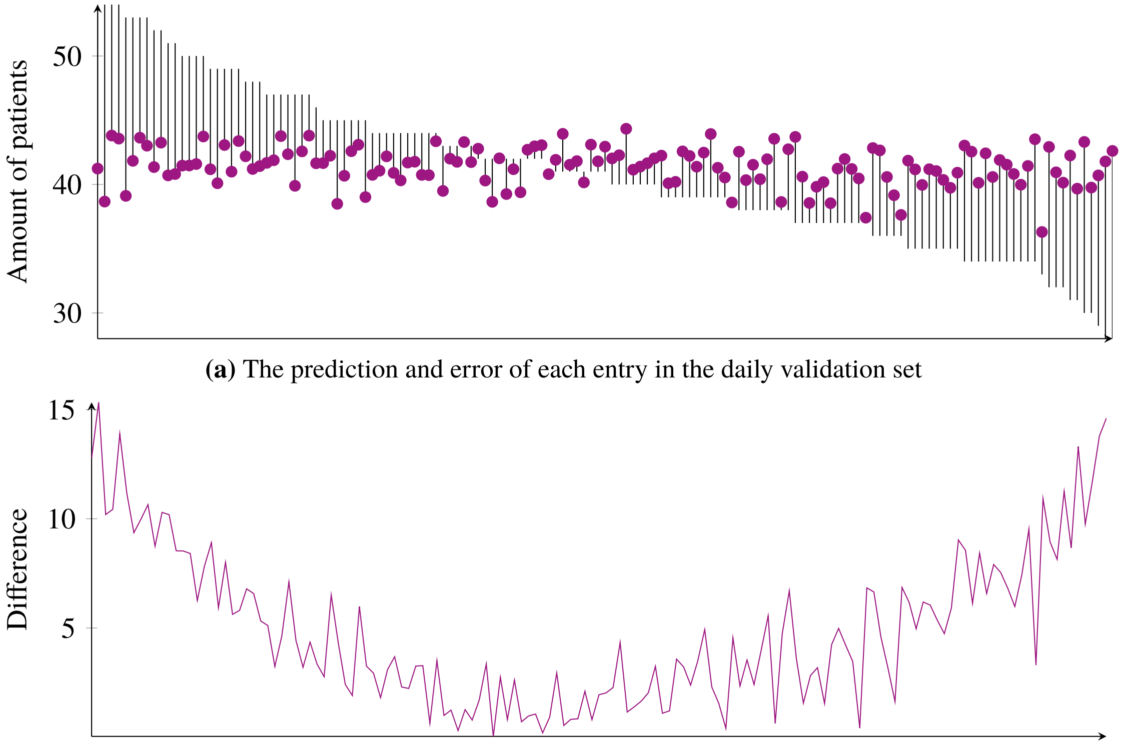
$$MAE_{avg} = \sum_{p=min}^{max} (|p - avg|) \cdot \frac{1}{max}$$

6.3 Results

In this section, we will talk about the results of our experiments. To retrieve most of the data, we trained each network five times and calculated the average of the results. Additionally, we present the results of a single run in detail.

6.3.1 Evaluation of the TNOP model

The TNOP model is intended to predict the total number of patients that appear over a day or week. The model deviates from the expected value on average by $MAE \approx 5.8$ patients on daily predictions. The relative deviation is $MAPE \approx 14.5\%$. Surprisingly, throughout all experiments, the validation set provides slightly better absolute results: Having an $MAE \approx 5.33$ patients and $MAPE \approx 14.79\%$.



(a) The prediction and error of each entry in the daily validation set

(b) The Difference between the expected and the predicted value for each entry in the daily validation set

Figure 6.1: Results of one daily TNOP experiement

In the years 2016 to 2019, the Universitätsklinikum Freiburg had 22 to 66 patients daily, with a daily average of 41 patients. Resulting in an expected baseline of $\text{MAE}_{\text{avg}} \approx$ 7.8 patients and $\text{MAPE}_{\text{avg}} \approx$ 19.14%. This shows that, on average, the TNOP model is ≈ 2.47 patients closer to the expected value than the baseline. This result is comparable with [LBX+21], which also predicted a workload of an emergency room and had a MAPE of 14.14% when using a multi-layer perceptron and a time-lag of 1 day.

Figure 6.1a provides a visual representation of the validation set of a single experiment. The x-axis represents the 127 entries in the validation dataset, sorted from the highest to the lowest expected value. The y-axis represents the number of patients. For each entry, the colored point represents the prediction made by the artificial neural network. Purple indicates a daily prediction, while blue indicates a weekly prediction. Additionally, each prediction has a line pointing at the expected value it tried to predict, making the length of the line a visual representation of the absolute error MAE. The data is sorted from highest to lowest for the expected values. Figure 6.1b shows the relative difference (MAPE) for each entry in the validation set. The data is sorted identically to Figure 6.1a.

The model predicts all values between 35 and 45, while the data provides examples between 22 and 66. Nevertheless, the model adapted to the data, even if only slightly. It is shown by a linear trendline on the data. In all experiments, this trendline has a negative gradient < -0,01. This shows that the prediction is higher when the expected value is higher and lower when the expected value is lower, but it is a marginal difference from the mean.

The problem is that the adaption variance is relatively high. The model is very accurate for patient numbers often encountered and wildly inaccurate for values it encountered rarely.

Figure 6.1b shows again that the difference is higher on the ends of the plot, where the extreme patient numbers (high and low) are. Figure 6.2 shows the distribution of daily workloads in the dataset. Extremely high and extremely low workloads rarely appear in the dataset. Data that appears rarer than others is more difficult for an artificial neural network to learn, making the results regarding that data more inaccurate.

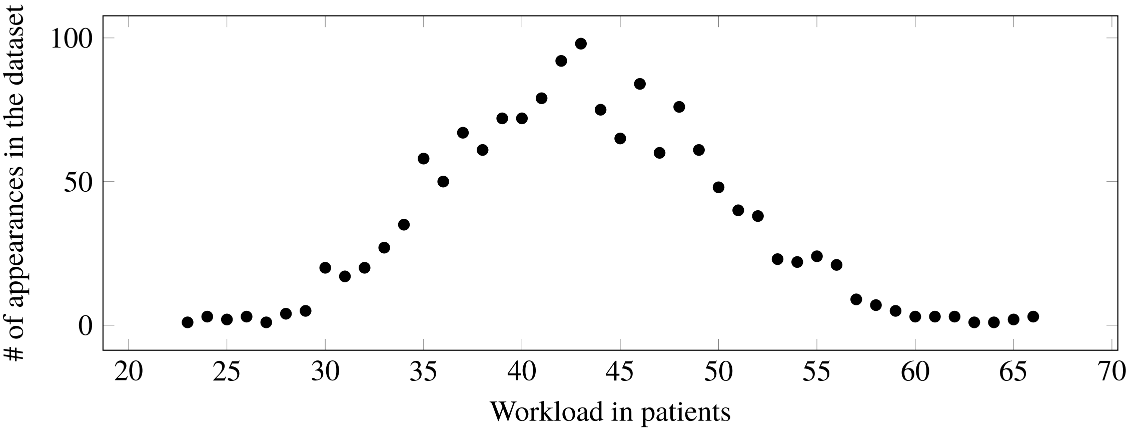
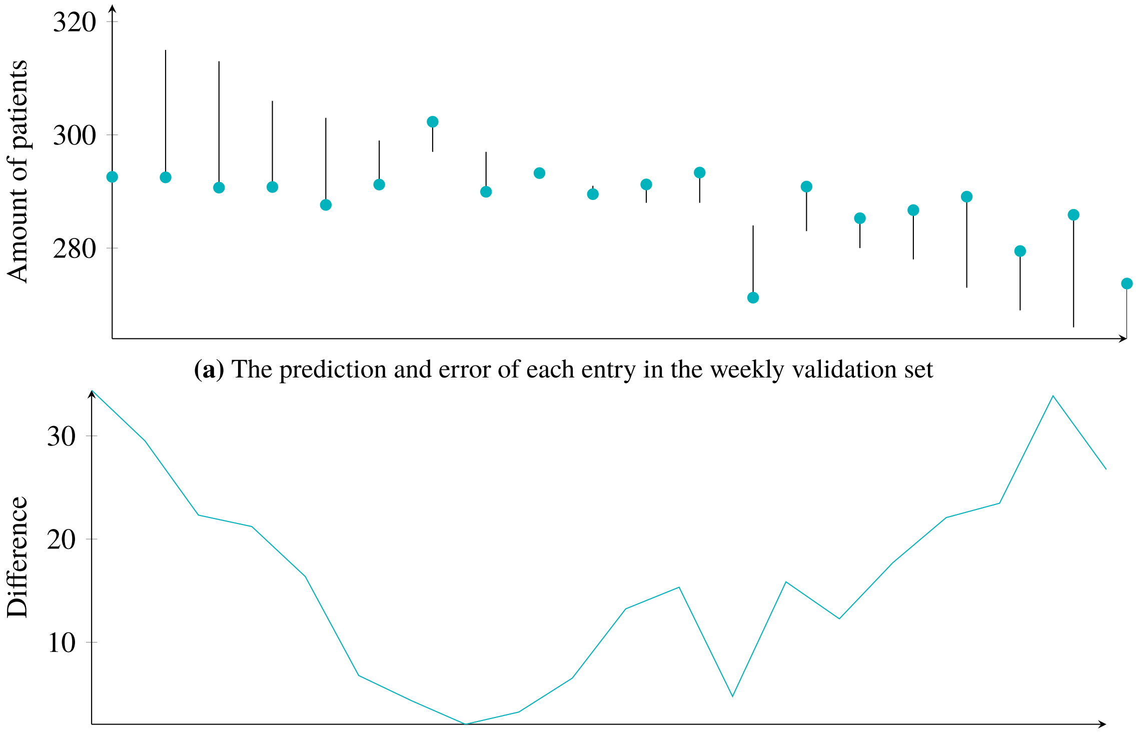


Figure 6.2: The daily appearance of different workloads in the hospital

Additionally, as shown in Section 5.5, the network needs many epochs to predict something else than the average for each input. The artificial neural network initially only predicted the mean as described in Section 5.4.2. With more epochs, the size of the prediction area grows from a single value to a range around the mean. However, if the network is further trained with the amount of data we have, increasing this range further would only result in overfitting. Therefore the only way to increase the quality of the model from this point is to increase the dataset, especially with more data regarding the extreme workloads.

Another support for this hypothesis is the difference between the training set's prediction quality and the verification set. As the verification set only contains 10% of the total data, it is less likely to contain extremes that increase the average deviation.



(**a**) The prediction and error of each entry in the weekly validation set

(**b**) The difference between the expected and the predicted value for each entry in the weekly validation set

Figure 6.3: Results of one weekly TNOP experiment

The prediction accuracy increases when the model is trained with weekly data instead of daily. Figure 6.3 provides the details of exemplary validation sets, similar to Figure 6.1. At first, the deviation seems to be higher than the one of the daily models. Nevertheless, that is a misconception. As the weekly prediction works with the sum of 7 days of patients in each entry, each entry has significantly more patients than any daily entry. This results in a more significant absolute error, even when the prediction quality is similar or better.

The weekly prediction has an average deviation of $MAE \approx 27$ patients (or $MAE \approx 21$ patients in the validation set). Relatively, this is a deviation of $MAPE \approx 8.29\%$ from the expected value (or $MAPE \approx 7.45\%$ in the validation set).

From 2016 to 2019, the Universitätsklinikum Freiburg had 145 to 395 patients per week, with an average of 287 patients per week. This average is more than 3.5 times bigger than the daily average, while the weekly MAE is not even doubling. Additionally, the relative error is nearly half on weekly predictions than on daily predictions. The weekly baseline MAE_{avg} even increases to $\approx$ 40.61 patients weekly. The weekly baseline also has a slightly smaller relative error with $\text{MAPE}_{\text{avg}} \approx 17.96\%$

Figure 6.4 shows that the weekly dataset has a significantly better distribution of the workloads. It has only two extreme outliers, and each number of patients appears between one and seven times. This is a better base for the artificial neural network to train on, as visible in the results. The adaption to the data is not perfect, but it is significantly better than the daily one.

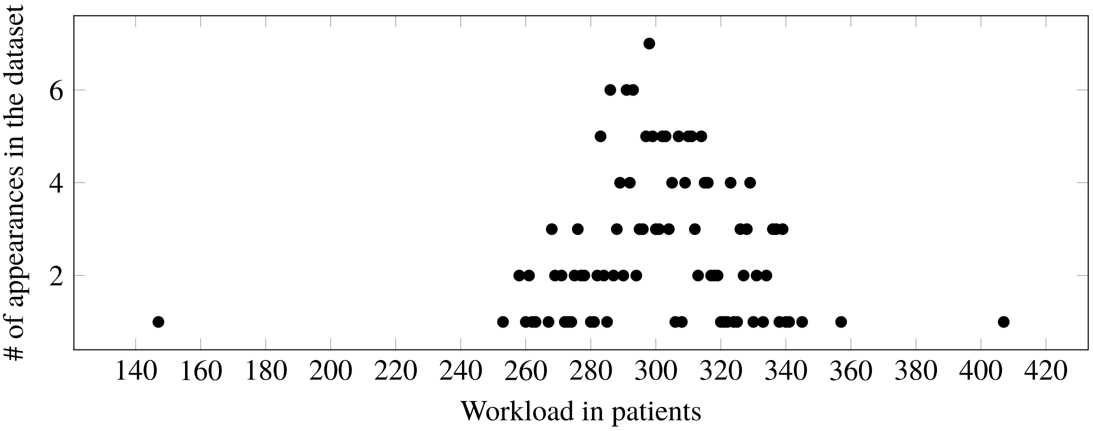
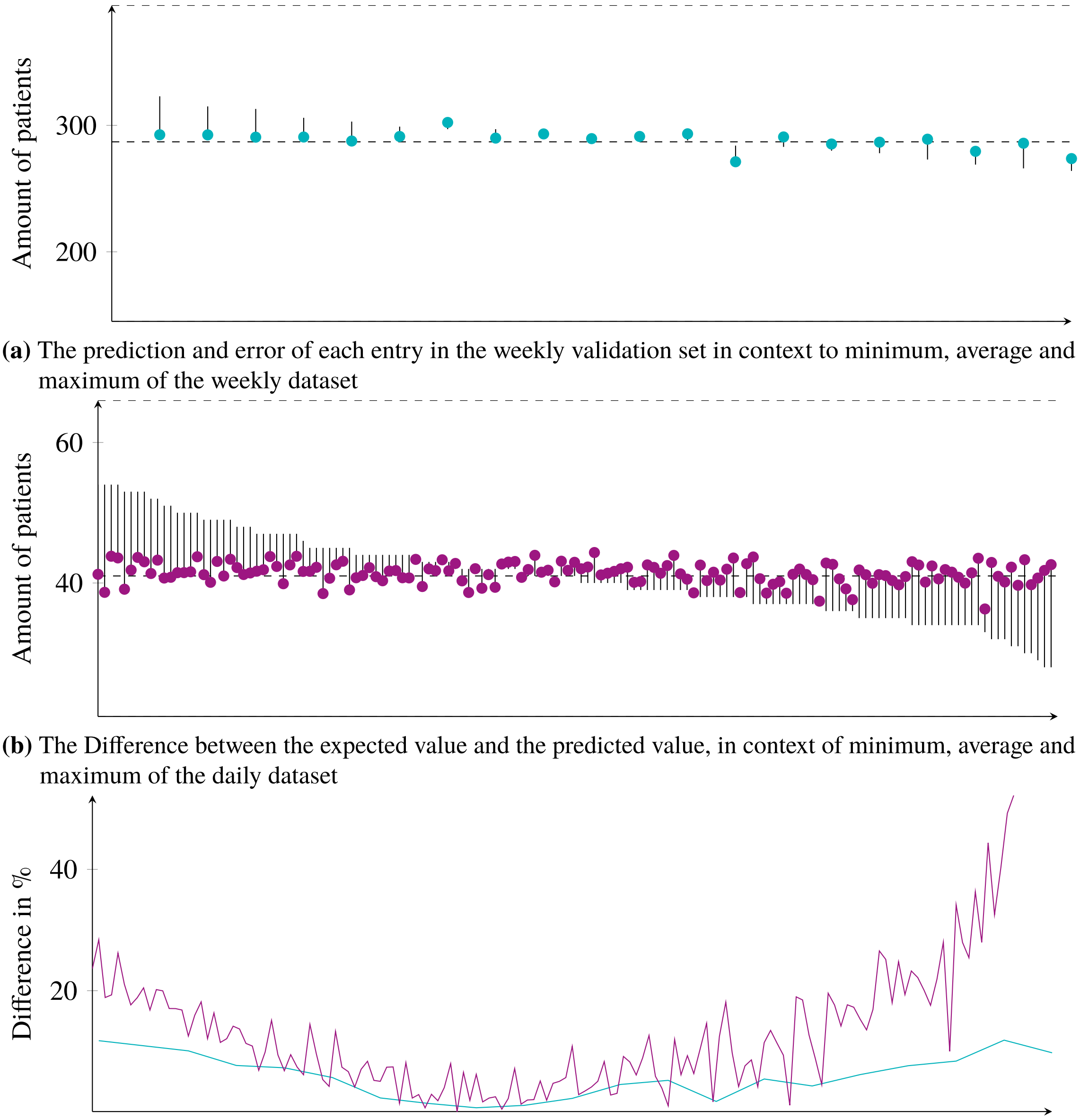


Figure 6.4: The weekly appearance of different workloads in the hospital

Figure 6.5 provides a more understandable comparison of the two networks. Figure 6.5a and Figure 6.5b are copies of the plots shown above, with their y-axis representing the range of the respective datasets. This creates a better comparison of the error and possible Search Space for the model. It shows that the weekly prediction provides better results relative to its Search Space. It is also visible in Figure 6.5c, which compares the mean absolute percentage error between both models with each other.



(a) The prediction and error of each entry in the weekly validation set in context to minimum, average and maximum of the weekly dataset

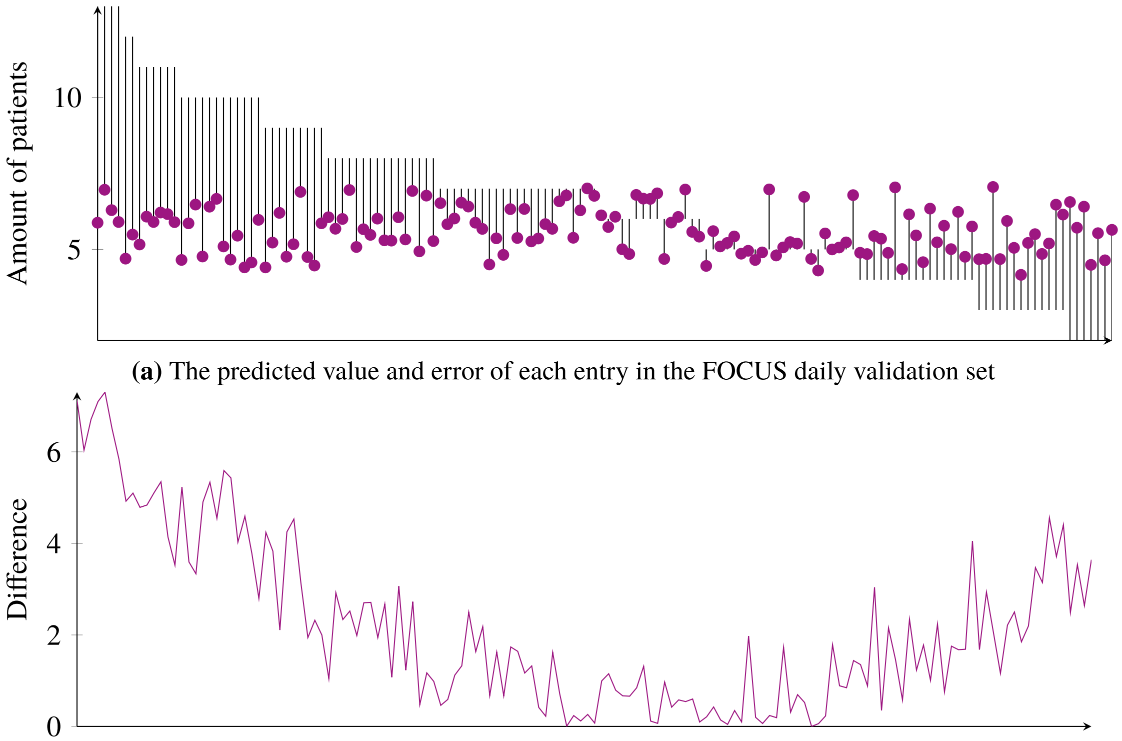
(b) The Difference between the expected value and the predicted value, in context of minimum, average and maximum of the daily dataset

(c) The Difference between daily and weekly TNOP predictions in percentage deviation

Figure 6.5: Comparision of the daily and weekly TNOP network

6.3.2 Evaluation of the FOCUS Network

The FOCUS network only tries to predict the number of patients that can be sorted into the category ICD-XIX. When the timeframe is set to daily, the training prediction deviates from the expected value by $MAE \approx 2.23$ patients. The validation dataset has slightly higher precision. It deviates on average only by $MAE \approx 2.21$ patients from the expected value.



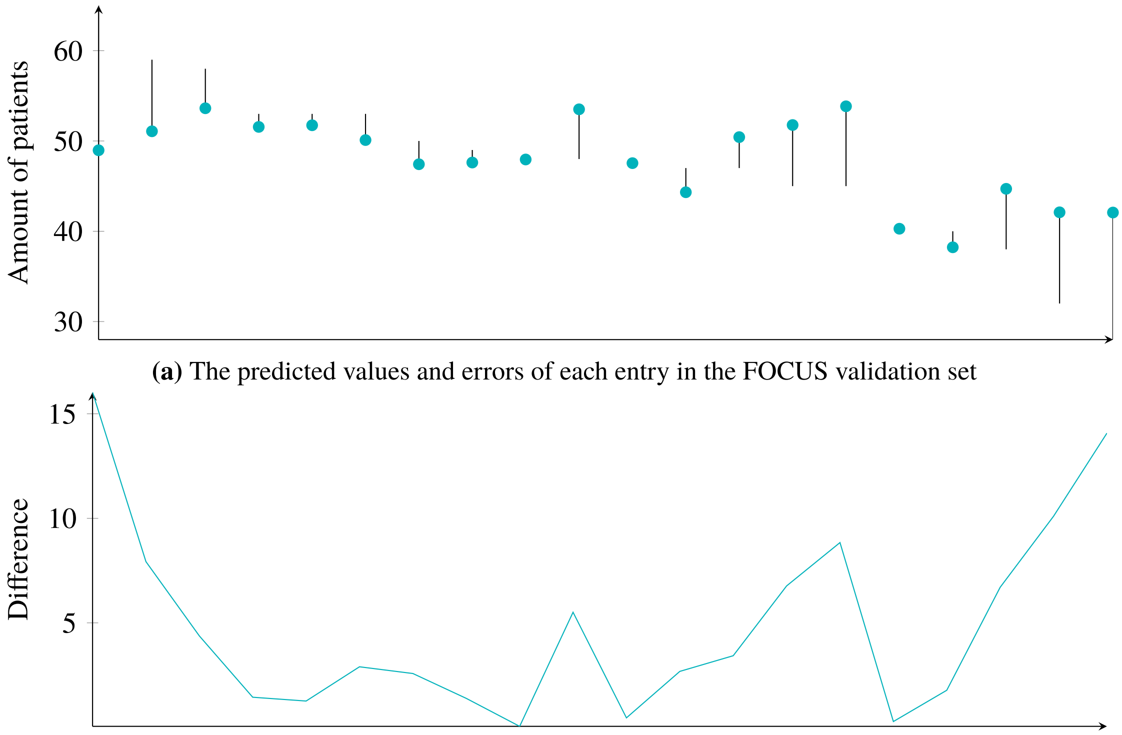
(a) The predicted value and error of each entry in the FOCUS daily validation set

(b) The comparison between the expected value and the predicted value of each entry in the FOCUS validation set

Figure 6.6: Exemplary verification accuracy of daily predictions of the FOCUS model

Compared to the baseline, this model is significantly better than the baseline of $\text{MAE}_{\text{avg}} \approx 6.71$ patients. Neither the MAPE of the test set nor the MAPE_{avg} can be calculated, as this would need a division by zero. The Keras library compensates for this error in an undocumented way, resulting in extremely high MAPEs (>10000%), which seem unreasonable. The validation set does not always contain data points with zero patients. In those cases, the MAPE can be calculated. It results in a $MAPE \approx 39.03\%$, the highest MAPE of all experiments.

The validation results of an exemplary artificial neural network training can be seen in Figure 6.6. Figure 6.6a shows that high patient numbers heavily influence the validation set. It is the same pattern as in the TNOP model. Both models have problems with extreme data entries. Figure 6.6b shows that the model mainly adapted to the patient numbers close to the average, while the outliers are predicted with high errors.



(a) The predicted values and errors of each entry in the FOCUS validation set

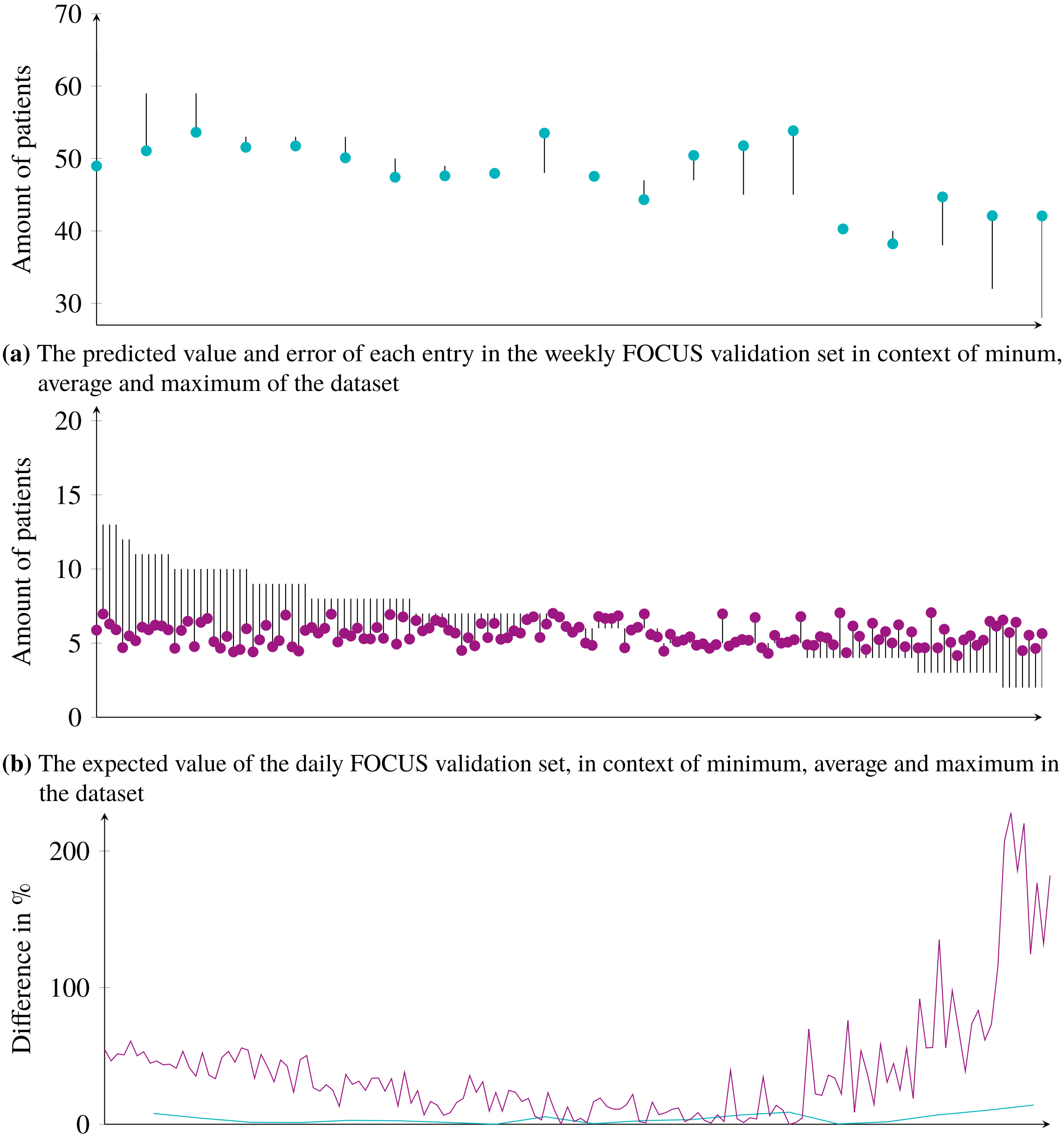
(b) The Difference between the expected value and the predicted value

Figure 6.7: Exemplary verification accuracy of weekly predictions of the FOCUS model

On weekly predictions, the mean absolute error is approximate $MAE \approx 4.51$ patients during training and $MAE \approx 6.55$ patients during the validation. This results in a deviation of $MAPE \approx 9.67\%$ in the training set and $MAPE \approx 14.45\%$ in the validation set. With 27 to 70 ICD-19 patients in the dataset and a mean of 46 patients.

Figure 6.7a shows the prediction ability of the artificial neural network without the need for a trend line. Furthermore, the pattern that the extreme values have a significantly worse prediction quality also stays true in this model. Figure 6.7b is again a plot of the relative deviation (MAPE). Similar to the TNOP model.

To compare the daily and weekly models, we set them again relative to their data ranges in Figure 6.8a and Figure 6.8b. Similar to the TNOP model, the weekly FOCUS model predicts the workload more accurately than the daily FOCUS model. This can be seen especially in Figure 6.8c, which again shows the difference in the two MAPE values.



(a) The predicted value and error of each entry in the weekly FOCUS validation set in context of minum, average and maximum of the dataset

(b) The expected value of the daily FOCUS validation set, in context of minimum, average and maximum in the dataset

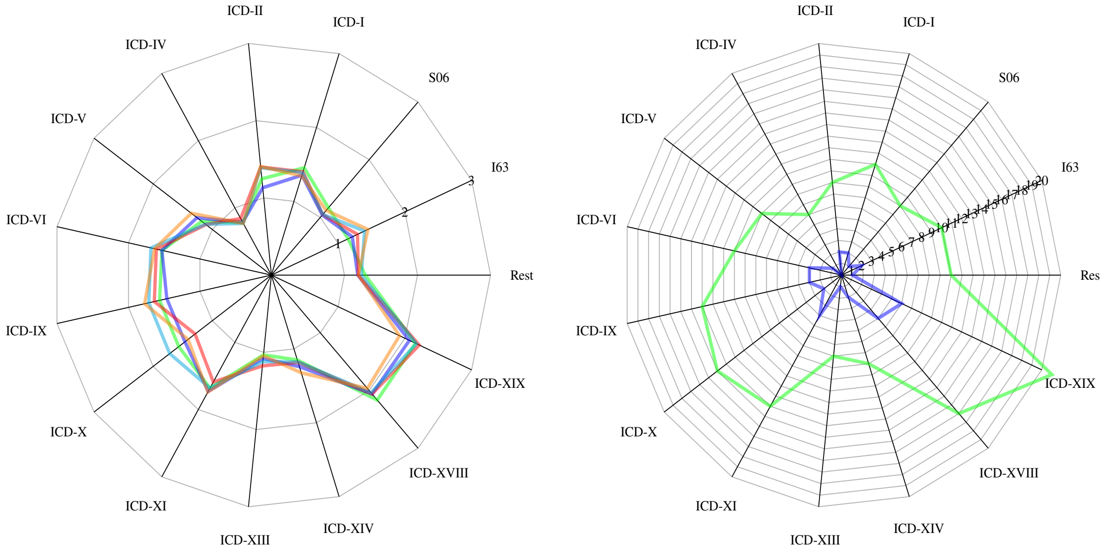
(c) The relative error of the daily and weekly FOCUS validation sets in comparision

Figure 6.8: Comparision of the daily and weekly FOCUS model

6.3.3 Evaluation of the FF Network

After tuning the FF network, the prediction accuracy reaches its top at $MAE \approx 1.49$ patients deviation over all categories on daily predictions. The validation set is a little bit better: It predicts with only an average error of $MAE \approx 1.42$ Patients overall categories. This can be compared to a baseline of $\text{MAE}_{\text{avg}} \approx 3.98$ patients. The absolute individual performance of each category can be seen in Figure 6.9a.

The radar graphs in this chapter show the prediction quality of each prediction category for exemplary models. Each line outwards represents one category indicated by the label outside the circle, while each circumference marks another tick (one patient or one percent). The center indicates zero, and each color represents a single experiment.



(a) daily MAE

(b) mean and maximum amount of patients for each category in the dataset

Figure 6.9: Prediction error in the validation set of the FF model by category (daily prediction)

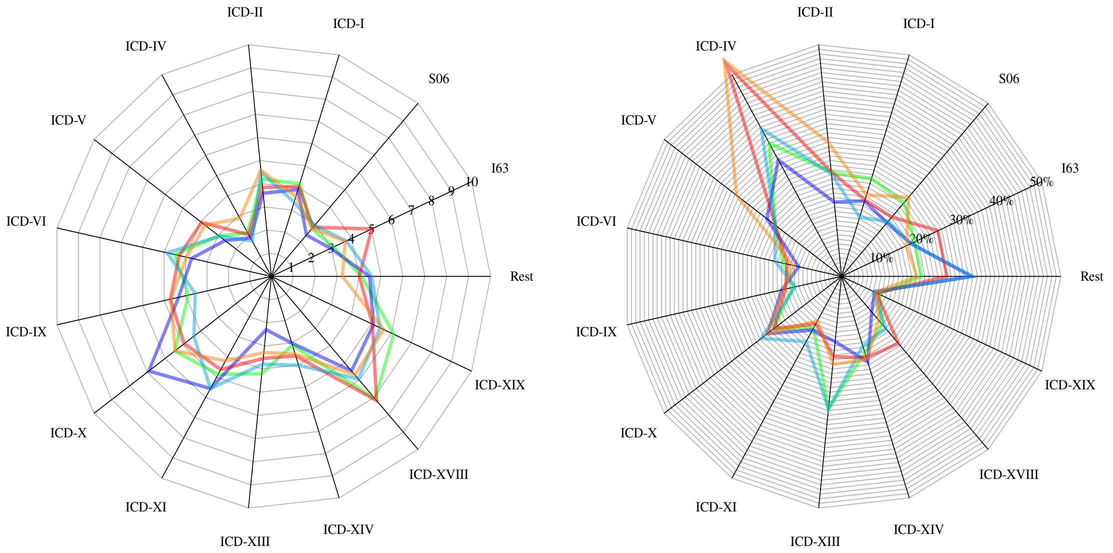
Figure 6.9a shows the daily mean average error (MAE), with the maximum and average value for each category in Figure 6.9b as a comparison. The minimum is not visible as it is zero for all categories. This frequent appearance of zeros in the expected set is why the daily prediction has no relative error, similar to the daily FOCUS model.

All 15 categories can be predicted with an average error of fewer than 2 Patients per day. The most significant error is in the categories ICD-XVIII (up to 2.1 patients) and ICD-XIX (up to 2.2 patients). Those are also the categories with the biggest Search Space.

Interestingly, we can compare the FOCUS model to the prediction of category ICD-XIX in the FF model. Both predictions have an average daily error of ≈ 2.2 patients in training and validation. More minor differences could be caused by the initiation of the model or the chosen sample for validation. This shows that the prediction quality stays similar when trained for a single category or all of them simultaneously.

On weekly predictions, the average absolute error is higher. In training, the expected error is ≈ 4.28 patients. The error in the validation run is ≈ 5.20. The baseline here is a deviation of $\text{MAE}_{\text{avg}} \approx 7.19$ patients and $\text{MAPE}_{\text{avg}} \approx 46.19\%$.

The mean absolute error of each category looks similar to the daily prediction, just scaled up. It can be seen in Figure 6.10a. Next to it is the mean absolute relative error (Figure 6.10b) for the weekly predictions. Except for the category ICD-IV, all predictions stay under 30 %. ICD-IV is extreme, with a mean average error of up to over 50%, making it the only prediction worse than the baseline. This happens because ICD-IV is the smallest category, having only 0 to 14 patients a week, with an average of 5 patients. This is possible another calculation error of Keras, as ICD-IV is the only weekly category that still contains zeros.

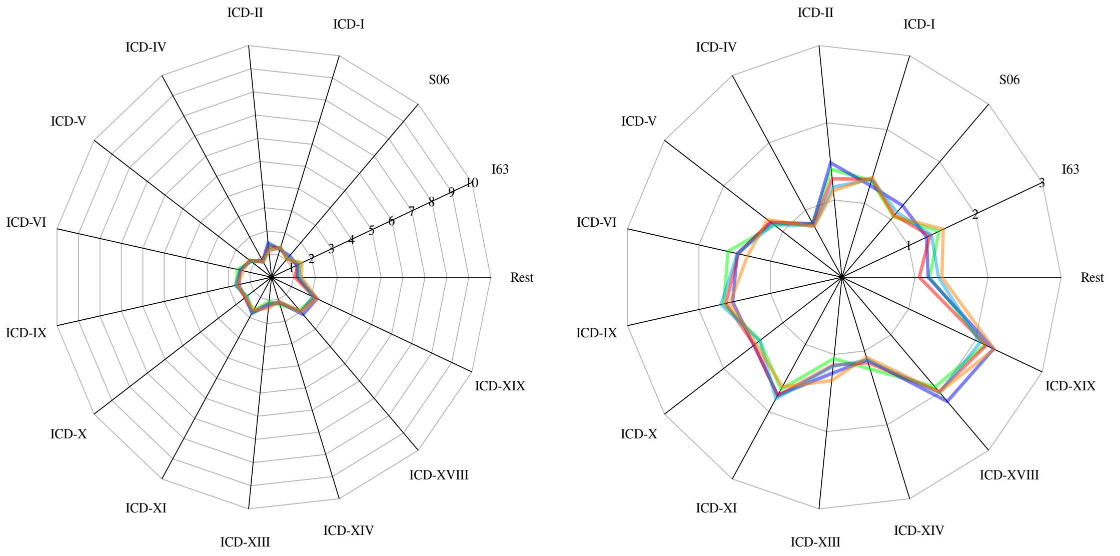


(a) weekly MAE

(b) weekly MAPE (up to 50%)

Figure 6.10: Prediction error in the validation set of the FF model by category (weekly prediction)

6.3.4 Evaluation of the LSTM Network

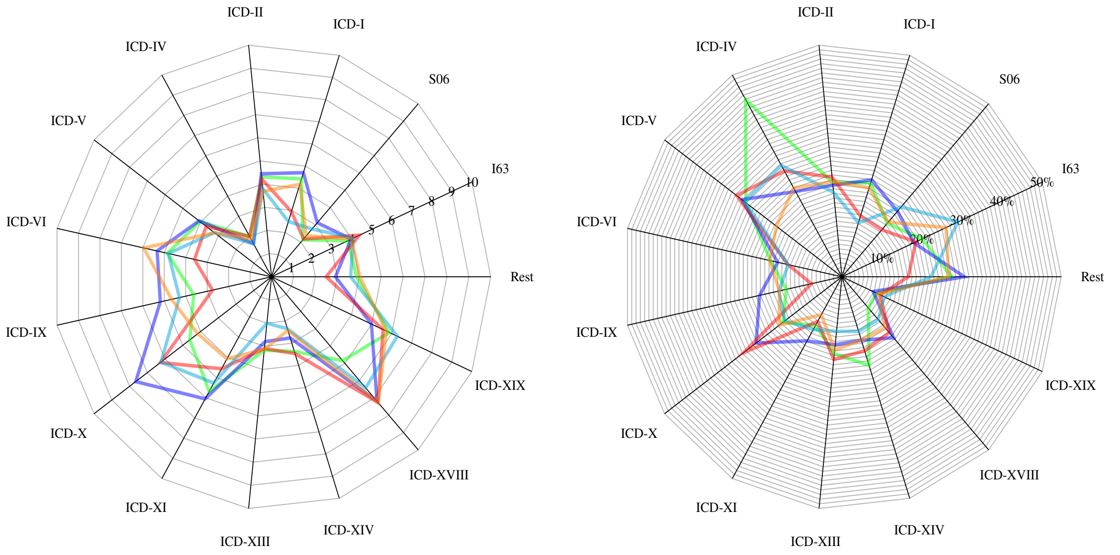


(a) MAE up to an error of 10 patients

(b) MAE up to an error of 3 patients

Figure 6.11: Prediction error in the validation set of the LSTM model by category (daily prediction)

Figure 6.11 shows the mean absolute error for the LSTM network for daily predictions. Figure 6.11a provides a view comparable to the visualization of the daily mean average error of the FF model. In contrast, Figure 6.11b is a zoomed-in version for better readability. Both contain the results of five individually trained graphs tested with an unknown ten % validation split.



(a) weekly MAE

(b) weekly MAPE (up to 50%)

Figure 6.12: Prediction error in the validation set of the LSTM model by category (weekly prediction)

The LSTM network can predict the emergency room workload with a maximum error of ≈ 2.2 over all experiments. Like the daily FF model, the high amount of zeros in the data breaks the MAPE metric, resulting in a vector filled with infinities. As it uses the same data as the FF model, it has the same baseline of $\text{MAE}_{\text{avg}} \approx 3.98$ patients.

The weekly prediction has an average mean absolute error of ≈ 4.01 patients for the training set and ≈ 4.05 patients for the validation set. Being ≈ 26.58% and ≈ 26.01% the respective mean absolute percentage errors. The baseline here is a deviation of $\text{MAE}_{\text{avg}} \approx 7.19$ patients and $\text{MAPE}_{\text{avg}} \approx 46.19\%$.

6.3.5 Overview

Table 6.1 shows an overview of all results, their Search Space, and baseline. *inf* marks fields that we cannot calculate because of zeros, while - marks fields that a single number cannot represent. The minimum, average, and maximum for the FF and LSTM model are 15 different values each, which cannot be represented here.

Overall, the TNOP model performs better than the others. Its vMAPE is far better than the one of the FF and LSTM model and the FOCUS model may be overfitting.

Model	MAE	MAPE	vMAE[3]	vMAPE[3]	min	max	avg	MAE_avg	MAPE_avg
TNOP (daily)	5.8	14.50%	5.33	14.79%	22	66	41	7.80	19.14%
TNOP (weekly)	27.0	8.29%	21	7.45%	145	395	287	40.60	17.95%
FOCUS (daily)	2.23	inf	2.21	39.03%	0	21	6	6.71	inf
FOCUS (weekly)	4.51	9.67%	6.55	14.45%	27	70	46	7.0	15.32%
FF (daily)	1.49	inf	1.42	inf	-	-	-	3.98	inf
FF (weekly)	3.88	inf	4.16	28.00%	-	-	-	7.19	46,19%
LSTM (daily)	2.20	inf	2.1	41,48%	-	-	-	3.98	inf
LSTM (weekly)	4.01	26.58%	4.05	26.01%	-	-	-	7.19	46.19%

Table 6.1: Overview over all Model results

[3] vMAE and vMAPE are the MAE and MAPE of the validation set

6.4 Discussion

6.4.1 TNOP and FOCUS

The first two networks we created for this thesis do not predict the separate ICD categories but a single amount of patients. To be able to compare them, the MAPE is the most interesting metric: The weekly deviation in the training set is quite similar with $MAPE \approx 8.29\%$ and $MAPE \approx 9.67\%$. The verification set paints another image, as the FOCUS network has $MAPE \approx 14.45\%$ twice the deviation than the TNOP model. Considering this difference, the FOCUS network may be partly overfitted.

The daily networks are more difficult to compare, as the zeros in the dataset make the MAPE metric unreliable.

The TNOP model is the broadest model we trained. The prediction provides the least amount of information, but it predicts with the highest precision. While the FOCUS model overall has worse performance than the TNOP model, the performance scales with the data frequency similar to the TNOP model. This shows that all models could be improved by expanding the dataset, especially in extreme cases that are already rare.

6.4.2 FF AND LSTM

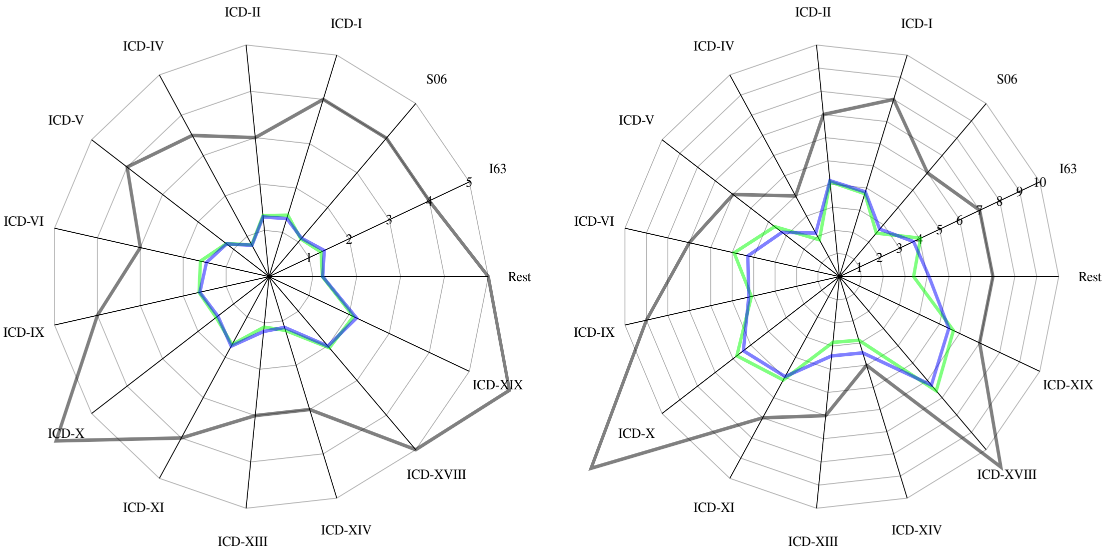
The main target of this thesis is to predict the workload of an emergency room separated by the different categories. The FF model and the LSTM model are the models that fulfill this aim.

The FF model uses a relatively direct approach, creating a mapping from 19 input variables to 15 output categories. This only looks at environmental influences of the same day the network predicts the workload. The LSTM model uses those 19 environmental influences over three days leading up to the predicted day.

Despite this difference in the dataset and the different architecture of both artificial neural networks, the results are nearly identical. Figure 6.13 shows both models in comparison. The green line represents the average of the five FF experiments, and the blue line the average of the five LSTM experiments.

The lines are nearly aligned, and the differences are not dominated by one of the networks. The differences are so minor that they are probably a result of the random nature of machine learning. This similarity is a small validation, as two different models converge to the same results using the same data. The only difference is the number of epochs the networks need to reach this prediction quality. While the FF network needs around 300 epochs, the LSTM network can reach this quality in around 64 Epochs.

The networks can still be compared to the baseline, which is also part of the plot. It is represented by the black line.



(a) Daily Comparison

(b) Weekly Comparison

Figure 6.13: Comparison of the MAE of the LSTM model (blue), the FF model (green) and the baseline (black)

The models rarely have a mean absolute error greater than two patients on daily predictions, which would be more than sufficient for staff planning. However, this average is not the whole picture. The TNOP and FOCUS models show that the predictions stay in an area around the average of the predicted values. While they predict highly precise workloads in this area, a very accurate prediction of an average workload is indistinguishable from a bad prediction of an extreme workload.

All this concludes that the models would profit from expanding the data. A more considerable amount of total data, meaning data from more than four years, would allow training the model more without overfitting to reduce the error for extremes.

6.4.3 More Data and other Aproaches

The artificial neural networks do not predict as accurately as hoped for. As said several times, the prediction quality could increase when the dataset is extended with more data. Especially with more data regarding the, currently, rare workloads. The weekly predictions showed that such changes could increase prediction quality.

Another solution could be to change the machine learning approach. Other algorithms. Artificial neural networks are the technology tested in this thesis. However, they need comparably big datasets to approach a global optimum. Osisanwo et al. showed that Artificial neural networks, together with Decision Trees and JRip algorithms, have the most significant performance loss on small datasets compared to ger datasets [OAA+17].

On the other hand, Osisanwo et al. found that Support Vector machines and random forests had the most negligible quality loss when applied to a smaller dataset. Therefore those two machine learning algorithms may be better suited for a hospital emergency room prediction [OAA+17].

Another idea could be to train a base model with data from multiple hospitals in multiple locations. The resulting model can be used as a base to train a specialized model for a single hospital. This process is called transfer learning and "attempts to improve on traditional machine learning by transferring knowledge learned in one or more source tasks and using it to improve learning in a related target task" [TS10].

6.5 Requirements

In Chapter 3 we presented the requirements of this thesis. This chapter aims to reiterate them to show if our models sufficiently fulfill those requirements.

Most artificial neural networks predict the workload of the hospital emergency room in the Universitätsklinikum Freiburg. The FOCUS model only predicts parts of the workload. This fulfills the first requirement *Workload Prediction*.

The second requirement, *Prediction of Multiple Kinds of Diagnoses*, is only fulfilled by the daily and weekly LSTM and FF model. Still, the thesis overall fulfills the requirement.

All implementations are trained to predict a daily workload and a weekly workload. While the daily timeframe is better suited for staff planning, the weekly timeframe helped identify the shortcomings of the daily one. Therefore, the requirement *Sensible Time Frame* is fulfilled.

The LSTM and FF predictions provide a vector containing 15 different categories and a workload prediction for each category. Those categories are created in cooperation with a medical professional of the Universitätsklinikum Freiburg to ensure their applicability. This fulfills the requirement*Sensible Granularity*.

While the overall prediction quality is better than the baselines, the vast difference in predicting values close to the average and extreme values makes the system unreliable for real-world applications. Therefore, *Quality of Analysis* is at most partly fullfilled.

The requirements *Secret Data* and *In the City of Freiburg* are Specifications, which are fulfill by design. The network was developed using data from the Universitätsklinikum Freiburg without direct access to the data.

7 Conclusion

This thesis evaluates artificial neural networks as prediction tools for hospital emergency room workloads based on environmental influences. Examples of environmental influences are weather, air pollution, and holidays.

Those environmental influences are combined with the dataset from the Universitätsklinikum Freiburg. It provides the day and diagnosis of all stationary emergency room patients from 2016 to 2019. It is reformated to represent the number of patients for each category that this thesis intends to predict, then normalized and cleaned to be used as a training dataset for the artificial neural networks. This process created only a small dataset with 1461 entries for daily predictions and 209 entries for weekly predictions.

We construct four networks, all trained with daily and weekly data. The TNOP network is a simple multi-layer predictor that predicts the sum of all categories and the total number of patients in the emergency room for that timeframe. Likewise, the FOCUS network is also a multi-layer perceptron. Rather than all patients, it predicts the workload of a single perceptron category. The third multi-layer perceptron is the FF network. While similar in architecture to the others, it has a more complex output. It provides a vector with the predicted workload of all prediction categories instead of a single value. It is the same output as the LSTM model, which has its own architecture and uses multiple days of data for the prediction.

All models predict significantly better than their baseline, with the TNOP network having the most accurate predictions. However, the predictions based on the weekly datasets are better than those found daily. Overall, the trend is that if each dataset entry contains more cases, the prediction quality increases. The FOCUS network performs similarly to the same prediction category in the FF and LSTM network. FF and LSTM are near identical in prediction quality.

However, while the average predictions are excellent, the models have difficulties predicting extreme big or small workloads. The predictions are kept close to the average, making it unreliable as a real-life application.

Still, it shows that artificial neural networks and other machine learning approaches generally can predict such a chaotic and complicated system as the demand for emergency services, even if the data used for training needs to be expanded significantly before a real-life application is feasible.

7.1 Limitations

7.1.1 Pre COVID-19 Data

To make the model training easier, data, after corona started was explicitly ignored. This date could have decreased the quality of the model, as the pandemic was a significant time that probably has changed people's behavior towards the emergency room and other areas of their life. Such a change could result in new patterns. Different patterns for the same data in one dataset may reduce the quality of the trained model.

This does not thread the validity of this thesis itself but makes the resulting model inapplicable for predictions today. It can happen to any social correlation over time, but COVID-19 was a catalyst for an extreme change in this regard.

To solve such problems, a model that is actively used in hospitals should be updated regularly.

7.1.2 Stationary only

The provided hospital data has diagnosis data only for patients threaded at the hospital stationary. First, that introduces a bias, which removes all illnesses and injuries that people go to the hospital for but are treated on the spot. Examples here are broken bones, light burns, or minor wounds.

Ambulant cases are more critical for staff planning, as they cover more than 60% of the cases in Freiburg from 2016 to 2019. Nevertheless, they could not be used for the training process as diagnosis data was not provided for those cases. This further decreases the size of the dataset and reduces the value of the final model.

To fix this problem, the data for training the networks should be from a hospital that tracks diagnosis data from ambulant and stationary patients alike.

7.1.3 Small Data set

As already stated, the data originates from a single hospital. Additionally, it is only data from a period of 4 years. This creates a relatively small dataset with only 1461 entries. The model may find a pattern that does not exist or is over-fitted to this small amount of data. The relatively high amount of features only worsens this. Twenty features is not a lot compared to other deep neural networks, but those have millions of data points, while we use less than 2000.

This could be solved similarly to **??**. More data from multiple hospitals would be a solution. It would increase the amounts of data points significantly and, therefore, the quality of the model. Nevertheless, the same restrictions that **??** mentioned apply here.

7.2 Future Work

This thesis showed that artificial neural networks could predict hospital emergency room workload. Furthermore, while the prediction is limited, it could be improved using a few different approaches.

First, the machine learning algorithm could be changed to be more suitable for small datasets. Some ideas here are Random Forests or Support Vector Machines. Those may have fewer problems with extreme values.

This thesis's data to train the neural networks originates from a single city. Like the weather and hospital data, most of it is from a single source. This creates a specific dataset that is only applicable to the same city. The results can not safely be generalized and applied to other cities, especially those with different politics, cultures, and climates. All those factors can influence the decision to go to the emergency room, while the data the networks are trained on stays the same. However, the network could be generally trained with data from multiple cities and hospitals and then specialized to a single area. This process is called transfer learning.

Also, it would be interesting to evaluate how effective the predictions can influence the staff planning. One could conduct a study in which the staff planning process of multiple hospitals is evaluated and improved using a prediction model with a given precision. It would be interesting to know if there is a significant impact.

The current implementation could be expanded to yield improved results. One could experiment with more artificial neural network architectures or identify more effective input variables. A feature analysis of the input variables currently used could also provide interesting insights. It would also be interesting to increase the impact of the timelag, testing different timelag lengths or different inputs for each day in the timelag, like using workloads during the timelag period as additional input.

Bibliography

Sources

[BLB+18] E. J. Brandl, T. A. Lett, G. Bakanidze, A. Heinz, F. Bermpohl, M. Schouler-Ocak. “Weather conditions influence the number of psychiatric emergency room patients”. In: *International Journal of Biometeorology* 62 (5 May 2018), pp. 843–850. DOI: 10.1007/s00484-017-1485-z (cit. on p. 5).

[Bre96] L. Breiman. “Bagging predictors”. In: *Machine Learning* 24.2 (Aug. 1996), pp. 123–140. DOI: 10.1007/bf00058655 (cit. on p. 20).

[BRG94] H. B. Burke, D. B. Rosen, P. H. Goodman. “Comparing artificial neural networks to other statistical methods for medical outcome prediction”. In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN’94)*. Vol. 4. IEEE, 1994, pp. 2213–2216. ISBN: 0-7803-1901-X. DOI: 10.1109/ICNN.1994.374560. URL: http://ieeexplore.ieee.org/document/374560/ (cit. on p. 7).

[Cha17] S. Chatterjee. *Good Data and Machine Learning*. Aug. 24, 2017. URL: https://towardsdatascience.com/data-correlation-can-make-or-break-your-machine-learning-project-82ee11039cc9 (cit. on p. 24).

[CMM83] J. G. Carbonell, R. S. Michalski, T. M. Mitchell. *AN OVERVIEW OF MACHINE LEARNING*. Elsevier, Jan. 1983, pp. 3–23. DOI: 10.1016/B978-0-08-051054-5.50005-4. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780080510545500054 (cit. on p. 6).

[CSSA95] J. Castellsague, J. Sunyer, M. Saez, J. M. Anto. “Short-term association between air pollution and emergency room visits for asthma in Barcelona.” In: *Thorax* 50 (10 Oct. 1995), pp. 1051–1056. ISSN: 0040-6376. DOI: 10.1136/thx.50.10.1051. URL: https://thorax.bmj.com/lookup/doi/10.1136/thx.50.10.1051 (cit. on pp. 2, 5, 27).

[Dev22] N. Developers. *numpy.ndarray — NumPy v1.22 Manual*. 2022. URL: https://numpy.org/doc/stable/reference/generated/numpy.ndarray.html (cit. on p. 40).

[Fin99] T. L. Fine. *Feedforward Neural Network Methodology*. 1999. ISBN: 0-387-98745-2 (cit. on p. 8).

[FS96] Y. Freund, R. E. Schapire. “Experiments with a new boosting algorithm”. In: *Proceedings of the Thirteenth International Conference on Machine Learning*. Bari, July 3, 1996 (cit. on p. 20).

[GKG+98] B. Z. Garty, E. Kosman, E. Ganor, V. Berger, L. Garty, T. Wietzen, Y. Waisman, M. Mimouni, Y. Waisel. “Emergency Room Visits of Asthmatic Children, Relation to Air Pollution, Weather, and Airborne Allergens”. In: *Annals of Allergy, Asthma & Immunology* 81 (6 Dec. 1998), pp. 563–570. ISSN: 1081-1206. DOI: 10.1016/S1081-1206(10)62707-X (cit. on p. 27).

[GWS20] J. Gao, H. Wang, H. Shen. “Machine Learning Based Workload Prediction in Cloud Computing”. In: *2020 29th International Conference on Computer Communications and Networks (ICCCN)*. IEEE, Aug. 2020. DOI: 10.1109/icccn49398.2020.9209730 (cit. on pp. 19, 21, 22).

[HDR19] S. Hayou, A. Doucet, J. Rousseau. “On the impact of the activation function on deep neural networks training”. In: *International conference on machine learning*. PMLR. 2019, pp. 2672–2680 (cit. on p. 40).

[HLY+08] J. I. Halonen, T. Lanki, T. Yli-Tuomi, M. Kulmala, P. Tiittanen, J. Pekkanen. “Urban air pollution, and asthma and COPD hospital emergency room visits”. In: *Thorax* 63 (7 July 2008), pp. 635–641. ISSN: 0040-6376. DOI: 10.1136/thx.2007.091371. URL: https://thorax.bmj.com/lookup/doi/10.1136/thx.2007.091371 (cit. on pp. 2, 5, 27, 34, 35).

[Ho95] T. K. Ho. “Random decision forests”. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. IEEE Comput. Soc. Press, 1995. DOI: 10.1109/icdar.1995.598994 (cit. on p. 20).

[Joa20] R. S. Joachim Steinwendner. *Neuronale Netze programmieren mit Python*. Rheinwerk Verlag GmbH, May 28, 2020. 479 pp. ISBN: 3836274507. URL: https://www.ebook.de/de/product/38548759/joachim_steinwendner_roland_schwaiger_neuronale_netze_programmieren_mit_python.html (cit. on p. 8).

[KASJ11] S. M. R. Kazemi-Bajestani, A. Amirsadri, S. A. A. Samari, A. Javanbakht. “Lunar phase cycle and psychiatric hospital emergency visits, inpatient admissions and aggressive behavior”. In: *Asian Journal of Psychiatry* 4 (1 Mar. 2011), pp. 45–50. DOI: 10.1016/j.ajp.2010.12.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S1876201810001668 (cit. on p. 5).

[KBK11] A. Krenker, J. Bester, A. Kos. “Introduction to the Artificial Neural Networks”. In: *Artificial Neural Networks - Methodological Advances and Biomedical Applications*. InTech, Apr. 2011. DOI: 10.5772/15751 (cit. on p. 9).

[Kla21] C. R. Klaus Pohl. *Basiswissen Requirements Engineering*. Dpunkt.Verlag GmbH, Apr. 1, 2021. ISBN: 3864908140. URL: https://www.ebook.de/de/product/39842397/klaus_pohl_chris_rupp_basiswissen_requirements_engineering.html (cit. on p. 11).

[Kri21] K. Krieglstein. *Satzung der Ethikkommission des Universitätsklinikum Freiburgs*. Oct. 2021. URL: https://www.uniklinik-freiburg.de/fileadmin/mediapool/10_andere/ethikkommission/pdf/21-10-29-satzung.pdf (cit. on p. 12).

[Kro] B. Krollner. *10-GM-2022 ICD-10-GM-2022 - ICD10*. URL: https://www.icd-code.de/icd/code/ICD-10-GM.html (cit. on pp. 6, 13, 31, 32, 35).

[Kub17] M. Kubat. *An Introduction to Machine Learning*. Springer International Publishing, Sept. 2017, pp. 1–348. ISBN: 978-3-319-63912-3. DOI: 10.1007/978-3-319-63913-0. URL: http://link.springer.com/10.1007/978-3-319-63913-0 (cit. on pp. 6, 40, 41).

[LBX+21] J. Lu, P. Bu, X. Xia, N. Lu, L. Yao, H. Jiang. “Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases”. In: *Environmental Science and Pollution Research* 28 (23 June 2021), pp. 29701–29709. ISSN: 0944-1344. DOI: 10.1007/s11356-021-12658-7. URL: https://link.springer.com/10.1007/s11356-021-12658-7 (cit. on pp. 16, 17, 21, 22, 34, 43, 45).

[LHO97] M. Lipsett, S. Hurley, B. Ostro. “Air pollution and emergency room visits for asthma in Santa Clara County, California.” In: *Environmental Health Perspectives* 105 (2 Feb. 1997), pp. 216–222. ISSN: 0091-6765. DOI: 10.1289/ehp.97105216. URL: https://ehp.niehs.nih.gov/doi/10.1289/ehp.97105216 (cit. on pp. 5, 27).

[LPS+19] L. M. T. Luong, D. Phung, P. D. Sly, T. N. Dang, L. Morawska, P. K. Thai. “Effects of temperature on hospitalisation among pre-school children in Hanoi, Vietnam”. In: *Environmental Science and Pollution Research* 26 (3 Jan. 2019), pp. 2603–2612. DOI: 10.1007/s11356-018-3737-9 (cit. on pp. 2, 5).

[MDH06] R. N. McLay, A. A. Daylo, P. S. Hammer. “No Effect of Lunar Cycle on Psychiatric Admissions or Emergency Evaluations”. In: *Military Medicine* 171 (12 Dec. 2006), pp. 1239–1242. ISSN: 0026-4075. DOI: 10.7205/MILMED.171.12.1239. URL: https://academic.oup.com/milmed/article/171/12/1239-1242/4578201 (cit. on p. 5).

[Nas17] V. Nasteski. “An overview of the supervised machine learning methods”. In: *HORIZONS* 4 (Dec. 2017), pp. 51–62. DOI: 10.20544/HORIZONS.B.04.1.17.P05. URL: http://uklo.edu.mk/filemanager/HORIZONTI%202017/Serija%20B%20br.%204/6.An%20overview%20of%20the%20supervised.pdf (cit. on pp. 6, 7).

[NLBW71] J. H. Noble, M. E. Lamontagne, C. Bellotti, H. Wechsler. “Variations in Visits to Hospital Emergency Care Facilities: Ritualistic and Meteorological”. In: *Care*. Vol. 9. 1971, pp. 415–427. URL: https://www.jstor.org/stable/3762514 (cit. on pp. 2, 5).

[NW72] J. A. Nelder, R. W. M. Wedderburn. “Generalized Linear Models”. In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), p. 370. DOI: 10.2307/2344614 (cit. on p. 20).

[OAA+17] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, J. Akinjobi. “Supervised machine learning algorithms: classification and comparison”. In: *International Journal of Computer Trends and Technology (IJCTT)* 48.3 (2017), pp. 128–138 (cit. on p. 58).

[OBL+19] T. O’Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, et al. *KerasTuner*. 2019. URL: https://github.com/keras-team/keras-tuner (cit. on p. 41).

[OK83] G. M. Oderda, W. Klein-Schwartz. “Lunar Cycle and Poison Center Calls”. In: *Journal of Toxicology: Clinical Toxicology* 20 (5 Jan. 1983), pp. 487–495. ISSN: 0731-3810. DOI: 10.3109/15563658308990614. URL: http://www.tandfonline.com/doi/full/10.3109/15563658308990614 (cit. on p. 5).

[Org] W. H. Organisation. *International Classification of Diseases (ICD)*. URL: https://www.who.int/classifications/classification-of-diseases (cit. on pp. 6, 13).

[Par10] H. A. Partsch. *Requirements-Engineering systematisch*. Springer Berlin Heidelberg, 2010. DOI: 10.1007/978-3-642-05358-0 (cit. on p. 11).

[PCZ+20] J. Peng, C. Chen, M. Zhou, X. Xie, Y. Zhou, C.-H. Luo. “Peak Outpatient and Emergency Department Visit Forecasting for Patients With Chronic Respiratory Diseases Using Machine Learning Methods: Retrospective Cohort Study”. In: *JMIR Medical Informatics* 8.3 (Mar. 2020), e13075. DOI: 10.2196/13075 (cit. on pp. 20–22, 34, 35).

[RBH02] M. Rusticucci, M. L. Bettolli, M. D. L. A. Harris. “Association between weather conditions and the number of patients at the emergency room in an Argentine hospital”. In: *International Journal of Biometeorology* 46 (1 Feb. 2002), pp. 42–51. DOI: 10.1007/s00484-001-0113-z (cit. on pp. 5, 17, 21, 22).

[Rud16] S. Ruder. *An overview of gradient descent optimization algorithms*. 2016. DOI: 10.48550/ARXIV.1609.04747 (cit. on p. 40).

[SAL+17] J. D. Sonis, E. L. Aaronson, R. Y. Lee, L. L. Philpotts, B. A. White. “Emergency Department Patient Experience”. In: *Journal of Patient Experience* 5.2 (Sept. 2017), pp. 101–106. DOI: 10.1177/2374373517731359 (cit. on p. 11).

[SB98] R. S. Sutton, A. G. Barto. *Reinforcement learning : an introduction*. MIT Press, 1998, p. 322. ISBN: 9780262193986 (cit. on p. 7).

[SCK+19] S. Sohn, W. Cho, J. A. Kim, A. Altaluoni, K. Hong, B. C. Chun. “’Pneumonia weather’: Short-term Effects of Meteorological Factors on Emergency Room Visits Due to Pneumonia in Seoul, Korea”. In: *Journal of Preventive Medicine and Public Health* 52 (2 Mar. 2019), pp. 82–91. DOI: 10.3961/jpmph.18.232 (cit. on p. 5).

[Scu11] C. Scuffy. “Belief in lunar effects”. PhD thesis. University of Minnesota, 2011. URL: https://conservancy.umn.edu/handle/11299/187487 (cit. on p. 28).

[Seb21] V. M. Sebastian Raschka. *Machine Learning mit Python und Keras, TensorFlow 2 und Scikit-learn*. MITP Verlags GmbH, Mar. 12, 2021. 768 pp. ISBN: 374750213X. URL: https://www.ebook.de/de/product/40283639/sebastian_raschka_vahid_mirjalili_machine_learning_mit_python_und_keras_tensorflow_2_und_scikit_learn.html (cit. on pp. 40, 41).

[Sim19] M. Simon. “Personalbesetzung - Dichtung und Wahrheit”. In: *Pflegezeitschrift* 72.3 (Feb. 2019), pp. 17–19. DOI: 10.1007/s41906-019-0006-6 (cit. on p. 1).

[SSB+81] J. M. Samet, F. E. Speizer, Y. Bishop, J. D. Spengler, B. G. Ferris. “The Relationship between Air Pollution and Emergency Room Visits in an Industrial Community”. In: *Journal of the Air Pollution Control Association* 31 (3 Mar. 1981). ISSN: 0002-2470. DOI: 10.1080/00022470.1981.10465214 (cit. on p. 27).

[SSL+93] J. Schwartz, D. Slater, T. V. Larson, W. E. Pierson, J. Q. Koenig. “Particulate Air Pollution and Hospital Emergency Room Visits for Asthma in Seattle”. In: *American Review of Respiratory Disease* 147 (4 Apr. 1993), pp. 826–831. ISSN: 0003-0805. DOI: 10.1164/ajrccm/147.4.826. URL: http://www.atsjournals.org/doi/abs/10.1164/ajrccm/147.4.826 (cit. on pp. 5, 27).

[ST05] P. H. Sydenham, R. Thorn. *Handbook of measuring system design*. Wiley, 2005, pp. 901–908. ISBN: 0470021438 (cit. on p. 9).

[Swa00] “MAPE (mean absolute percentage error)MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)”. In: *Encyclopedia of Production and Manufacturing Management*. Ed. by P. M. Swamidass. Boston, MA: Springer US, 2000, pp. 462–462. ISBN: 978-1-4020-0612-8. DOI: 10.1007/1-4020-0612-8_580. URL: https://doi.org/10.1007/1-4020-0612-8_580 (cit. on p. 43).

[TBC+21] Y. d. Tadano, E. T. Bacalhau, L. Casacio, E. Puchta, T. S. Pereira, T. Antonini Alves, C. M. Ugaya, H. V. Siqueira. "Unorganized Machines to estimate the number of hospital admissions due to respiratory diseases caused by PM10 concentration". In: *Atmosphere* 12.10 (2021), p. 1345. DOI: 10.3390/atmos12101345 (cit. on pp. 18, 19, 21, 22).

[TBST21] E. Tuba, N. Bačanin, I. Strumberger, M. Tuba. "Convolutional Neural Networks Hyperparameters Tuning". In: *Artificial Intelligence: Theory and Applications*. Springer, 2021, pp. 65–84 (cit. on p. 40).

[TS10] L. Torrey, J. Shavlik. "Transfer Learning". In: *Handbook of Research on Machine Learning Applications and Trends*. IGI Global, 2010, pp. 242–264. DOI: 10.4018/978-1-60566-766-9.ch011 (cit. on p. 59).

[Vap98] Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Sept. 16, 1998. 762 pp. ISBN: 0471030031. URL: https://www.ebook.de/de/product/3602628/vapnik_statistical_learning_theory.html (cit. on p. 20).

[WH00] R. Wirth, J. Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1. Manchester. 2000, pp. 29–40 (cit. on pp. 23, 24).

[YSHZ19] Y. Yu, X. Si, C. Hu, J. Zhang. "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures". In: *Neural Computation* 31.7 (July 2019), pp. 1235–1270. DOI: 10.1162/neco_a_01199 (cit. on p. 9).

[Zan06] A. Zanobetti. "Air pollution and emergency admissions in Boston, MA". In: *Journal of Epidemiology & Community Health* 60 (10 Oct. 2006), pp. 890–895. ISSN: 0143-005X. DOI: 10.1136/jech.2005.039834. URL: https://jech.bmj.com/lookup/doi/10.1136/jech.2005.039834 (cit. on pp. 5, 27).

[ZG09] X. Zhu, A. B. Goldberg. "Introduction to Semi-Supervised Learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3 (1 Jan. 2009), pp. 1–130. ISSN: 1939-4608. DOI: 10.2200/S00196ED1V01Y200906AIM006. URL: http://www.morganclaypool.com/doi/abs/10.2200/S00196ED1V01Y200906AIM006 (cit. on p. 7).

Legal Sources

[ArbZG] B. Deutschland. *Arbeitszeitgesetz (arbzg)*. June 1994. URL: https://www.gesetze-im-internet.de/arbzg/BJNR117100994.html (cit. on p. 12).

[DSGVO] E. Union. *VERORDNUNG (EU) 2016/679*. Apr. 2016. URL: https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX%3A32016R0679 (cit. on pp. 14, 30).

[MBO-Ä] Bundesärztekammer. *(muster-)berufsordnung für die in Deutschland ... - bundesaerztekammer.de*. Dec. 2018. URL: https://www.bundesaerztekammer.de/fileadmin/user_upload/downloads/pdf-Ordner/MBO/MBO-AE.pdf (cit. on p. 14).

[PDSG] B. Deutschland. *Gesetz zum Schutz Elektronischer Patienten ...* Oct. 2020. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/P/PDSG_bgbl.pdf (cit. on pp. 12, 14).

All links were last followed on June 08, 2022.

A ICD-10 Top Level Categories

- **ICD- I (A00-B99)** Certain infectious and parasitic diseases
- **ICD- II (C00-D48)** New formations
- **ICD- III (D50-D90)** Diseases of the blood and blood-forming organs and certain disorders involving the immune system
- **ICD- IV (E00-E90)** Endocrine, nutritional and metabolic diseases
- **ICD- V (F00-F99)** Mental and behavioral disorders
- **ICD- VI (G00-G99)** Nervous system diseases
- **ICD- VII (H00-H59)** Eye-related diseases
- **ICD- VIII (H60-H95)** Ear-related diseases
- **ICD- IX (I00-I99)** Diseases of the circulatory system
- **ICD- X (J00-J99)** Respiratory system diseases
- **ICD- XI (K00-K93)** Diseases of the digestive system
- **ICD- XII (L00-L99)** Diseases of the skin and subcutaneous tissue
- **ICD- XIII (M00-M99)** Diseases of the musculoskeletal system and connective tissue
- **ICD- XIV (N00-N99)** Diseases of the genitourinary system
- **ICD- XV (O00-O99)** Pregnancy, birth and postpartum
- **ICD- XVI (P00-P96)** Certain conditions that originate in the perinatal period
- **ICD- XVII (Q00-Q99)** Congenital malformations, deformities and chromosomal anomalies
- **ICD- XVIII (R00-R99)** Symptoms and abnormal clinical and laboratory findings not classified elsewhere.
- **ICD- XIX (S00-T98)** injuries, poisoning and certain other consequences of external causes
- **ICD- XX (V01-Y84)** External causes of morbidity and mortality
- **ICD- XXI (Z00-Z99)** Factors influencing health status and leading to health care utilization.
- **ICD- XXII (U00-U99)** Key numbers for special purposes

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

place, date, signature