GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Analysis of Machine Learning Prediction Quality for Automated Subgroups within the MIMIC III Dataset

by

Jakob Vanek

A Master's Thesis
Submitted to the Institute of Computer Science
Goethe-Universität Frankfurt am Main
In Partial Fulfillment of the Requirements
For a Master's Degree
July 2023

Supervisor: Prof. Dr. Lena Wiese

GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Master Thesis

Analysis of Machine Learning Prediction Quality for Automated Subgroups within the MIMIC III Dataset

Author: Jakob Vanek
Matr.No.: 5879812
s0266343@stud.uni-frankfurt.de
Wirtschaftsinformatik
Supervisor: Prof. Dr. Lena Wiese
Institute of Computer Science
Goethe-Universität Frankfurt a. M.

Date: July 17, 2023

Bitte dieses Formular zusammen mit der Abschlussarbeit abgeben!

Erklärung zur Abschlussarbeit

gemäß § 34, Abs. 16 der Ordnung für den Masterstudiengang Wirtschaftsinformatik vom 01. April 2019

Hiermit erkläre ich

____

(Nachname, Vorname)

Die vorliegende Arbeit habe ich selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst.

Ebenso bestätige ich, dass diese Arbeit nicht, auch nicht auszugsweise, für eine andere Prüfung oder Studienleistung verwendet wurde.

Zudem versichere ich, dass die von mir eingereichten schriftlichen gebundenen Versionen meiner Masterarbeit mit der eingereichten elektronischen Version meiner Masterarbeit übereinstimmen.

Frankfurt am Main, den

____

Unterschrift der/des Studierenden

Abstract

The motivation for this master's thesis is to explore the potential of predictive data analytics in the field of medicine. For this, the MIMIC-III dataset offers an extensive foundation for the construction of prediction models, including Random Forest, XGBOOST, and deep learning networks. These models were implemented to forecast the mortality of 2,655 stroke patients.

The first part of the thesis involved conducting a comprehensive data analysis of the filtered MIMIC-III dataset.

Subsequently, the effectiveness and fairness of the predictive models were evaluated. Although the performance levels of the developed models did not match those reported in related research, their potential became evident. The results obtained demonstrated promising capabilities and highlighted the effectiveness of the applied methodologies. Moreover, the feature relevance within the XGBOOST model was examined to increase model explainability.

Finally, relevant subgroups were identified to perform a comparative analysis of the prediction performance across these subgroups. While this approach can be regarded as a valuable methodology, it was not possible to investigate underlying reasons for potential unfairness across clusters. Inside the test data, not enough instances remained per subgroup for further fairness or feature relevance analysis. In conclusion, the implementation of an alternative use case with a higher patient count is recommended.

The code for this analysis is made available via a GitHub repository and includes a frontend to visualize the results.

Abstrakt

Das Ziel dieser Masterarbeit ist es, das Potenzial prädiktiver Datenanalyse im Bereich der Medizin zu erforschen. Hierzu bietet MIMIC-III Datensatz eine umfangreiche Grundlage für die Erstellung von Vorhersagemodellen, darunter Random Forest, XGBOOST und Deep-Learning-Netzwerke. Diese Modelle wurden eingesetzt, um die Sterblichkeit von 2.655 Schlaganfallpatienten vorherzusagen.

Im ersten Teil der Arbeit wurde eine umfassende Datenanalyse des gefilterten MIMIC-III Datensatz durchgeführt.

Anschließend wurden die Effektivität und Fairness der entwickelten Vorhersagemodelle bewertet. Obwohl das Leistungsniveau der entwickelten Modelle nicht an die in verwandten Forschungsarbeiten berichteten Werte heranreichte, wurde ihr Potenzial dennoch deutlich. Die erzielten Ergebnisse zeigten vielversprechende Anwendungsmöglichkeiten auf. Zudem wurde der Einfluss der Features auf das XGBOOST Modell untersucht, um die Erklärbarkeit des Vorhersagemodells zu erhöhen.

Schließlich wurden relevante Untergruppen identifiziert, um eine vergleichende Analyse der Vorhersageleistung zwischen diesen Untergruppen durchzuführen. Dieser Ansatz kann zwar als nützliche Methode angesehen werden, doch war es nicht möglich, die Gründe für mögliche Unfairness zwischen den Clustern zu untersuchen. Innerhalb der Testdaten blieben nicht genügend Instanzen pro Untergruppe für eine weitere Fairness- oder Merkmalsrelevanzanalyse übrig. Abschließend wird die Implementierung eines alternativen Anwendungsfalls mit einer höheren Patientenzahl empfohlen.

Der Code für diese Analyse wird über ein GitHub-Repository zur Verfügung gestellt und enthält ein Frontend zur Visualisierung der Ergebnisse.

Contents

1 Introduction

1.1 The Potential of Prediction Models in Medicine

Predictive data analytics are going to have a tremendous impact on the field of medicine. Especially recent advancements in computer science, like the development of Machine Learning (ML) models based on electronic health records, can immensely support healthcare professionals in their daily work.

Novel tools in healthcare have the potential for various applications, including healthcare research, early diagnosis, and real-time monitoring of risks [1]. In the context of this thesis, the specific focus lies on mortality prediction, as advancements in this field can have a direct impact on saving lives. By accurately predicting mortality risks, healthcare providers can make informed decisions and optimize healthcare services to improve patient outcomes. This has the potential to bring about significant changes in medical treatment and ultimately enhance overall healthcare delivery.

However, the development of these predictive tools must be accompanied by careful examination and reflection. Models like deep learning networks, which are often referred to as Artificial Intelligence (AI), make it increasingly difficult to retrace the calculations on which they were based. For such "black-box" models, the decision process from input to output is often not retractable [1]. These unclear decision processes introduce potential risks that are specific to this technology. Especially, unclear accountability, a lack of transparency, potential unfairness, and inadequate data protection are some of the main concerns regarding predictive AI tools [2].

While the benefits of predictive models, such as improved patient management, are widely acknowledged, the risks associated with their deployment emphasize the need for responsible implementation. As these aspects are especially relevant in the healthcare sector, some researchers advocate for establishing "uniform international standards, not at least from a medical ethics perspective" [3].

The societal impact of predictive AI models is currently the subject of intense public discussion and the relevance of these concerns is underscored by the introduction of new regulations, such as the European AI Act[1]. Herein, the need for careful oversight in the field of artificial intelligence is recognized. Specifically, AI systems deployed in the context of medical assistance are classified as high-risk. The designation further highlights the importance of ensuring the safety, effectiveness, and fairness of AI technologies used in healthcare settings.

In line with this, the World Health Organization has published guidelines addressing the use of digital tracking devices. They emphasize the necessity of safeguarding patients' rights while also ensuring acceptable working conditions for healthcare workers [4].

A final important aspect to consider is the issue of equal treatment, which is a fundamental

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206

principle of the medical system. There is a concern that advancements in computational profiling techniques may inadvertently reinforce existing biases and prejudices. This could further "deepen existing inequalities and reinforce already entrenched practices of discrimination" [3]. This underscores the crucial requirement for the medical field to not only pursue performance advancements in the development of digital solutions but also prioritize the explainability and fairness of such approaches.

Should predictive algorithms prove to be unfair, the confidence and trust in these emerging tools would experience a significant decline. These concerns account for the increased importance of an enhanced perspective on data analytics. To this end, the inclusion of fairness metrics and feature relevance can improve the explainability of such models.

1.2 Research Question

The practical focus of this thesis is a use case of mortality prediction of stroke patients in the Intensive Care Unit (ICU). The data originates from the publicly available Medical Information Mart for Intensive Care dataset (MIMIC-III, v1.4).

The primary objective of this thesis is to evaluate and compare various prediction models in terms of their performance and fairness. The aim is to determine whether accurate mortality prediction is achievable for stroke patients. Moreover, different subgroups shall be detected for which the predictive quality might differ. Clustering methods and fairness metrics may be supportive of predictive approaches by making such relevant subgroups visible.

The overarching research question to be addressed is: Does the MIMIC-III dataset provide adequate suitability for investigating these research objectives, and can potential subgroups be effectively identified within the dataset?

Given that this is not a medical thesis, the emphasis is placed on the comparison of prediction models rather than drawing specific medical conclusions. Consequently, this thesis aims to primarily benefit the field of machine learning and, in turn, contribute to research in the healthcare sector.

1.3 Thesis Outline

In the following Chapter 2 the related research and expected results are introduced. Next, Chapter 3 explains the technical pre-processing steps that were realized for the setup of the use case. Chapter 4 summarizes a descriptive data analysis of the dataset. In addition, the feature correlations are presented. This is followed by a clustering analysis in Chapter 5. In Chapter 6 the results for different prediction models are discussed. The chapter is concluded with an analysis of feature importance. Next, Chapter 7 introduces a fairness analysis based on three examples. The final analysis step is implemented in Chapter 8, where a methodology to investigate subgroups is proposed. In conclusion, Chapter 9 lists potential steps for future research and summarizes the results of this thesis.

2 Related Work

This section provides a summary of previous research that was influential for this thesis. Additionally, the expected results of this thesis are outlined.

2.1 Influential Research

Previous research on the MIMIC-III dataset can be broadly categorized in three directions.

First, there is research based on the hourly prediction of the occurrence of sickness, death, or complications like sepsis. Therein, the goal is to achieve earlier predictions compared to scoring methods that are currently implemented in hospitals. In this regard, hourly time-series data is commonly employed as a basis.

A second direction is the prediction of the onset of an illness using general data averages rather than hourly data. However, achieving accurate predictions in this context necessitates a comprehensive understanding of the medical factors involved and the progression of the illness.

The last research approach is to predict the general development of a patient. For this, different use cases are possible: length of stay, probability of a secondary stay (relapse), and mortality prediction. This last case describes the intention of this thesis: the prediction of death within the hospital stay, within 30 days, and within 365 days.

Up until now, most papers that are based on the MIMIC-III dataset either focus on prediction models with hourly time-series data, as can be seen with Moor et al. 2021 [5], or on natural language processing approaches. Furthermore, a lot of research exists regarding the general prediction of mortality, length of stay, and readmission rates, as can be seen with Purushotam et al., 2017 [6]. Their paper also offered a precise descriptive analysis of the complete MIMIC-III dataset.

When looking at research with a focus on specific illnesses, research regarding heart disease patients is prevalent. However, the methodology therein can be transferred to the stroke use case in this thesis. In this way, a paper by Vazquez et al., 2021 [7] provided a useful reference. They thoroughly examined risk markers for gender subgroups of patients with acute coronary syndrome. As of yet, the focus on underlying factors across subgroups, like gender differences, is not common in this field of research. Another insightful example of a heart disease use case is presented by Barrett et al. 2019 [8], wherein patient mortality was predicted with deep learning networks.

In comparison to that, little research was found for the prediction of mortality of stroke patients. A possible reason for this is that there are fewer cases compared to other illnesses, such as heart failure. Moreover, previous stroke-related research mainly focused on Natural Language Processing (NLP). For example, a model, that was also published through PhysioNet, has been developed to derive the NIHSS (National Institutes of Health Stroke Scale) scoring system from free-text patient discharge summaries [9].

One noteworthy paper in relation to stroke within the MIMIC-III dataset was published by Li et al., in 2022 [10]. They developed a nomogram for mortality prediction of stroke patients. The performance of their model is used as a reference point for the machine learning models in this thesis.

The relative scarcity of research dedicated to stroke is astonishing, considering that it remains one of the primary contributors to global mortality [11]. Stroke comprises various subtypes, each exhibiting distinct patterns of fatality among men and women. In terms of overall stroke-related deaths, Rexrode et al. state that, in 2019, stroke accounted for 6.2% of all female deaths and 4.4% of all male deaths [12]. This discrepancy in prevalence between genders presents an interesting dimension for investigating equal prediction quality across subgroups. The limited amount of stroke-related research within the MIMIC-III dataset, coupled with the critical significance of this illness, was decisive for the selection of this use case.

2.2 Expected Results

Throughout this thesis, multiple data analysis tools and methods, like correlations, clustering, and classification, are applied to the MIMIC-III dataset. It is expected, that these methods can provide solid and reliable results, as there is already a number of comparable research based on the MIMIC-III dataset.

A particular focus lies on the performance and fairness of the implemented prediction models. The AUROC (Area Under the Receiver Operating Characteristic) score is a widely used metric to evaluate the performance of a classification model. It ranges from 0 to 1, where a score of 0.5 represents a random classifier and a score of 1 indicates a perfect classifier.

In a paper by Purushotham et al., this score was estimated to be about 0.75 for comparable machine learning models [6]. They further claim, that their own deep learning model performed between 0.87 to 0.94 depending on the feature set. It must be noted, that their prediction task differed slightly, as they predicted general mortality throughout the complete dataset. However, similar research within MIMIC-III regarding heart failure led to comparable results [8].

Finally, the paper by Li et al. [10], with a focus on stroke mortality, claims that a model simply based on the Oasis score shows up an AUROC score of approximately 0.70, while their nomogram achieved about 0.80. These values can be seen as potential benchmarks for the prediction models in this thesis.

3 Use Case Setup

This chapter describes the selection and filtering of the dataset. Especially the fundamental pre-processing steps that were implemented for this use case are presented in detail. Furthermore, the chapter introduces the supplementary frontend that was developed as part of this research.

3.1 The MIMIC-III Dataset

This master thesis is founded on data from the Medical Information Mart for Intensive Care dataset (MIMIC-III, v1.4) [13]. The data was collected in the Beth Israel Deaconess Medical Center, Boston, between 2001 and 2012. It comprises over 58,000 hospital admissions of over 45,000 individual, deidentified patients who stayed in critical care units. A multitude of features was measured per patient on an hourly basis. As was previously shown, MIMIC-III has already been successfully implemented in previous research papers [5] [6]. Moreover, the dataset has been the foundation for multiple public PhysioNet challenges[2]. Thus, this complete dataset presents a reliable source with sufficient size for machine learning.

While in theory, the data of the MIMIC-III dataset is publicly accessible, it may only be distributed by PhysioNet[3], where a user needs to apply for access. PhysioNet also demands users to finish the “CITI Data or Specimens Only Research” training[4], which emphasizes responsible data handling.

Lastly, a credentialed user is required to accept the Data Use Agreement, wherein one is further requested to publish the code that was employed for dataset analysis. Facilitating this request, the complete code of this thesis is made available through a **GitHub repository**[5], ensuring accessibility and transparency.

As a first step of this thesis, the MIMIC-III dataset was imported into PostgreSQL[6], which proved suitable as a local Relational Database Management System (RDBMS). Necessary database setup files for this process were distributed by PhysioNet[7]. This setup procedure is described in detail in a provided paper [14].

3.2 Filtering and Pre-Processing

3.2.1 Patient Selection

The initial filtering of patients determines, which admission are at all suitable. The complete process is also visualized in Figure 1, in the subsequent section.

[2] `https://physionet.org/about/challenge/moody-challenge`

[3] `https://physionet.org/content/mimiciii/1.4/`

[4] `https://physionet.org/about/citi-course/`

[5] `https://github.com/JayVeezy1/MA_thesis`

[6] `https://www.postgresql.org/`

[7] `https://github.com/MIT-LCP/mimic-code/tree/main/mimic-iii`

While there are 58,976 unique admissions in the dataset, many patients are underage or need to be excluded because of missing data. Moreover, only patients were kept for which actual measurements of vital signs, also referred to as "chart_events-data", were available. This resulted in approximately 40,000 relevant admissions. This filtering step was conducted in line with Moor et al. [15]. Furthermore, their guidelines for structured work with MIMIC-III were mostly adopted for this thesis. They also provided the code for their paper and offered valuable pre-processing references.

Next, the feature "icustay_id" was selected as the relevant key to remove duplicate patients. Otherwise, duplicate instances would occur when using hospital admissions as the primary key, as some patients undergo multiple transfers to the intensive care unit within a single hospital stay. In addition, all ICU stays with less than 24 hours were removed as they do not offer sufficient data. After these filtering steps, there were 32,220 unique icustay_ids left. This filtering step was based on concepts from previous research by Alistair et al. [16].

In numerous publications, an additional filtering step was implemented based on the hospital database system software utilized. Over the course of data collection, the old "CareVue" system was succeeded by the "metavision" system, which resulted in distinct chart_event-ids and item_ids. As a result, it is a tremendous challenge to map the roughly 12,000 "CareVue" variables, encompassing vital signs and lab events, to the 2,000 "metavision" variables. However, a separation would have been regretful, as the 18,837 "CareVue" patients make up a large amount of the MIMIC-III dataset. Still, the complete integration of those two systems was not conducted, as it is a major task in itself. Thus, only for the selected features, described in the following section, the respective "CareVue" features were mapped to the "metavision" features. Future researchers may expand this mapping to additional variables or work exclusively with the "metavision" system.

3.2.2 Use Case Selection

As mentioned in the previous sections, the selected use case for the master thesis is the mortality prediction of stroke patients. The final filtering led to 2,655 patients with stroke.

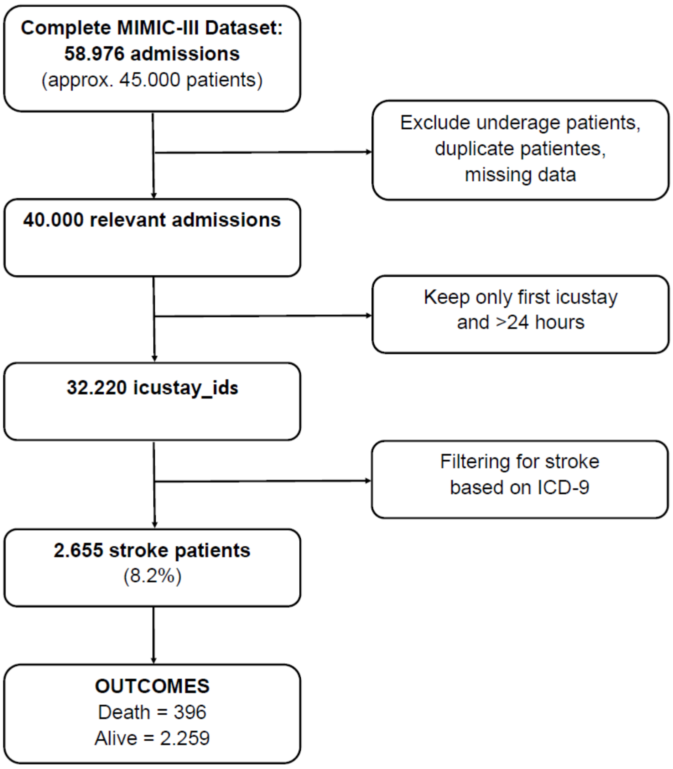

Complete MIMIC-III Dataset:
58.976 admissions
(approx. 45.000 patients)
Exclude underage patients, duplicate patientes, missing data
40.000 relevant admissions
Keep only first icustay and >24 hours
32.220 icustay_ids
Filtering for stroke based on ICD-9
2.655 stroke patients
(8.2%)
OUTCOMES
Death = 396
Alive = 2.259

Figure 1: Overview of Filtering Steps

Patients were filtered for their illnesses based on the ICD9-codes[8] of their diagnosis at admission. ICD-9 codes, which stand for the International Classification of Diseases, Ninth Revision, are a standardized index used for classifying and coding medical diagnoses and procedures. It is possible to change the use case through the selection of ICD9-codes in the supplement file "selection_ICD9_codes.py". This step is fundamental and requires some medical background knowledge. In addition, related papers, for example by Woodfield et al. [17], offer helpful guidance on this topic.

It should be noted, that this selection can have some pitfalls, as some ICD9-titles might seem like they have medical relevance, but they were used differently in the actual diagnosis of patients. For example, not all cerebrovascular ICD9-code titles that sound like they indicate stroke can be accounted as stroke cases. The code "43883 - Facial Weakness" might be considered a valid medical indicator. However, after cross-referencing

[8] https://www.cdc.gov/nchs/icd/icd9.htm

the “diagnosis_text” for all the patients with “43883”, it became clear that those were mainly heart attack patients. Thus, “43883” was removed as a selector for stroke. Of course, if one of those patients showed up with another actually valid ICD9-code, like “430 - Subarachnoid hemorrhage”, then they were kept in the dataset. This example shows, that it is crucial to select the correct ICD9-codes, as a reliable patient selection is one of the most important steps of the analysis.

The following Table 1 offers an overview of the implemented ICD9-codes based on official guidelines[9].

Stroke Type	ICD9 Codes
hemorrhagic	430, 431, 432, 4329
ischemic (& transient ischemic attack):	433, 4330, 4331, 4332, 434, 4340, 43400, 43401, 4341, 43411, 435, 4350, 4351, 4353, 4359, 436
other (& late effects of stroke)	437, 4370, 4371, 4372, 4373, 4374, 438, 4381, 43811, 4382, 43820, 4383, 4384, 4385, 4388, 43882, 43885

Table 1: ICD9-Codes to identify Stroke Types

This filtering is in line with the ICD9 guidelines and should make it possible to differentiate between hemorrhagic and ischemic cases. Hemorrhagic stroke encompasses various types of stroke characterized by internal bleeding within the brain. On the other hand, ischemic stroke refers to cases where an artery is blocked by a blood clot, resulting in a disruption of blood supply to specific regions of the brain.

However, the occurrence of hemorrhagic cases seems higher than expected. Using the diagnosis as a filter for stroke types might be problematic, as it seems that the medical staff sometimes used ICD9-code 430 “general stroke” for ischemic and hemorrhagic cases. This might explain why there are unexpectedly more hemorrhagic cases than ischemic cases, even though ischemic stroke commonly accounts for 80% of all cases [10].

The differentiation between hemorrhagic and ischemic stroke might still be used as a filter. However, for the subsequent analysis, the complete dataset, containing all stroke types, is used. It is important to keep in mind, that ischemic stroke and hemorrhagic stroke have different occurrences and specific characteristics. Thus, it may be beneficial to improve the filtering between these cases. Moreover, different subtypes of these stroke types might offer another promising field of further research. For example, research has shown that while hemorrhagic stroke was more common among men, cardioembolic stroke, which can be more fatal, was more prevalent amongst women [11].

[9] `https://health.mo.gov/data/mica/CDP_MICA/StrokeDefofInd.html`

3.2.3 Feature Pre-Processing

As the MIMIC-III dataset contains over 12,000 available features, identified by their related “item-ids”, these are too many features for any useful analysis. Thus, a user has to select, which features are relevant for the respective use case. Regardless of the user selection, patient demographics and the admission diagnosis are always included. Patient vitals and lab results should be selected after the estimation of relevance. In total, a selection of approximately 40 of the most relevant features is recommended.

For each patient, these pre-selected features are extracted from the original dataset into an individual .csv file. The features included for this use case are displayed below in Tables 3 and 4.

This process is currently solved hardcoded in the setup function 'export_final_dataset' in the supplemented code. For future development, a solution similar to the factorization table, which is described further below, might offer a more flexible approach to this feature selection.

This pre-selection naturally requires some medical knowledge, thus it is also recommended to use previous research as a guideline. A user can later select inside the script or inside the frontend, which of these pre-selected features to actually use for the correlation analysis and the creation of prediction models. Overall, the topic of the pre-selection of features, as well as the actual feature selection for the model creation, still has potential for future improvements.

As another part of the pre-processing pipeline, there are two steps for outlier removal. At first, for each patient and each feature in the time series data, values that are bigger or smaller than 100 times the mean of the respective feature are removed. This intends to remove extreme values like measurement errors. Secondly, the central data table “average_patient_cohort” is derived from the time-series data. From this table, patients whose average values are higher than 10 times the mean of the respective feature are removed. This removes a small amount of patients, who diverge too strongly from mean values.

At last, this “average_patient_cohort” contains all relevant patients with their feature averages. A previous comparison of prediction results from studies where only mean values per Patient were used [7] [10], with research where hourly time-series data were implemented [6] [15] was conclusive. Based on this, it became apparent that mean values are indeed sufficient for initial mortality prediction. Thus, the implementation of this table containing feature averages was chosen as the main data source for the following analysis steps.

Furthermore, the categorical features were factorized with the supplementary table “factorization_table.xlsx”, a section of which is displayed below. The factorized values for each feature were selected manually with the subsequent correlation analysis in mind.

feature	unfactorized_value	factorized_value
admission_type	ELECTIVE	0
admission_type	EMERGENCY	1
admission_type	URGENT	2
ethnicity	UNKNOWN/NOT SPECIFIED	0
ethnicity	WHITE	1
ethnicity	ASIAN	2
ethnicity	HISPANIC OR LATINO	3
ethnicity	BLACK	4
ethnicity	OTHER	5
insurance	Government	0
insurance	Self Pay	1
insurance	Medicaid	2
insurance	Medicare	3
insurance	Private	4
marital_status	SINGLE	-1
marital_status	DIVORCED	-1
marital_status	SEPARATED	-1
marital_status	WIDOWED	-1
marital_status	no_data	0
marital_status	UNKNOWN (DEFAULT)	0
marital_status	MARRIED	1
marital_status	LIFE PARTNER	1
stroke_type	ischemic	-1
stroke_type	other_stroke	0
stroke_type	hemorrhagic	1

Table 2: Factorization Table

The factorization has the advantage, that the correlations of most categorical features, which are discussed in detail in the subsequent Chapter 4.2, can be interpreted more clearly. For example, this can be seen for the feature "admission_type", which shows up a positive correlation to death. This indicates, that the mortality rate is higher for more urgent admission.

In addition to this, the factorization of categorical data is crucial for the performance of subsequent prediction models, as recall rates were only at approximately 0.10 without it.

However, there are certain features, for which factorization may not be the most appropriate approach. For instance, the marital status is mapped to -1 for single, divorced, or widowed individuals. Further, the value 0 is assigned for "no data", and 1 for anyone in a relationship. Surprisingly, the mentioned correlation analysis indicates that singles have a higher survival rate. This finding contradicts the research by Goulart et al. [18], which suggests that individuals with a lower socioeconomic position, including widowed and divorced individuals, tend to have lower long-term survival rates after experiencing a stroke.

One possible explanation for this discrepancy is that the aggregation of original feature values into the combined group “-1” was not an appropriate approach. Overall, the factorization of “marital_status” does not yield the expected correlation results, highlighting the need for further research in this area.

This raises the general question of whether the factorization based on a supplementary table is a reliable solution. Therefore, one-hot encoding was also implemented as an alternative. With this process, each categorical value is transformed into a new, binary column. It presents a simpler approach, which is also less prone to errors. In hindsight, encoding of categorical features seems more promising and the factorization process should be reconsidered.

3.2.4 Feature Selection

The selection of features can be managed by a user either inside the supplementary table “feature_preprocessing_table.xlsx” or directly inside the frontend. Based on the results of the subsequent correlation analysis, the following 17 features are selected as the main independent features. The use case of this thesis focuses on the prediction of in-hospital mortality. However, in Table 3, further dependent variables, which are available for alternative prediction tasks, are listed.

Continuous	Categorical	Dependent Variables
age	admission_type	death_in_hosp
Anion Gap	electivesurgery	death_3_days
Bicarbonate	ethnicity	death_30_days
Chloride (whole blood)	gender	death_180_days
Creatinine	mechvent	death_365_days
gcs	stroke_type	
Heart Rate		
O2 saturation pulseoxymetry		
oasis		
Sodium (whole blood)		
White Blood Cells		

Table 3: Initial Feature Selection

One feature of note is the Oasis score[10] (Oxford Acute Severity of Illness Score), which can be implemented as a scoring system for patient mortality risk. The calculation of

[10] `https://alistairewj.github.io/project/oasis/`

the score is based on ten other variables, like ventilation or pre-ICU length of stay. This feature shows up the highest correlation to "death_in_hospital" and is also very influential for prediction models. Some researchers reckon that the scoring system could already be employed as a standalone tool to monitor patients' health levels independently [19].

The technical process to derive the score from these base features has been implemented by previous research and was made available via GitHub[11]. The corresponding SQL scripts were integrated for the setup process of this analysis.

Furthermore, the features displayed in the subsequent Table 4 are available for future, alternative implementation. Some of these features have already been examined for correlation. However, they were not included in the later prediction models, as their relation to mortality was not clearly established.

Moreover, there are some features, that have not been included in this analysis, as of yet. They can be selected in the frontend for additional avenues of research.

Lastly, some features can only be included for filtering. It is not recommended to utilize these features for correlation or the creation of a prediction model, as they are solely intended for filtering purposes. Either they have no death-related influence, like the "icustay_id", or their format is not usable for standard predictors, such as "diagnosis_text", which might be better suited for a Natural Language Processing implementation.

Only for Correlations	**Unused Features**	**Only for Filtering**
dbsource	Arterial Blood Pressure	subject_id
gauges_total	Glucose (whole blood)	hadm_id
patientweight	insurance	icustay_id
marital_status	religion	icustays_count
sepsis_flag	diabetes_flag	infarct_type
cancer_flag	Respiratory Rate	diagnosis_text
obesity_flag		dob (date of birth)
drug_abuse_flag		dod (date of death)
hypertension_flag		intime
		outtime
		preiculos

Table 4: Further available Features

In general, it is not recommended to incorporate an excessive number of features and it was later discovered that this initial selection of 17 features proved to be overly abundant. The implemented frontend makes it possible to conveniently analyze different feature selections. With this, it became evident that a refined selection consisting of only nine features yields improved clustering outcomes.

[11] https://github.com/caisr-hh/Dayly-SAPS-III-and-OASIS-scores-for-MIMIC-III

3.2.5 Cache Solution

Once the patients were filtered for the intended use case, the data of each single patient was exported from the PostgreSQL database. Saving the time series data as .csv files enables the concurrent import into the Python[12] environment. While this leads to some data redundancy, the required memory space is still reasonable.

However, the process of reading single files for each analysis can be time-consuming. Thus a basic cache system was developed to improve runtime performance. For this, the "Patients"-Class was constructed, to save the related patient objects as a serialized "pickle"-object. This cached file can be imported more quickly. It only takes approximately five seconds to execute the method "load_patients_from_cache" for all 2,655 patients, while the loading time for the .csv files took approximately one minute.

Currently, the analysis is founded on the previously introduced "average_patient_cohort", which can be saved as a single .csv and thus has minimal loading time. However, for the creation of this cohort, the reload of all patients is necessary. This is required when a new feature is added to the pre-selection or a different scaling method is applied. Moreover, this caching solution supports any potential future analysis based on time-series data.

3.3 Frontend

As a central part of this thesis, a frontend was developed to enable the visualization of the complete analysis.

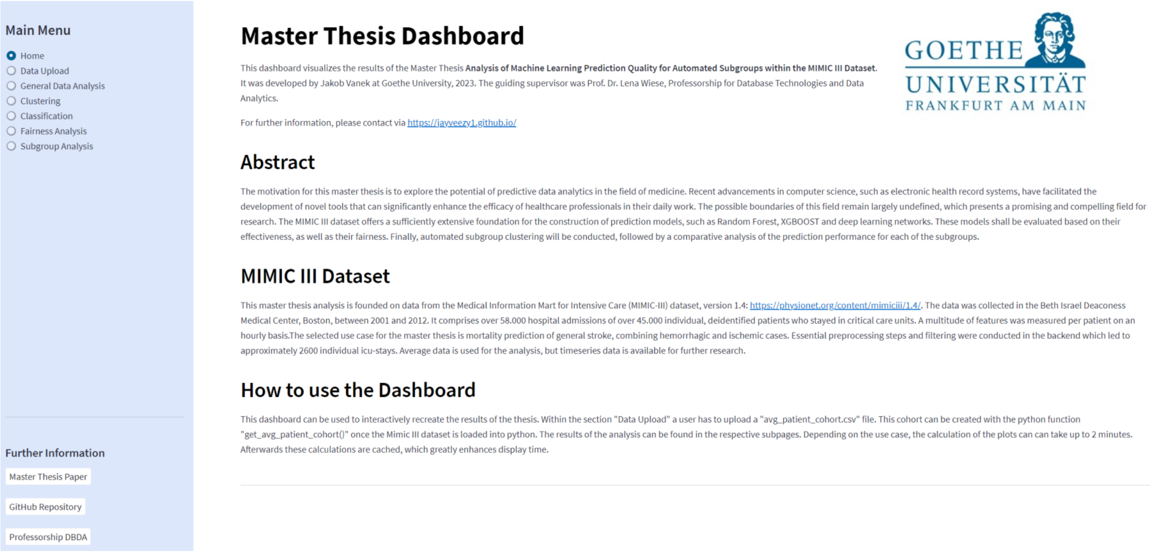
The first approach for this was to integrate the dashboard for Automated Subgroup Detection and Fairness Analysis, also referred to as"ASDF-Dashboard"[13]. This tool was developed by Schäfer and Wiese to enable the automatic detection of subgroup imbalance within a binary classifier [20]. The dashboard provides an in-depth analysis of fairness and it is possible to upload the classified MIMIC-III data therein. However, the current version of the dashboard solely focuses on fairness analysis and offers no general data analysis or integrated classification methods. Thus, the current "ASDF-Dashboard" was not implemented for the visualization of this master thesis.

Instead, a lightweight Streamlit[14] dashboard was developed to offer a quick and user-friendly insight. The frontend can be easily opened locally, by running the "main.py" function of the code, which is distributed through the previously mentioned GitHub repository.

[12] https://docs.python.org/3/

[13] https://github.com/jeschaef/ASDF-Dashboard

[14] https://docs.streamlit.io/



Master Thesis Dashboard

This dashboard visualizes the results of the Master Thesis **Analysis of Machine Learning Prediction Quality for Automated Subgroups within the MIMIC III Dataset**. It was developed by Jakob Vanek at Goethe University, 2023. The guiding supervisor was Prof. Dr. Lena Wiese, Professorship for Database Technologies and Data Analytics.

For further information, please contact via https://jayveezy1.github.io/

Abstract

The motivation for this master thesis is to explore the potential of predictive data analytics in the field of medicine. Recent advancements in computer science, such as electronic health record systems, have facilitated the development of novel tools that can significantly enhance the efficacy of healthcare professionals in their daily work. The possible boundaries of this field remain largely undefined, which presents a promising and compelling field for research. The MIMIC III dataset offers a sufficiently extensive foundation for the construction of prediction models, such as Random Forest, XGBOOST and deep learning networks. These models shall be evaluated based on their effectiveness, as well as their fairness. Finally, automated subgroup clustering will be conducted, followed by a comparative analysis of the prediction performance for each of the subgroups.

MIMIC III Dataset

This master thesis analysis is founded on data from the Medical Information Mart for Intensive Care (MIMIC-III) dataset, version 1.4: https://physionet.org/content/mimiciii/1.4/. The data was collected in the Beth Israel Deaconess Medical Center, Boston, between 2001 and 2012. It comprises over 58.000 hospital admissions of over 45.000 individual, deidentified patients who stayed in critical care units. A multitude of features was measured per patient on an hourly basis.The selected use case for the master thesis is mortality prediction of general stroke, combining hemorrhagic and ischemic cases. Essential preprocessing steps and filtering were conducted in the backend which led to approximately 2600 individual icu-stays. Average data is used for the analysis, but timeseries data is available for further research.

How to use the Dashboard

This dashboard can be used to interactively recreate the results of the thesis. Within the section "Data Upload" a user has to upload a "avg_patient_cohort.csv" file. This cohort can be created with the python function "get_avg_patient_cohort()" once the Mimic III dataset is loaded into python. The results of the analysis can be found in the respective subpages. Depending on the use case, the calculation of the plots can can take up to 2 minutes. Afterwards these calculations are cached, which greatly enhances display time.

Figure 2: Frontend Landing Page

The frontend contains a page for each of the subsequent chapters in this thesis. It especially enables quick descriptive data analysis and correlations. Furthermore, it offers multiple clustering algorithms and their visualization. Next, different classification methods can be used to develop prediction models. Finally, the frontend includes fairness analysis, either based on manually selected subgroups, or across subgroups derived from clustering methods.

4 Dataset Exploration

In this chapter, a descriptive analysis of the stroke use case within the MIMIC-III dataset is implemented. This is concluded by a correlation analysis of the previously selected features.

4.1 General Statistics of the Stroke Use Case

The following descriptive perspective shall give a broad insight into the dataset. The stroke use case contains 2.655 unique ICU stays, which make up about 8.2% of the available MIMIC-III dataset.

As previously mentioned, the filtering for stroke subtype based on ICD9-codes proved to be unreliable. Unfortunately, there is no official feature available to clearly differentiate those two stroke types. Nonetheless, there were 569 ischemic cases found, which is sufficiently close to the nomogram research by Li and Li [10], who identified 767 ischemic cases. The divergence might be explained by differing previous filters and the aforementioned unclear categorization based on the ICD9-codes. It should be highlighted that the missing disclosure of the ICD9 codes, which were utilized within other studies for filtering purposes, posed a hindrance to reproducing the work of these researchers.

The subsequent Table 5 offers an overview of influential features within the selected dataset. There are only minimal disparities in the distribution between genders, with the majority of cases occurring in individuals aged 66 years and above. However, the prevalence of hypertension and mechanical ventilation (mechvent) are noteworthy.

Of particular significance is the striking disparity in the distribution of ethnicities within the dataset. Approximately 70% of patients self-report as “white” ethnicity. This notable skew in representation must be acknowledged throughout subsequent analyses. Such unequal representation within the dataset can potentially introduce biases.

Moreover, further investigation revealed that the high number of missing values for the “O2 saturation pulseoxymetry” feature predominantly occurred in the patients from the “CareVue” database system. This observation highlights a potential data quality issue specific to this system. The cause of this data gap warrants further research and evaluation to determine whether it stems from mapping inconsistencies or other factors.

Variables	**Classification**	**Count**	**Missing Values**
total_count	icustay_ids	2655	-
dbsource	both	8	0
dbsource	carevue	1415	0
dbsource	metavision	1232	0
stroke_type	hemorrhagic	1556	0
stroke_type	ischemic	569	0
stroke_type	other_stroke	530	0
death_in_hosp	0	2259	0
death_in_hosp	1	396	0
age	(17.999, 42.33]	211	0
age	(42.33, 66.67]	1019	0
age	(66.67, 91.0]	1425	0
ethnicity	UNKNOWN/NOT SPECIFIED	279	0
ethnicity	WHITE	1846	0
ethnicity	ASIAN	78	0
ethnicity	HISPANIC OR LATINO	112	0
ethnicity	BLACK	252	0
ethnicity	OTHER	88	0
gender	F	1353	0
gender	M	1302	0
Heart Rate	(48.999, 79.67]	1273	10
Heart Rate	(79.67, 110.33]	1331	-
Heart Rate	(110.33, 141.0]	40	-
O2 saturation pulseoxymetry	(70.999, 82.33]	3	1413
O2 saturation pulseoxymetry	(82.33, 93.67]	39	-
O2 saturation pulseoxymetry	(93.67, 105.0]	1199	-
oasis	(6.999, 24.0]	452	20
oasis	(24.0, 41.0]	1785	-
oasis	(41.0, 58.0]	397	-
cancer_flag	0	2247	0
cancer_flag	1	408	0
diabetes_flag	0	2023	0
diabetes_flag	1	632	0
hypertension_flag	0	1138	0
hypertension_flag	1	1517	0
mechvent	0	1230	0
mechvent	1	1425	0

Table 5: Feature Distribution for the Stroke Use Case within MIMIC-III

Table 6 enables a closer examination of the distribution of mortality within the dataset. It is noteworthy, that the category "death in hospital" contains mortality cases with varying time frames, up to half a year.

case	total	death_3_days	death_30_days	death_180_days	death_365_days
death_in_hosp	396	203	188	5	0
death_not_in_hosp	580	26	244	212	98
total_deaths	976	229	432	217	98
total_deaths_perc	37%	9%	16%	8%	4%
alive	1679	-	-	-	-
total	2655	-	-	-	-

Table 6: Occurrences of Death in Stroke Dataset

Only minor gender differences can be detected for "death in hospital"-mortality within the dataset. The mortality rate for female patients is around 15%, with 203 to 1353 cases. Male mortality is around 14.82%, with 193 to 1302 cases. This remains similar for cumulated annual mortality. Here, female mortality is around 38% and male mortality around 35%. This data is in line with previous stroke research inside MIMIC-III [10]. However, these overall statistics do not indicate any underlying gender differences for stroke, which are also discussed within the literature. This may be explained by the relatively small sample size of only 2,655 instances within the MIMIC-III use case.

4.2 Feature Correlations

The following correlation analysis helps identify the most relevant features for the mortality of stroke patients within the MIMIC-III dataset.

The statistical significance of each feature was tested and their p-values are symbolized inside the brackets. The p-value roughly indicates the probability of uncorrelated features producing a dataset with a correlation at least as extreme as the original datasets. Thus a high p-value indicates, that the 0-Hypothesis "the features are not correlated" is more probable. Statistical significance can be assumed if the p-value is below 0.05, which is represented with a single asterisk (*). Even higher confidence levels are 0.01 (**) and 0.001 (***).

Continuous variables were evaluated using Pearson's Correlation Coefficient, with the significance test conducted using Pearson's R. Binary or flag features also utilized Pearson's Correlation Coefficient, with the Chi-Squared test applied for significance testing. Categorical features with more than two values employed Theil's U for correlation calculation, and the Chi-Squared test for significance analysis.

As previously mentioned, the Oasis score exhibits the highest correlation with the dependent variable, at approximately 0.4. Thus, the Oasis score can serve as a valuable benchmark for the subsequent development of predictive models, providing a baseline for assessing the models' performance.

Another feature with a comparably high correlation coefficient of 0.36 is "mechvent", which represents the need for mechanical ventilation of a patient. This fits the negative correlation of "O2 saturation pulseoxymetry", which measures a patient's 02-saturation. These features clearly show how crucial a stroke patient's oxygen supply is.

Furthermore, the "gcs" attribute represents the Glasgow Coma Scale[15], which is a widely used scale to assess patients with acute brain injury. Here a higher value correlates with better chances for survival.

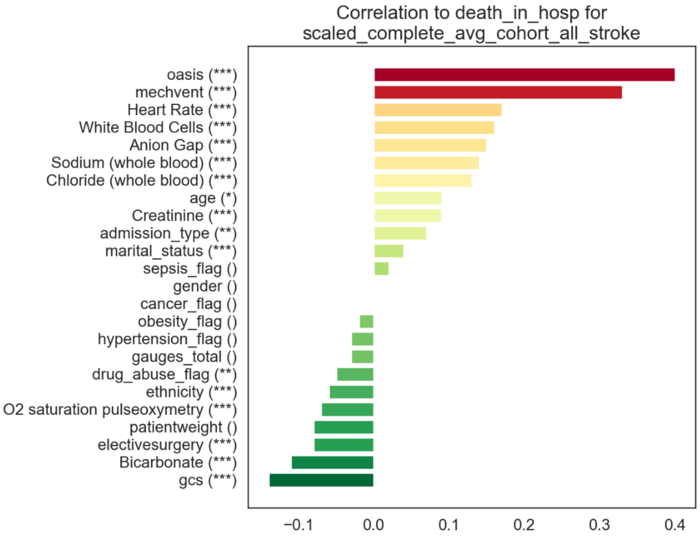


Figure 3: Correlations of relevant Features for Complete Stroke Use Case

As can be seen in Figure 3, the remaining features have a less strong correlation towards the dependent variable, mostly below 0.2. Still, they may be used in a combined manner for the prediction of patient mortality.

However, the analysis of feature correlations also revealed some unexpected findings. While certain vital signs such as "Heart Rate" and "Bicarbonate" align with existing research regarding mortality risk, there were other features that yielded surprising results.

One feature that was anticipated to have a more pronounced influence on patient outcomes was "gauges_total". The use of gauges is a common method employed to reduce brain pressure and is crucial for stabilizing patients. It was expected that patients requiring a more significant reduction in brain pressure would be at higher risk. However, despite a potential connection to survival, the statistical analysis does not reveal a strong

[15] https://www.ncbi.nlm.nih.gov/books/NBK513298/

significance for this feature. This unexpected finding suggests that other factors may have a more dominant role in determining patient outcomes.

Moreover, the limited influence observed for binary features, that indicate secondary illnesses, was not anticipated. For instance, cancer is widely recognized to have a significant correlation with patient mortality. However, the analysis suggests that these flag features may not have an immediate impact on the outcome variable “death_in_hosp”. An explanation for this might be that cancer is not a direct factor for in-hospital survival but only has an impact on survival rates after the hospital stay.

In line with this assumption, the following Figure 4 demonstrates higher correlations of comorbidities with “death-within-365-days”. Additionally, a higher correlation with age is also evident for annual mortality.

Nevertheless, it was initially expected that age and secondary illnesses such as hypertension or drug abuse would also exert a stronger correlation to in-hospital mortality.

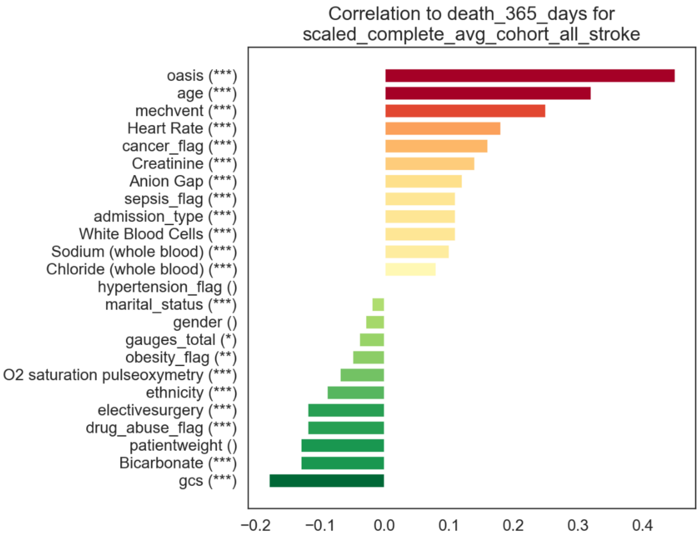


Figure 4: Correlations to Death within 365 Days

One feature, that unexpectedly shows no statistical significance in its correlation to any of the dependent variables is “patientweight”. A possible explanation for this is, that this feature may not have reliable values in one of the two database systems, either “CareVue” or “metavision”.

To investigate further, Figures 5 and 6 were generated to compare the correlations within each dataset separately. Both analyses confirm the general lack of statistical significance for the “patientweight” feature. Similarly, another feature that exhibits no statistical significance within the “CareVue” system is ”O2 saturation pulseoxymetry”. These findings suggest that there may be missing values or mapping errors associated with these features in the “CareVue” system. It is recommended to reevaluate the mapping process and assess data quality issues that may contribute to these results.

Despite these specific differences, the overall correlations within these two subsets appear to be similar. This consistency implies that the relationships between the remaining features and mortality risk are relatively stable across the different database systems.

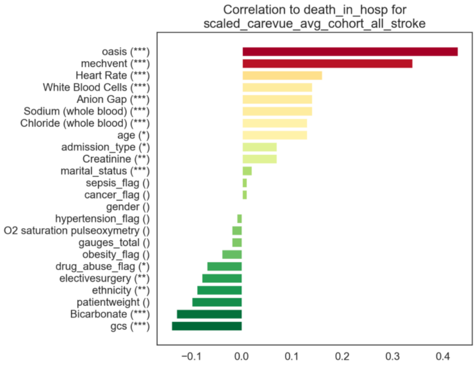


Figure 5: Subset Carevue Database

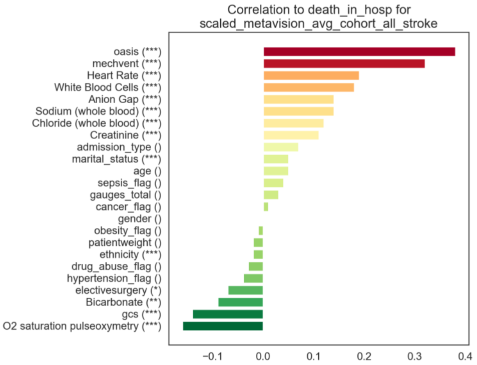


Figure 6: Subset Metavision Database

In conclusion of the correlation analysis, it became evident that certain variables can be utilized for the subsequent prediction model. While most vital signs exhibit a relatively weak correlation individually, the combination of multiple features may support reliable predictions.

4.3 Data Visualization

When confronted with high-dimensional data, employing dimensionality reduction techniques can be a valuable approach for visualizing the dataset. One reliable method for this purpose is the PaCMAP algorithm [21]. With this, high-dimensional data is reduced to three dimensions, revealing discernible clusters. These dimensions do not represent a real feature, but the resulting shape of data instances can still offer valuable insights.

This can be seen within the following visualizations. Figure 7 incorporates multiple categorical features, resulting in the noticeable separation into distinct clusters. Conversely, Figure 8 demonstrates the general proximity of most patients when considering mainly continuous vital signs. Herein, the division into the two groups can be attributed to the presence of the “gender” variable, which represents the sole categorical feature. Additionally, the “death” cases have been highlighted in both visualizations as they serve as the dependent variable.

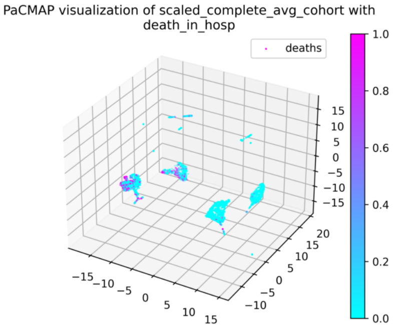


Figure 7: PaCMAP multiple Features

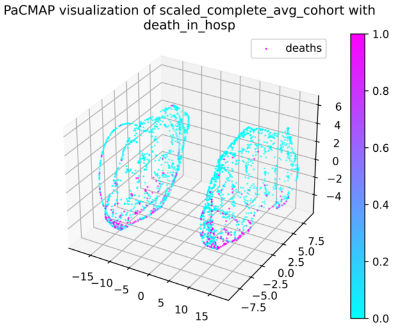


Figure 8: PaCMAP continuous Features

This method of dataset visualization is also implemented in the following chapter to display the clustering results.

5 Clustering Analysis

This chapter compares the clustering results of multiple clustering algorithms for the stroke use case within the MIMIC-III dataset.

The kMeans[16] and kPrototype[17] algorithms are widely used clustering methods that require manual selection of the cluster count. Furthermore, the DBSCAN[18] (Density-Based Spatial Clustering of Applications with Noise) approach is utilized, as it offers an automated approach for determining the optimal number of clusters. In addition, the SLINK[19] (single-linkage) clustering method is introduced at the end of this chapter as an alternative approach.

A standard metric for evaluating the quality of clusters is the sum of squared errors (SSE), which calculates the squared distance between each data point and its corresponding cluster center. The SSE provides a measure of how compact the clusters are, with lower values indicating tighter and more well-defined clusters.

Another metric for cluster assessment is the silhouette score, which offers a more comprehensive evaluation by considering both the cohesion within clusters and the separation between clusters. It ranges from -1 to 1, with 1 being the optimal value indicating well-separated and internally homogeneous clusters. For kMeans and kPrototype the selection of an appropriate number of clusters can be guided by these key indicators. However, it is crucial to consider the trade-off between a high silhouette score, or a low SSE, and the interpretability of the clusters. Therefore, it is advisable to avoid selecting an excessive number of clusters, as this could reduce the interpretability of the results.

For the clustering process, the categorical variables were one-hot encoded to make them compatible with the kMeans and DBSCAN clustering algorithms. By converting the categorical variables into binary indicators, their values can be effectively utilized. For the kPrototype clustering algorithm, encoding was not necessary, as it accommodates both continuous and categorical variables [22].

[16] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`

[17] `https://github.com/nicodv/kmodes`

[18] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html`

[19] `https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html`

5.1 Comparison of Clustering Algorithms

The optimization of clustering results, displayed in the three figures below, shows that kMeans and DBSCAN exhibit slightly higher silhouette scores compared to kPrototype. In addition, DBSCAN holds the advantage of being an unsupervised clustering method. However, it tends to generate an excessive number of clusters, which may restrict interpretability. Therefore, the most compelling choice for clustering based on this feature selection is kMeans.

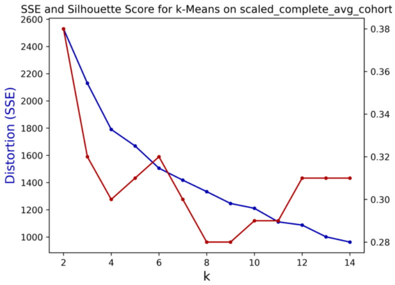


Figure 9:
Optimization for kMeans

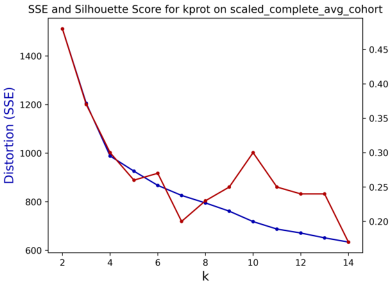


Figure 10:
Optimization for kPrototype

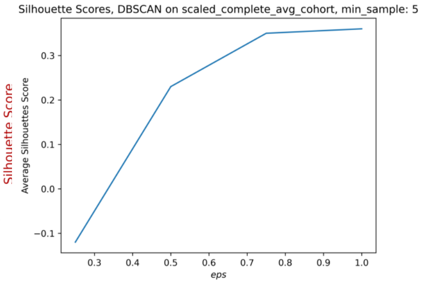


Figure 11:
Optimization for DBSCAN

On a side note, visualizing the SSE for DBSCAN was not feasible, as of yet. Additionally, DBSCAN can be further optimized by tuning the epsilon parameter and the "number of neighbors" parameter. A thorough exploration of these parameters is left for future research.

Following the analysis of the silhouette scores, the optimal clustering results can be displayed upon PaCMAP visualizations. It becomes evident that conducting an analysis with more than 15 clusters, as observed in the case of DBSCAN, does not yield easily interpretable results. The higher number of clusters makes it challenging to discern meaningful patterns or distinctions between the clusters.

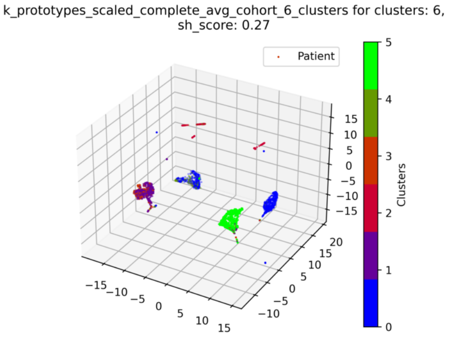


Figure 12: kPrototype Clustering

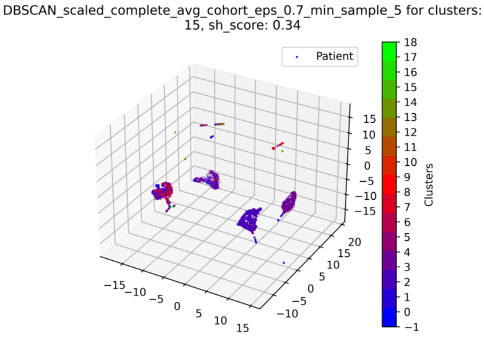


Figure 13: DBSCAN Clustering

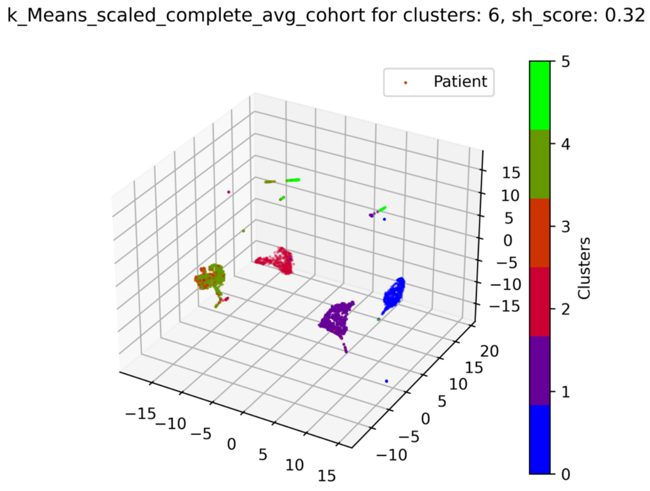


Figure 14: kMeans Clustering

It is important to note that even the highest obtained clustering result only yielded a silhouette score of 0.32, which falls below the desired range. A silhouette score of at least 0.5 is commonly considered more favorable for distinct clusters.

With the following Table 7, it is possible to further investigate the composition of the resulting kMeans clusters. Herein, cluster three is noteworthy as it only contains positive cases. This cluster is congruent with many mortality cases, which were highlighted within Figure 7, in the previous chapter. Closer inspection also shows that in this cluster, there are mostly instances with high Oasis scores, as well as older patients. A similar comparison of clusters is conducted within the subgroup analysis in Chapter 8.

Variables	**Classification**	**complete_set**	**cluster_0**	**cluster_1**	**cluster_2**	**cluster_3**	**cluster_4**	**cluster_5**
total_count	icustay_ids	2655	481	657	556	229	529	203
death_in_hosp	0	2259	476	648	425	0	529	181
death_in_hosp	1	396	5	9	131	229	0	22
age	(-0.001, 0.33]	211	33	60	41	14	36	27
age	(0.33, 0.67]	1019	152	265	205	60	213	124
age	(0.67, 1.0]	1425	296	332	310	155	280	52
dbsource	carevue	1415	0	657	0	229	529	0
dbsource	both	8	1	0	4	0	0	3
dbsource	metavision	1232	480	0	552	0	0	200
stroke_type	ischemic	569	109	133	124	34	124	45
stroke_type	other_stroke	530	125	177	82	19	100	27
stroke_type	hemorrhage	1556	247	347	350	176	305	131
oasis	(-0.001, 0.33]	464	146	260	5	0	7	46
oasis	(0.33, 0.67]	1799	322	380	400	118	439	140
oasis	(0.67, 1.0]	372	7	5	151	111	83	15

Table 7: Cluster Comparison

5.2 Clustering Results for the Reduced Feature Set

As was previously mentioned, a second selection of fewer features led to improved results. The remaining nine features are:

- Anion Gap,
- ethnicity,
- gcs,
- gender,
- Heart Rate,
- O2 saturation pulseoxymetry,
- oasis,
- Sodium (whole blood),
- White Blood Cells.

Overall, this narrowed-down feature selection yields clustering outcomes with higher silhouette scores. Notably, both kMeans and DBSCAN show significant improvements in their clustering performance. Nevertheless, even the highest achievable silhouette score for kMeans, which reaches approximately 0.47, cannot be regarded as a groundbreaking outcome. Moreover, kPrototype still demonstrates lower scores, which indicates that the approach is less suitable for this use case.

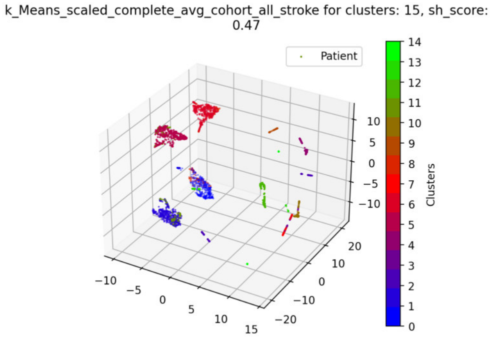


Figure 15: Optimized kMeans Clustering

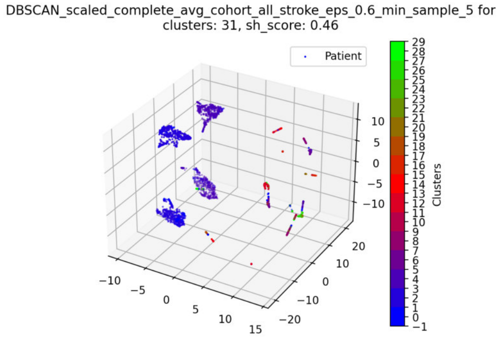


Figure 16: Optimized DBSCAN Clustering

5.3 SLINK Clustering Alternative

In addition to the previous methods, the agglomerative SLINK algorithm was implemented in the frontend. Research by Schäfer and Wiese concluded that the SLINK clustering provided the most reliable results for three different use cases [20]. In the context of stroke mortality, the algorithm produces clustering results that are nearly comparable to those of kMeans.

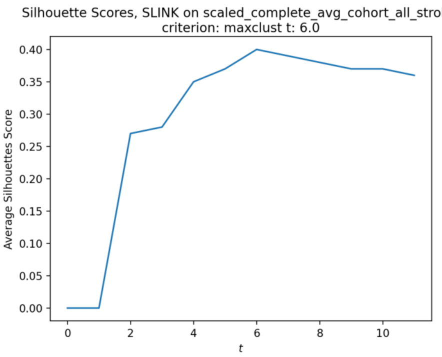


Figure 17: Optimized SLINK Clustering

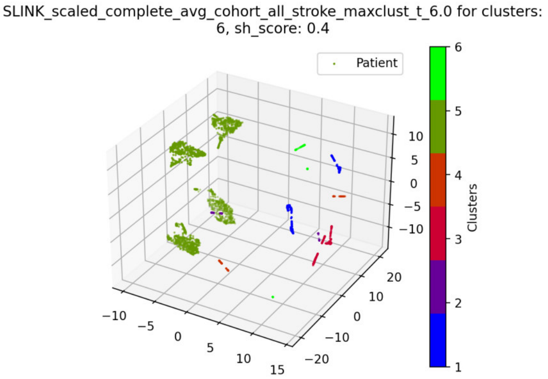


Figure 18: SLINK Clustering

In addition, this approach may offer valuable insights through the analysis of the corresponding dendrogram displayed below in Figure 19. By leveraging this feature, it may be possible to further optimize the cluster count and additionally enhance the interpretability of the results.

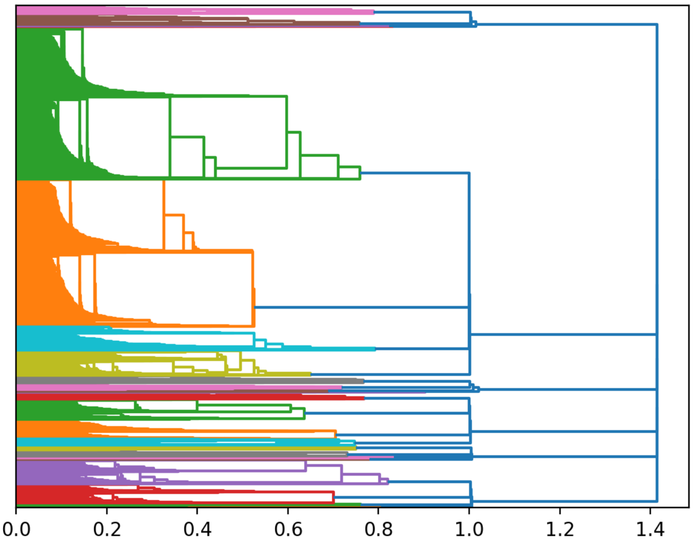


Figure 19: SLINK Dendrogram

Moreover, there are multiple separation criteria available, based on which the clustering can be implemented. However, these alternative selections and a more thorough analysis of the dendrogram were not included, as kMeans was deemed sufficient for the scope of this thesis. The optimization of the SLINK algorithm requires further investigation and thus this alternative was not selected for the following process.

In summary, kMeans continues to be the preferred clustering method, even for this improved feature selection, due to its robust performance and simple implementation. As a result, it serves as the foundation for subgroup detection in Chapter 8.

6 Stroke Mortality Prediction

Following the exploration of the dataset, the performance results of multiple prediction models are compared for the stroke use case within MIMIC-III in this chapter. In addition, the models are also implemented on a reduced feature set. Lastly, the feature relevance within the XGBOOST model is examined more closely.

6.1 Prediction Results Baseline

The features for the classification models were chosen in line with the initial selection of 17 features from the previous correlation analysis. While there are many more features available in MIMIC-III, using more features does not necessarily improve prediction quality but increases complexity and makes it harder to interpret results.

The chosen prediction metrics in the classification report are standard guidelines to evaluate prediction quality. They are calculated based on the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) with the following formulas [23]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

At first, a basic, exemplary Random Forest classifier is constructed based on raw data to create a baseline for the prediction models. This initial test can indicate if it is at all possible to reach conclusive results and already delivers some important insights. Figure 20 displays the resulting confusion matrix for this rudimentary model. Based on this matrix, it is possible to retrace the absolute numbers of true and predicted cases.

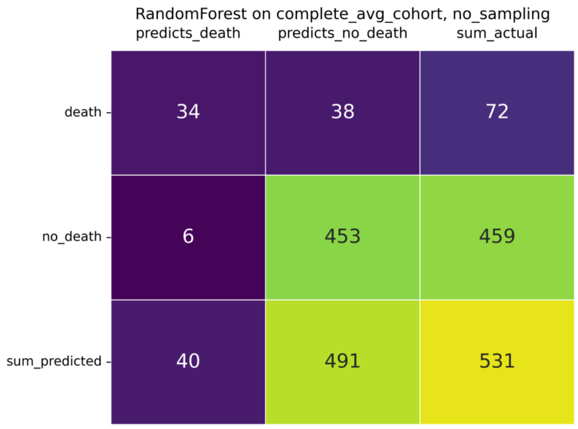


Figure 20: Confusion Matrix for Random Forest, not scaled

On a side note, for a more detailed analysis of the subsequent models, it is also possible to consult these confusion matrices. However, as the actual model evaluation is based on the performance metrics from the corresponding classification reports, the confusion matrices are omitted for most of the following models. They can optionally be visualized within the supplemented frontend.

Based on the values from the previous confusion matrix and the formulas displayed above, the following classification report is calculated for this model.

	precision	recall	f1-score	support
0 (no_death)	0.92	0.99	0.95	459
1 (death)	0.85	0.47	0.61	72
accuracy			0.92	531
macro avg	0.89	0.73	0.78	531
weighted avg	0.91	0.92	0.91	531

Table 8: Classification Report for Random Forest, not scaled

The classification report in Table 8 summarizes the prediction quality for this approach, which is based on raw data without scaling. This model already enables a seemingly high accuracy of 0.92 and a recall rate of 0.47. While the prediction quality of such a simple classifier is not sufficient, it can be seen as a baseline for further models.

However, there are two noteworthy aspects to consider. First, although the accuracy may appear high at over 90%, this can be attributed to the challenge of dealing with highly unbalanced data. The model tends to consistently classify patients as “no_death” since there is a significantly higher number of patients, who do not die. As a result, the apparent prediction quality appears to be high. Therefore, in the context of this medical use case, the recall rate for death cases assumes greater importance. In this case, it lays only at 0.47.

The other aspect is, that the feature “oasis” is used for these predictions. As was shown in the previous correlation analysis, it has a very high correlation to the dependent variables. When removing “oasis” from the selected features, recall decreases further to 0.36. However, this score was developed mainly based on the MIMIC-III dataset and thus its applicability on other, more general datasets, remains a point for future research. The strong dependence on one feature can also be seen as critical. Hence, the inclusion of “oasis” for prediction models is an open point for further discussion.

The relatively low recall rate, compared to previous research, is a motivation to improve the prediction model. Consequently, the following pre-processing methods are implemented to improve this Random Forest classifier.

The scaling of continuous features is a common practice in data preprocessing to ensure that features with large variances do not dominate the analysis compared to features with lower variances. By scaling all features to a range between -1 and 1, their influences are effectively equalized. An alternative to this is to scale feature values between 0 and 1, which is referred to as normalization. The chosen method for this step was min-max normalization where the scaled values, or z-values, are scaled between 0 and 1, based on the following formula:

$$scaled\_value = \frac{(current\_value - feature\_minimum)}{(feature\_maximum - feature\_minimum)} \tag{5}$$

In addition, for most classifiers, it is not sensible to use categorical features in one column. Accordingly, one-hot encoding of these features is used to transform the single columns into multiple binary columns. These important pre-processing steps lead to the following classification report.

	precision	recall	f1-score	support
0	0.92	0.98	0.95	459
1	0.77	0.47	0.59	72
accuracy			0.91	531
macro avg	0.85	0.73	0.77	531
weighted avg	0.90	0.91	0.90	531

Table 9: Classification Report for Random Forest, scaled and encoded

While these adjustments did not directly improve the classification result, they were essential for subsequent steps. One explanation for the nonexistent effect is that the most important features, the Oasis score and mechanical ventilation, were not affected by this directly, so this basic Random Forest model did not change its decision process.

6.2 Comparison of Prediction Models

The following sections present the prediction results of four different classifier models that were developed for the stroke use case. Two versions of Random Forest models, an XGBOOST model, and a neural network approach are compared.

6.2.1 Random Forest with SMOTE Oversampling

The dataset exhibits a significant class imbalance, with a considerably lower number of "death" cases compared to "non-death" cases. This class imbalance poses a challenge for prediction algorithms. By employing appropriate sampling techniques, it becomes possible to mitigate the impact of class imbalance and enhance the performance of prediction algorithms.

One potential approach to address the class imbalance is through undersampling techniques, such as NearMiss[20], which aims to reduce the over-represented class until both classes are more balanced in occurrence. However, in this particular case, undersampling does not yield favorable results, as it leads to a significantly reduced dataset size, which negatively impacts overall prediction performance.

For this particular case, a more effective sampling method is oversampling, which involves generating new instances of the less frequent class, which are the "death" cases in this scenario. These new instances are created by synthesizing values based on the similarity to existing instances. The SMOTE algorithm is a widely recognized and reliable approach that utilizes kMeans clustering to facilitate the oversampling process [24]. As the dataset is composed of categorical and continuous features, the variant SMOTENC (Synthetic Minority Oversampling Technique for Nominal and Continuous data) algorithm[21] is deemed most useful. This variant is implemented in the code, and for simplicity, it is referred to as "SMOTE" in the following sections.

The subsequent results demonstrate a substantial improvement in prediction performance, which is achieved through the balancing strategy. Although there is a slight decrease in accuracy, there is a notable enhancement in the recall, which is now at 0.67. Therefore, SMOTE is consistently employed in all subsequent models, unless explicitly stated otherwise.

[20] `https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.NearMiss.html`

[21] `https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTENC.html`

	precision	recall	f1-score	support
0	0.94	0.88	0.91	459
1	0.47	0.67	0.55	72
accuracy			0.85	531
macro avg	0.71	0.77	0.73	531
weighted avg	0.88	0.85	0.86	531

Table 10: Classification Report for Random Forest, balanced with SMOTE

In addition to the aforementioned evaluation metrics, the overall model performance is commonly assessed using the AUC-ROC (Area Under the Curve of the Receiver Operating Characteristic) and its corresponding AUROC score. This metric illustrates the trade-off between the true positive rate and the false positive rate. Ideally, a value close to 1 is desired, indicating a high discriminatory power of the model.

Furthermore, it is beneficial to consider the AUPRC (Area Under the Precision-Recall Curve) when dealing with unbalanced data, where the emphasis on recall may outweigh precision [6]. The AUPRC provides insights into the trade-off between recall and precision, enabling an assessment of whether the model's accuracy is derived from correctly identifying positive cases (high recall) or simply predicting negative cases (higher precision). An optimal AUPRC score of 1 would indicate the accurate classification of each patient.

For both curves, the model's classification threshold is varied from 0 to 1, and the corresponding results are plotted on the graph. Although the specific threshold value may not always be directly adjustable in the classification model, this approach facilitates a comprehensive comparison among different classifiers. By observing the performance at various threshold values, it becomes possible to assess and compare the overall effectiveness of the classifiers.

For this Random Forest model, the obtained AUROC score of 0.882 indicates a robust overall prediction quality. However, as was discussed for the accuracy of the previous prediction model, this seemingly strong prediction performance could be attributed to the underrepresentation of positive cases and the large number of true negatives. This assumption is supported by the relatively modest AUPRC value of 0.528, which highlights the limited precision in identifying positive cases. These metrics collectively emphasize the importance of considering accuracy, precision, and recall together when evaluating the model's predictive capabilities.

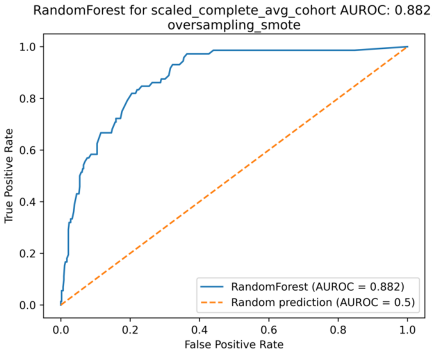


Figure 21:
AUROC of Random Forest with SMOTE

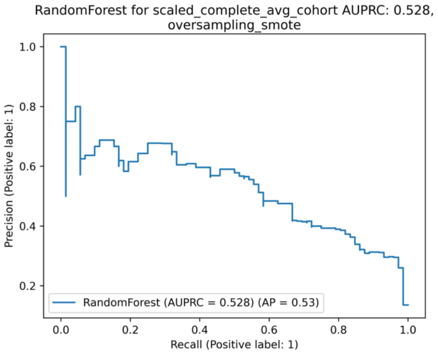


Figure 22:
AUPRC of Random Forest with SMOTE

On a side note, an additional advantage of using a Random Forest model can be the option to investigate feature relevance by examining the underlying Decision Trees. While this is not conducted in this particular study, it can be a valuable approach for alternative investigations into feature importance.

6.2.2 Optimizing Random Forest with GridSearchCV

As a next step, the Random Forest model was further enhanced by employing hyperparameter tuning. This approach involved iterating and varying the available parameters of the Random Forest algorithm. To accomplish this, GridSearchCV[22] was utilized. The algorithm systematically explores different parameter combinations to identify the optimal Random Forest model configuration that maximizes a selectable performance metric. For this specific model, the optimization of the recall metric was implemented, as it presents the most important indicator in this context.

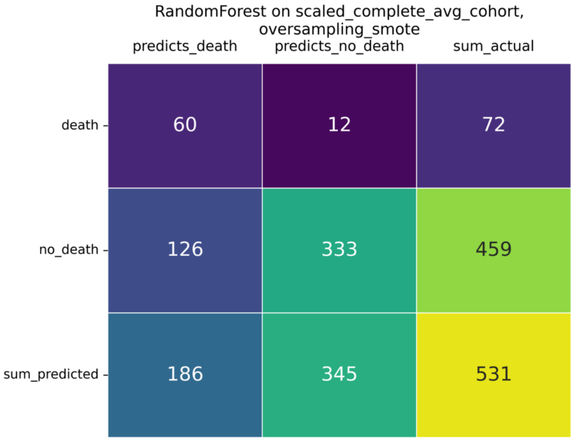
The optimization of recall leads to a significant improvement in the metric, increasing it to 0.83. However, this improvement comes at the expense of a decrease in precision, which drops to 0.32. Consequently, there is a notable decline in accuracy, with the value decreasing to 0.74.

[22] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

	precision	recall	f1-score	support
0	0.97	0.73	0.83	459
1	0.32	0.83	0.47	72
accuracy			0.74	531
macro avg	0.64	0.78	0.65	531
weighted avg	0.88	0.74	0.78	531

Table 11: Classification Report for Random Forest, optimized with GridSearchCV

When additionally examining the confusion matrix for this scenario, it becomes evident that the classifier predicts a total of 186 death cases, whereas the actual dataset contains only 72 instances of death. This indicates that the classifier tends to classify cases as death more frequently than the actual occurrence of death in the dataset. Hence, this classifier has a higher tendency to generate false positive predictions for death cases.



	predicts_death	predicts_no_death	sum_actual
death	60	12	72
no_death	126	333	459
sum_predicted	186	345	531

Figure 23: Confusion Matrix for Random Forest, optimized with GridSearchCV

Similarly, the following visualizations of AUROC and AUPRC show, that the overall model quality does not increase. The sole optimization of recall with GridSearchCV results in a lower AUROC score, as well as a lower AUPRC. This outcome highlights that solely focusing on recall does not yield a desirable model performance, either. Nevertheless, the GridSearchCV technique offers a powerful tool for further optimization and fine-tuning of Random Forest implementations.

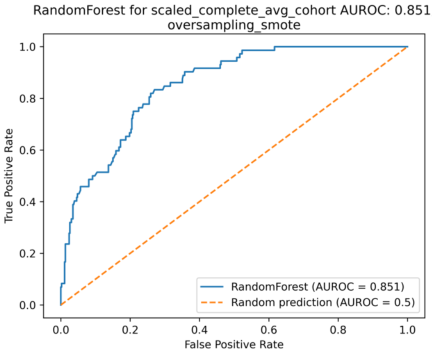


Figure 24:
AUROC of Random Forest, optimized with GridSearchCV

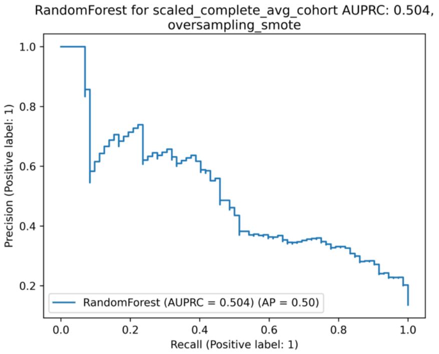


Figure 25:
AUPRC of Random Forest, optimized with GridSearchCV

6.2.3 XGBOOST Prediction Model

Another model that has shown promising results in various applications is eXtreme Gradient Boosting, which is also referred to as XGBOOST[23]. Similar to Random Forest, XGBOOST utilizes an ensemble of Decision Trees. However, what sets it apart is that each subsequent tree in the ensemble is built upon the classifications of the previous trees. Notably, XGBOOST was recommended by Vazquez et al. for predicting in-hospital mortality of patients with acute coronary syndrome, demonstrating its effectiveness in a related healthcare use case [7].

While the recall value is not as high as the GridSearchCV Random Forest option, the accuracy in this model did not sink. Overall, the model performance is similar to the oversampled Random Forest model, with an additional higher recall rate of 0.69.

	precision	recall	f1-score	support
0	0.95	0.86	0.90	459
1	0.44	0.69	0.54	72
accuracy			0.84	531
macro avg	0.69	0.78	0.72	531
weighted avg	0.88	0.84	0.85	531

Table 12: Classification Report for XGBOOST, balanced with SMOTENC

Similarly, the AUROC score for the XGBOOST model is comparable to that of the regular Random Forest model. However, the AUPRC score for XGBOOST is higher, at 0.59. This indicates a more desirable model performance.

[23] https://xgboost.readthedocs.io/en/stable/python/python_intro.html

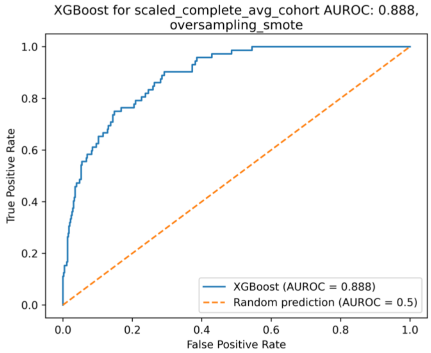


Figure 26: AUROC of XGBOOST

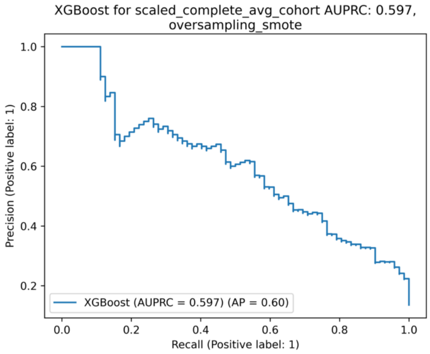


Figure 27: AUPRC of XGBOOST

Overall, the results of the XGBOOST model are promising. Still, there is potential for further optimizations and fine-tuning. For this use case, the model is chosen as the most reliable approach, based on machine learning methods.

6.2.4 Neural Network Prediction Model

Some research indicates that a deep learning solution might offer a better model performance, compared to common machine learning classifiers. Moreover, such an approach might be functional with less pre-processing of the data and could be based on time-series data. One suggested model was built with a network of gated recurrent units (GRUs) and for a comparable prediction task it resulted in an AUROC score of 0.874 and an AUPRC score of 0.471 [6]. It must be noted, that the results of neural networks vary with each execution, so the average over multiple results of the prediction metrics should be considered when evaluating these models.

To investigate this potential of deep learning models for the stroke use case, a "sequential"-model from the keras library[24] is implemented. The neural network was constructed with a "binary_crossentropy" loss function and trained using the "adam" optimizer to optimize accuracy. The selected network architecture consisted of a total of six layers. The first three layers comprised sixteen, twelve, and ten nodes, respectively, while the subsequent two layers contained eight nodes each. These inner nodes utilized rectified linear units, or "ReLU", as activation functions. The last layer consisted of a single node with a "sigmoid" activation function, producing a final output probability for mortality.

In terms of deep neural networks, the architecture presented here is considered relatively simple. However, it can be further enhanced through comprehensive network optimization, such as parameter tuning. Additionally, there are numerous other advanced architectures available, such as the mentioned gated recurrent units. The further development of more complex neural networks is omitted within this thesis. For one, these implementations might demand computation based on graphics cards (GPUs). However, the code of this thesis is intended to be usable with any hardware. Moreover, the goal of this comparison

[24] https://www.tensorflow.org/guide/keras/sequential_model

is primarily to demonstrate the potential of deep learning networks within a medical use case.

Finally, the training of the model was conducted for 175 epochs and the development of loss and accuracy is displayed in Figure 28. It can be seen that with each iteration of the model, the prediction quality inside the training dataset increases continuously. However, this progress quickly slows down and after approximately 150 epochs accuracy and loss stabilize. The final accuracy of the model, which results at 0.78, is based on the predictions of the test dataset.

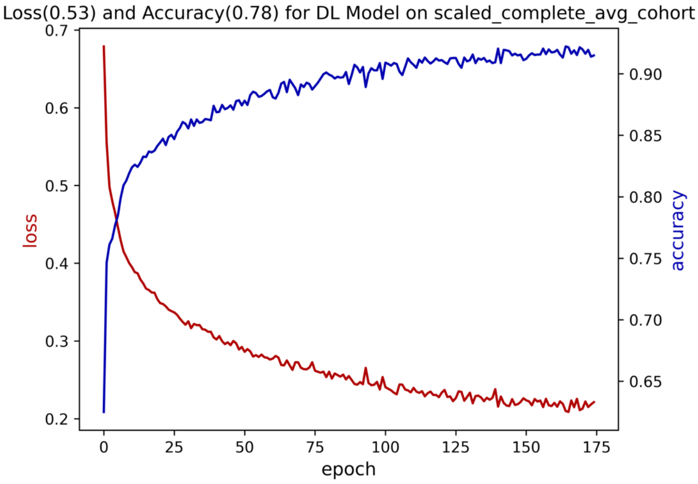


Figure 28: Neural Network Model Fit History

The following classification report summarizes the results of the neural network model. It yielded prediction results that are comparable to the previous approach of Random Forest with GridSearchCV.

	precision	recall	f1-score	support
0	0.97	0.77	0.86	459
1	0.37	0.83	0.51	72
accuracy			0.78	531
macro avg	0.67	0.80	0.68	531
weighted avg	0.89	0.78	0.81	531

Table 13: Classification Report for Neural Network, balanced with SMOTE

While the accuracy of the model appears slightly higher compared to the Random Forest alternative, it is important to again note the significant trade-off between recall and precision. This indicates, that the model tends to classify cases as “death” too frequently.

This results in a higher number of false positives, where patient mortality is predicted, when this may not be the real case. Therefore, achieving a better balance between recall and precision is crucial to improve the overall performance and reliability of the model. This should also have a positive effect on the relatively small AUPRC score, which is only at 0.329, as can be seen below in Figure 30. The following plots also illustrate that neural networks do not offer flexible threshold alteration. Instead, only a single threshold setting is available for these models.

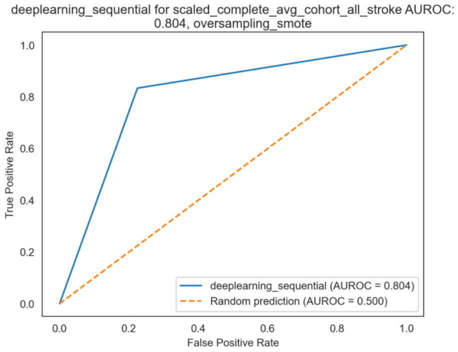


Figure 29: AUROC of Neural Network

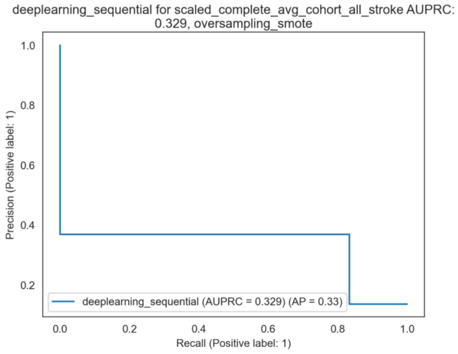


Figure 30: AUPRC of Neural Network

Moreover, it is worth noting that the model does not achieve similarly high scores as the examples reported in other research papers [6]. As previously mentioned, these shortcomings could potentially be attributed to the use of average data and the need for further optimization of the model.

6.2.5 Summary of Prediction Results

The following Table 14 provides a summary of the prediction results for each of the available model settings. It is important to note again that neural network models yield varying results with each implementation, which is why the metrics in the table may not align exactly with the analysis above. Despite not reaching as high scores, neural networks still show promise as an alternative approach to machine learning models.

classification_method	dependent_variable	auroc_score	auprc_score	accuracy	recall	precision
RandomForest	death_in_hosp	0.882	0.528	**0.853**	0.667	**0.471**
RandomForest	death_3_days	0.936	0.605	0.917	0.732	0.476
RandomForest	death_30_days	0.871	0.629	0.821	0.744	0.572
RandomForest	death_180_days	0.839	0.695	0.782	0.755	0.591
RandomForest	death_365_days	0.845	0.724	0.778	0.739	0.619
RandomForest_gridsearch	death_in_hosp	0.851	0.504	0.74	**0.833**	0.323
RandomForest_gridsearch	death_3_days	0.933	0.613	0.887	0.732	0.38
RandomForest_gridsearch	death_30_days	0.868	0.612	0.808	0.795	0.544
RandomForest_gridsearch	death_180_days	0.840	0.680	0.766	0.755	0.567
RandomForest_gridsearch	death_365_days	0.851	0.753	0.778	0.758	0.616
XGBoost	death_in_hosp	**0.888**	**0.597**	0.84	0.694	0.442
XGBoost	death_3_days	0.921	0.573	0.908	0.756	0.443
XGBoost	death_30_days	0.843	0.596	0.815	0.624	0.575
XGBoost	death_180_days	0.811	0.639	0.763	0.675	0.57
XGBoost	death_365_days	0.82	0.716	0.75	0.685	0.582
deeplearning_sequential	death_in_hosp	0.765	0.295	0.791	0.667	0.356
deeplearning_sequential	death_3_days	0.772	0.264	0.838	0.756	0.29
deeplearning_sequential	death_30_days	0.711	0.364	0.759	0.786	0.472
deeplearning_sequential	death_180_days	0.728	0.443	0.763	0.748	0.562
deeplearning_sequential	death_365_days	0.756	0.51	0.733	0.727	0.553

Table 14: Comparison of Model Results

Among the different dependent variables, the prediction results for "death within three days" are the most reliable. This finding suggests that the effects of stroke tend to be most severe within this timeframe. Predicting "death in hospital" may be less reliable as it also encompasses patients who stayed in the ICU for extended periods before passing away. It is also reasonable that the prediction for even longer time spans, such as 30 days or one year, becomes less reliable.

In conclusion, XGBOOST was selected as the most applicable model for the subsequent analysis, as it offers a sensible balance between recall and accuracy. The achieved results, although not yet optimal, provide valuable insights for further improvement. In comparison to the nomogram, proposed by Li and Li [10], the model can deliver similar and in parts better performance. The optimally desired thresholds of accuracy values above 0.9 and recall above 0.8 were nearly met for the "death_3_days" variable. However, they were not fully attained for the prediction of in-hospital mortality.

The overall suitability of the classification models presented in this study is subject to discussion. Despite the relatively high accuracy, it is important to note that recall rates below 0.8 are not desirable for mortality prediction [23]. Furthermore, the low precision suggests that the models often classify cases as potential deaths, leading to frequent false positives.

These shortcomings highlight the need for further improvements in the models' performance. As previously mentioned, it may be beneficial to refine feature selection and explore alternative pre-processing methods. Additionally, it is highly recommended to incorporate external validation of the developed model, as emphasized in previous research [5]. This suggestion has been duly acknowledged and in future studies, it can be explored whether the model can be applied to a different external data source for validation purposes. For example, MIMIC-VI[25], which is another dataset provided by PhysioNet, could be investigated for further validation.

6.2.6 Prediction Results for Reduced Feature Set

The same reduced set of features, as determined in the second, optimized clustering setup, was also employed for the classification models. In contrast to the clustering, the classification results did not improve when selecting fewer features. The impact of this feature reduction on the prediction results for "death in hospital" is summarized in the following table.

classification_method	**dependent_variable**	**auc_score**	**auc_prc_score**	**accuracy**	**recall**	**precision**
RandomForest	death_in_hosp	0.872	0.507	0.866	0.667	0.505
RandomForest_with_gridsearch	death_in_hosp	0.881	0.54	0.827	0.708	0.418
XGBoost	death_in_hosp	**0.894**	**0.612**	**0.868**	0.653	**0.511**
deeplearning_sequential	death_in_hosp	0.833	0.38	0.751	**0.847**	0.335

Table 15: Alternative Classification Results, Death in hospital

Upon comparing these classification reports, XGBOOST again emerges as the most promising option. Although the reduction of features did not result in an overall improvement in prediction quality, the XGBOOST model achieved a slightly more balanced trade-off between recall and precision compared to the previous model. Again, it is important to note that the GridSearchCV alternative and the deep learning network exhibit higher recall rates, albeit at the expense of significantly decreased precision. As this is not recommended, XGBOOST remains the most reliable choice for the task at hand.

Similar to the previous analysis, the dependent variable "death_3_days" offers an interesting use case, as it yields better overall results compared to "death in hospital". In terms of AUROC score and accuracy, most models exhibit higher values for the "death_3_days" case, indicating improved performance. However, it is crucial to consider the significant decrease in precision across all models, which is also reflected in the comparatively low AUPRC scores. This poses a challenge in selecting the most suitable model for this particular case, as even XGBOOST demonstrates notably low precision.

classification_method	**dependent_variable**	**auc_score**	**auc_prc_score**	**accuracy**	**recall**	**precision**
RandomForest	death_3_days	0.919	0.507	**0.908**	0.683	**0.438**
RandomForest_with_gridsearch	death_3_days	**0.93**	**0.591**	0.895	0.756	0.403
XGBoost	death_3_days	0.925	0.582	0.896	0.707	0.403
deeplearning_sequential	death_3_days	0.829	0.252	0.84	**0.829**	0.304

Table 16: Alternative Classification Results, Death within three Days

[25] https://physionet.org/content/mimiciv/2.2/

In summary, the reduction of features does not appear to have a clearly positive impact on the classification models. For XGBOOST, recall decreased slightly, while accuracy did rise. However, it is important to remember that the inclusion of too many features can complicate the interpretability and explainability of the classification results. Striking a balance between an optimal number of informative features and model interpretability is crucial. Overall, the XGBOOST model with fewer features appears to be slightly more balanced, so this setup is kept for the subsequent analysis.

6.3 Feature Relevance within XGBOOST

In this final section, an analysis of the feature relevance within the XGBOOST model is presented, concluding the chapter on mortality prediction.

When analyzing the results of machine learning models, it is crucial to engage in critical reflection. Particularly, because these models often function as "black boxes", making it challenging to trace the relationship between the input and the output. Understanding the influence of feature values is an initial and significant step toward comprehending prediction decisions and enhancing transparency.

Subsequently, the feature relevance within the XGBOOST model is investigated more closely with explanations based on Shapley values[26]. The SHAP (SHapley Additive exPlanations) methodology originally stems from a game theoretic approach and can be used to explain the output of machine learning models [25]. More precisely, Shapley values are local explanations of how feature values influenced the prediction decision for individual instances.

The following waterfall plots show, how the respective values per feature influenced the XGBOOST models' decision to classify a single patient.

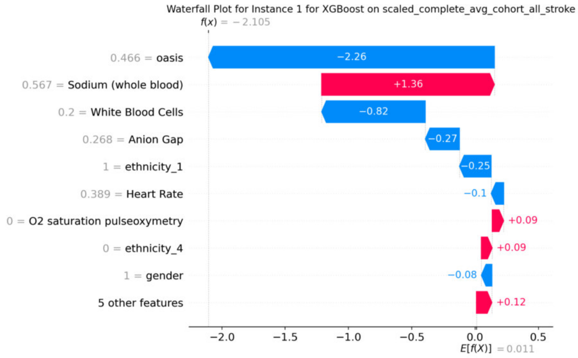


Figure 31: SHAP Waterfall for Instance One

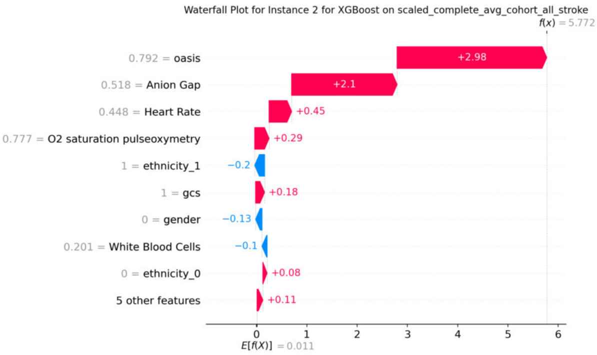


Figure 32: SHAP Waterfall for Instance Two

For the first instance, depicted in Figure 31, the overall Shapley value is negative, indicating a prediction of "no death". Similar to previous cases, categorical features such

[26] `https://shap.readthedocs.io/en/latest/index.html`

as ethnicity are one-hot encoded. It can be observed that ethnicity has minimal influence in this instance, aligning with its relatively low correlation with death. The continuous features are scaled within the range of 0 and 1. In this instance, the value of 0.466 for the scaled "oasis" feature, suggests that the patient has an Oasis score lower than the average patients. This corresponds to the negative Shapley value for the "oasis" feature in this particular case.

In contrast, instance number two, presented in Figure 32, exhibits an Oasis score of 0.792, which is significantly higher than the mean. Considering that "oasis" is positively correlated with death, the substantial positive influence of the Shapley value for this feature is logical. A positive influence in this context indicates that the patient is more likely to be predicted as belonging to the positive class. Therefore, the explanation for this feature is that a higher Oasis score indicates a greater risk of death. The same reasoning can be applied to the other features, and the sum of their Shapley values amounts to 5.772. This indicates that for this particular instance, the XGBOOST model predicts death with a high probability.

It should be noted that the explanation for an individual patient cannot be generalized to the entire dataset. Nevertheless, these explanations provide insights into the influence of certain features.

In addition, the subsequent figures can be used to facilitate an analysis of feature importance within the entire dataset.

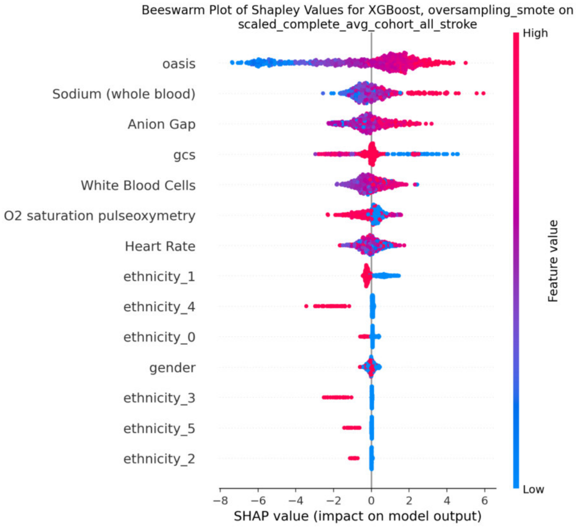


Figure 33: SHAP Beeswarm Plot

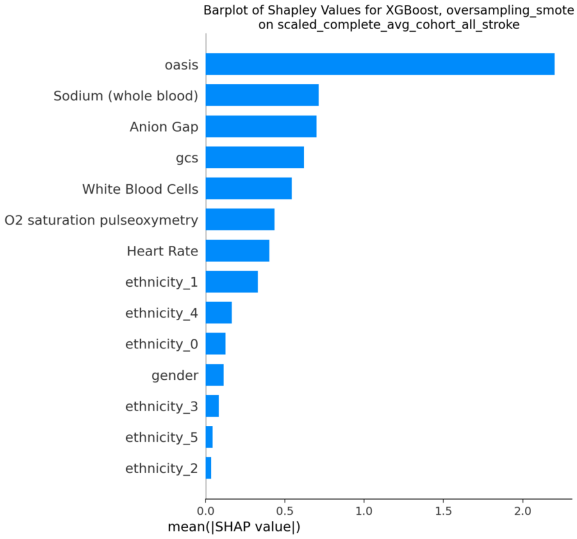


Figure 34: SHAP Barplot

A total of 150 instances were chosen for the beeswarm plot, with each instance represented by a dot reflecting its respective feature importance value. For example, in the case of the Oasis score, higher feature values, denoted by red dots, correspond to higher Shapley values, indicating a greater risk of death. A different pattern can be observed for the “gcs” feature, wherein lower feature values, indicated by blue dots, correspond to higher Shapley values. This suggests that lower Glasgow Coma Scale (GCS) scores are associated with an increased risk of death.

The accompanying bar chart presents the average feature influence across all selected instances. It provides an overview of the overall feature importance, with the top-ranked features being “oasis”, “Sodium (whole blood)”, and “Anion Gap”. These features appear to have the most significant relevance for the decisions of the XGBOOST classifier.

Lastly, a detailed examination can be performed for each respective feature. To illustrate this, the partial dependence plot below showcases the direct relationship between the Shapley values and the corresponding “oasis” values. One can observe a linear correlation that gradually diminishes at the lower and upper extremes. Once more, this indicates that higher oasis scores are associated with higher and more influential Shapley values, further highlighting the importance of the “oasis” feature in predicting mortality.

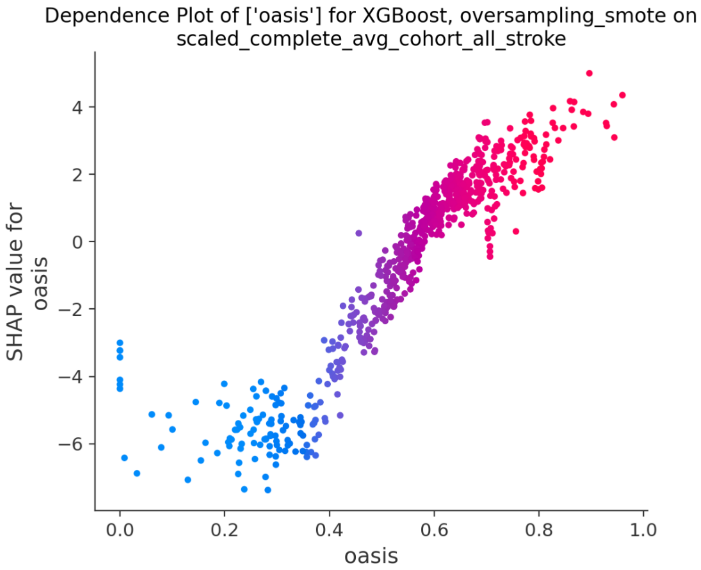


Figure 35: SHAP Partial Dependence Plot for Oasis

The Oasis score was selected as the main example in this chapter, as the strong correlation and importance of this feature have already been accounted for. Exploring the influence of other features on the model can offer additional valuable insights. It is noteworthy that the SHAP explanations presented here align with the overall correlation findings discussed in the previous chapter for most of the features. This consistency further enhances the reliability of the explanations and, consequently, the XGBOOST model itself. Such findings are helpful to increase trust in the prediction model.

On an additional note, it is advisable to conduct a thorough re-evaluation of any potential model explanations. This step is crucial as these explanations can sometimes be misleading. There are certain requirements that need to be met in order for the explanations to be considered valid and reliable [26]. For instance, “stability” indicates that instances with similar predictions should also have similar explanations. “Robustness” ensures that explanations are generally unaffected by small input perturbations. Another requirement is “consistency”, which states that local explanations for the same data points should be consistent across different prediction models. For the scope of this thesis, such detailed evaluations of explanations are not implemented. However, this remains an important area for future research.

7 Fairness Analysis

Based on the presented prediction results of the XGBOOST model, this chapter conducts a fairness analysis. First, the importance of fairness in prediction models is discussed and common metrics to evaluate fairness are introduced. Based on this, three different privileged groups are selected and investigated for diverging model performance.

7.1 Selection of Fairness Metrics

The significance of dependable data analysis in medicine has been increasingly recognized, particularly in light of the development of novel drugs. In the past, the efficacy tests of these drugs were often conducted with a disproportionate number of male participants, resulting in suboptimal medical treatment for female patients. As this issue gained greater public attention, there has been a growing emphasis on equal gender representation in clinical trials, and the field of gender medicine has emerged [27]. Analogously, in the medical domain, the utility of a prediction model would be compromised if it exhibited a consistent tendency towards inferior predictive accuracy for specific subgroups.

The most relevant risk that is investigated in the following analysis, is founded on the unequal size of subgroups. This so-called representational harm might be attributed to the representation bias in the selected dataset [28].

An initial solution for fairness problems within a dataset might be to remove potentially privileged attributes, like gender, ethnicity, caste, or religion, from the data. These attributes split a population into subgroups, which have historically faced systematic discrimination. This procedure is also known as “fairness-through-unawareness”. However, this approach is not recommended, as these attributes often correlate with other attributes. For example, ethnicity is often closely correlated to postal code. Thus the bias might simply be shifted towards a different feature.

A better approach is evaluating a model’s quality by closely analyzing fairness metrics. This approach is most common for applications with natural language processing [29], where social biases are easily incorporated into language structures. However, these metrics can also present a valuable tool for general data analysis, to reflect if certain subgroups exhibit differentiating prediction quality. This enhances the perspective of data analytics to not solely focus on predictive performance, but also incorporate social and ethical aspects of fairness. The focus on equal prediction quality is especially important when the model is implemented in a social domain, like this medical use case, where equality is imperative.

A fairness analysis can focus on group fairness, defining that the outcome for each subgroup of privileged attributes should have parity. The other aspect of fairness is individual fairness, which states, that similar individuals should have similar prediction outcomes and treatments [30].

The following fairness metrics can be used to incorporate these perspectives into the analysis. With these metrics, the fairness of different models can be compared on a statistical basis.

7.1.1 Metrics for Group Fairness

Accuracy Parity[27] presents a crucial metric to consider when evaluating the prediction quality for different subgroups. It requires that prediction accuracy remains consistent between privileged and unprivileged groups. The probability p of predictions $Y \in \{0, 1\}$ shall remain the same, regardless of the value of the sensitive attribute A. Dependent on this binary attribute, the unprivileged group is designated by 0, and privileged by 1.

$$p(\hat{\mathrm{Y}} = Y|A = 0) = p(\hat{\mathrm{Y}} = Y|A = 1) \tag{6}$$

The actual indicator for such fairness requirements can be calculated in two ways. The selected option for this thesis is the *parity difference* of the probabilities, for which the ideal value is 0.

$$metric\_difference = p(\hat{\mathrm{Y}} = Y|A = 0) - p(\hat{\mathrm{Y}} = Y|A = 1), \tag{7}$$

Alternatively, the *impact ratio*, which describes the proportion between the probabilities, should be close to 1.

$$metric\_ratio = \frac{p(\hat{\mathrm{Y}} = Y|A = 0)}{p(\hat{\mathrm{Y}} = Y|A = 1)}, \tag{8}$$

It is crucial to acknowledge that Accuracy Parity alone may not capture the complete picture of fairness in predictive models. In the context of stroke prediction, where identifying death cases is critical, it is necessary to evaluate the distribution of false positives and false negatives as well. To address this, metrics such as **Recall Parity** and **Precision Parity** can provide valuable insights. Recall Parity focuses on the true positive rate across different privileged groups, ensuring that the model's ability to detect positive cases, such as deaths, remains consistent irrespective of the privileged attribute. Precision Parity, which is also known as the predictive value metric, examines the positive predictive value across privileged groups. It assesses whether the precision of the model is dependent on the privileged attribute.

Another common criterion for ensuring fairness is the concept of **Demographic Parity**, which emphasizes the need for the positive predictions, $\hat{Y}$*=1*, to be independent of the privileged attribute A [31]. While Demographic Parity offers a more lenient form of fairness, it holds particular significance in the context of stroke prediction.

On a side note, this metric can also be measured with two different indicators. The Demographic Parity Difference, also known as Statistical Parity Difference, calculates the difference in prediction probabilities across subgroups, while the Disparate Impact Ratio compares the proportions between prediction probabilities.

[27] `https://afraenkel.github.io/fairness-book/content/05-parity-measures.html`

By considering these parity metrics, one can gain an initial understanding of the group fairness of a predictive model. The metrics provide insights into potential disparities and help assess if the model is behaving consistently across different subgroups.

However, this strict enforcement of group fairness might be unfair on an individual level. For example, if the model does not correctly predict a death case for a patient, because he belongs to a certain ethnicity for which the algorithm has to keep Demographic Parity. Hence, additional metrics for individual fairness are required.

7.1.2 Metrics for Individual Fairness

Metrics for individual fairness mainly focus on equal prediction rates for known positive cases. This ensures that individuals, who should be identified as positive cases (Y=*1*), are correctly predicted as such ($\hat{Y}$=*1*), regardless of their subgroup A. The decisive difference to the previous metrics is that one knows that the patient belongs to the positive cases, and tests if the correct prediction is made, independent of the subgroups.

The **Equalized Odds** metric serves as a rigorous metric for individual fairness. It requires both the true positive rate and the false positive rate to be equal across privileged attributes. The fulfillment of Equalized Odds indicates a highly fair prediction model, which is also referred to as high-level fairness [32].

$$p(\hat{\mathrm{Y}} = 1|A = 0, Y = y) = p(\hat{\mathrm{Y}} = 1|A = 1, Y = y), y \in 0, 1 \tag{9}$$

In addition, the principle of **Equal Opportunity** represents a less strict version, in a similar way Demographic Parity simplifies Accuracy Parity. It demands that only the true positive rate is equal across the privileged attributes. This metric has high relevance for the stroke use case, where the detection of actual deaths has priority [32].

$$p(\hat{\mathrm{Y}} = 1|A = 0, Y = 1) = p(\hat{\mathrm{Y}} = 1|A = 1, Y = 1) \tag{10}$$

7.1.3 Further Fairness Metrics

Generalized Entropy[28] presents a comprehensive approach that combines both individual and group fairness measures [33].

Initially, the concept of Generalized Entropy was used as an inequality index in an economic context. Low entropy describes a setting where there is little randomness, which is considered a more equal distribution. This concept can be adapted as a fairness metric, where the Generalized Entropy assesses the equal distribution of accurate predictions within each subgroup. In other words, it can be seen as the entropy, or randomness, of predictions, which should be minimal. The metric incorporates the parameter α, which determines the weight assigned to the prediction disparities. When

[28] `https://aif360.readthedocs.io/en/stable/modules/generated/aif360.metrics.ClassificationMetric.html#aif360.metrics.ClassificationMetric.generalized_entropy_index`

α is set to 1, the metric is commonly referred to as Theil-Index. The resulting entropy value ranges from 0 to 0.5, with values closer to 0 indicating a fairer model.

The following base formula calculates the utility or benefit b, for each individual i, as the difference between the predicted label $\hat{y}$ and the real occurrence y. It may be noted, that in the context of stroke, it is not correct to interpret the positive cases, which represent mortality, as inherently beneficial.

$$b_i = \hat{y}_i - y_i + 1 \tag{11}$$

Based on these benefit scores, the Generalized Entropy $\mathcal{E}$ of the model can be calculated according to three different formulas. The selection of the respective formula depends on the selected α value. In the subsequent analysis, the Theil-Index, where α is set to 1, is employed for the entropy metric. Furthermore, the parameter μ describes the mean of the benefit scores b.

$$\mathcal{E}(\alpha) = \begin{cases} \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left[\left(\frac{b_i}{\mu} \right)^{\alpha} - 1 \right], & \alpha \neq 0, 1, \\ \frac{1}{n} \sum_{i=1}^{n} \frac{b_i}{\mu} \ln \frac{b_i}{\mu}, & \alpha = 1, \\ -\frac{1}{n} \sum_{i=1}^{n} \ln \frac{b_i}{\mu}, & \alpha = 0. \end{cases} \tag{12}$$

The calculation of the entropy can then be conducted for each group and the difference is used to compare the between-group fairness.

These introduced fairness metrics can exhibit a trade-off between group fairness and individual fairness. Striving for higher group fairness could potentially result in lower individual fairness, and vice versa. Therefore, the choice of a suitable fairness metric is context-dependent and necessitates a thoughtful examination of the ethical and social implications specific to the application at hand.

7.2 Fairness Analysis of the XGBOOST Model

The fairlearn[29] library was implemented to calculate fairness metrics for the XGBOOST model. The library was also helpful to create visualizations of the differences between the subgroups.

Another widely recognized and comprehensive alternative for fairness analysis is the AIF360[30] library. However, during the evaluation process, it was observed that the calculations for certain fairness metrics were not consistently reliable. Upon closer examination, it was found that the division into privileged and unprivileged groups was prone to errors. Specifically, when multiple attributes were selected as privileged, the definition of the unprivileged group became overly strict. Only individuals who exactly matched the opposite of the selected privileged groups were considered as part of the unprivileged group. For example, if "asian" and "female" were chosen as privileged attributes, the resulting unprivileged group would consist only of individuals who were "non-asian" *AND* "male". However, the desired definition for the unprivileged group should include individuals who are either "non-asian" *OR* "male". As a quick solution to this issue was not readily available, the fairlearn package was selected due to its more user-friendly setup.

The following three examples illustrate the process of conducting a fairness analysis based on manually selected features. Specifically, the focus is on the features "gender" and "ethnicity", which are expected to exhibit varying prediction performances across different subgroups. These features are known to significantly divide groups and potentially impact the overall fairness of the model [27]. Moreover, there is compelling evidence suggesting differences in medical indicators and mortality rates among various gender and ethnic subgroups for stroke [11] [12]. Similarly, research focusing on heart illnesses detected differences between gender subgroups [7].

Following such a fairness analysis, it can become necessary to adjust a model to address fairness concerns. The actual realization of reworking models to enhance fairness is not included in this thesis. However, it presents a promising, alternative area for further research.

7.2.1 Sensitive Feature: Ethnicity

The feature that is selected for the first analysis of privileged groups is ethnicity. It can also be referred to as the "sensitive feature". Following the one-hot encoding of categorical features, the dividing feature is binary, with the class "1" representing the privileged group. Based on the previously introduced unequal distribution of ethnicities, the occurrence of "white" is considered "privileged" in the following step.

The following Figure 36 shows the distribution and performance metrics for each subgroup. The barplots enable easy visual comparison of these indicators, while the numerical

[29] `https://fairlearn.org/v0.8/user_guide/`

[30] `https://aif360.readthedocs.io/en/stable/index.html`

results are summarized in Table 17 next to it. For each of the subplots, the bar on the right represents the privileged class “1”, while the left bar displays the respective metric for all the remaining instances.

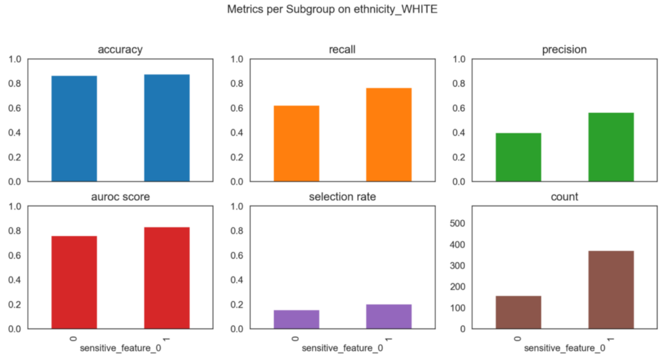


Figure 36: Metrics per Subgroup: White Ethnicity

Metrics	unprivileged	privileged	overall
accuracy	0.868	0.876	0.874
recall	0.625	0.768	0.736
precision	0.4	0.566	0.525
auroc score	0.76	0.832	0.816
selection rate	0.175	0.204	0.190
count	159	372	531

Table 17: Metrics per Subgroup: White Ethnicity

It can be seen, that the divergence in accuracy between the groups is minimal. However, notable differences can be observed in terms of recall and precision. The recall rate is significantly higher for the privileged class, “white”, at 0.768. This discrepancy is also reflected in the AUROC score, which is titled “roc_auc”. This indicates that the overall performance of the model is superior for the privileged class.

The selection rate, signifying the mortality rate, is slightly higher for the privileged class, at 20.5%. It must be noted, that this mortality rate in the test data is higher than the mortality rate in the original complete dataset because of the SMOTE oversampling method.

Finally, the count of the privileged class is noticeably higher, with 372 to 159 cases. This disparity in representation could potentially explain why the model performs better for white patients, as it has more data and a better understanding of this particular subgroup. It appears that other ethnicities are underrepresented in the MIMIC-III dataset when it comes to the stroke use case.

In addition to the performance metrics above, Table 18 displays the fairness metrics for the XGBOOST model. Consistent with the previous observations, the four parity metrics describing group fairness are lower for the unprivileged class.

This disparity is also evident in both fairness metrics for individual fairness. Notably, the high absolute value of 0.142 for the strict Equalized Odds metric, further strengthens the assumption of unfairness in the model.

At last, the entropy difference metric is computed by subtracting the Generalized Entropy of the privileged class from that of the unprivileged class. For the entropy of the predictions, a minimal value is desirable, so this metric has to be interpreted inversely to the previous metrics. Negative values of the difference metric imply a lower entropy within the unprivileged class, which would suggest better model performance for class “0”.

Metrics	Value
Accuracy Parity	-0.008
Recall Parity	-0.143
Precision Parity	-0.166
Demographic Parity	-0.047
Equalized Odds	-0.142
Equal Opportunity	-0.143
Entropy Difference	-0.057

Table 18: Fairness Metrics: White Ethnicity

Thus, in this case, the entropy difference is the only metric that does not indicate unfair model performance towards the unprivileged class. The value of -0.057 implies slightly lower entropy within the unprivileged class. However, the overall impression of this fairness analysis implies the existence of unfairness towards non-white patients.

The previous analysis primarily concentrated on the “1-vs-all” comparison between “white” ethnicity and the rest of the dataset. An alternative approach is to conduct a “1-vs-1” comparison for each specific ethnic subgroup. This involves selecting each ethnicity as the privileged class and examining the relevant performance metrics across all ethnic subgroups individually. By employing this, a more granular analysis can be performed to assess the disparities in predictive performance across all ethnicities. The results for this approach are displayed in Table 19.

However, an inherent challenge with the chosen stroke use case becomes evident. Within the test data, the number of instances for most of the ethnicity subgroups is insufficient to calculate the respective fairness metrics. The majority of these subgroups have a support size of less than 100. Consequently, a direct comparison of performance across ethnicities is not feasible. Nonetheless, this limitation strongly supports the assumption of representational bias within the dataset.

ethnicity	accuracy	recall	precision	auroc_score	size	reliable
OVERALL	0.874	0.736	0.525	0.832	531	yes
WHITE	0.876	0.768	0.566	0.832	372	yes
ASIAN	0.882	0.5	0.5	0.717	21	no
HISPANIC OR LATINO	0.842	0	0	0.444	19	no
BLACK	0.904	0	0	0.48	52	no
OTHER	0.96	1	0.667	0.978	25	no
UNKNOWN/NOT SPECIFIED	0.783	0.75	0.429	0.77	46	no

Table 19: Comparison of XGBOOST Performance across Ethnicity

7.2.2 Sensitive Feature: Gender

Next, the feature "gender" is selected with the attribute "male" as the assumed "privileged" class. This selection is based on suspected gender inequality within medical research [12].

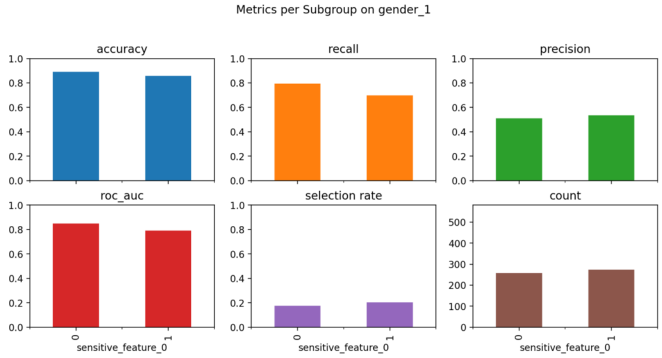


Figure 37: Metrics per Subgroup: Male Gender

Metrics	unprivileged	privileged	overall
accuracy	0.891	0.858	0.874
recall	0.793	0.698	0.736
precision	0.511	0.536	0.525
auroc score	0.848	0.793	0.816
selection rate	0.175	0.204	0.190
count	257	274	531

Table 20: Metrics per Subgroup: Male Gender

Contrary to expected results, accuracy and recall are higher for the unprivileged class. The same is true for the AUROC score. These metrics do not support the assumption, that male patients might be privileged and have better prediction results.

Furthermore, based on the selection rate, male patients have a slightly higher occurrence of death in the test data. At the same time, the prediction quality for these patients is not as high. This reveals, that the model does not predict male cases as well, even though it has higher support for these. Once again, it must be noted that the selection rates displayed above were impacted by the oversampling method and the separation into training and test data. Thus, they are not representative of the actual mortality rates, which are displayed in the descriptive data analysis, in Chapter 4. However, a higher prediction quality for female patients can still be detected when not implementing the sampling method.

Lastly, the sizes of each class are relatively similar, which suggests a reduced likelihood of representational bias.

These indicators can again be supported by the fairness metrics displayed in Table 21.

As mentioned before, it is surprising to observe a higher recall rate for female patients, as indicated by the Recall Parity of 9.5%. Demographic Parity is slightly lower for female patients but Equalized Odds and Equal Opportunity are again higher for the "unprivileged" group. Entropy within the female class is also lower, which further indicates a worse prediction quality for male patients.

While the observed discrepancies between the two groups are not remarkably high, there are indeed noticeable differences. Conducting a more comprehensive analysis of feature relevance for each subgroup could provide valuable insights into the underlying factors influencing the model's decision-making process for male and female cases.

Metrics	Value
Accuracy Parity	0.033
Recall Parity	0.095
Precision Parity	-0.025
Demographic Parity	-0.029
Equalized Odds	0.079
Equal Opportunity	0.095
Entropy Difference	-0.051

Table 21: Fairness Metrics: Male Gender

One plausible explanation can be found in the subsequent chapter on subgroup analysis. In this chapter, a detailed examination of clustering results reveals a noteworthy subgroup consisting of exclusively female, white patients. It is observed that all of these patients originate from the "CareVue" system and lack O2 measurements or Oasis scores. This pattern of missing crucial features could potentially account for the higher-than-average detection of mortality cases among this group. The manner in which patients with missing data are handled can significantly impact the prediction model.

When excluding the "CareVue" patients, a different pattern of unfairness can be observed, aligning more closely with the anticipated results.

Metrics	unprivileged	privileged	overall
accuracy	0.840	0.792	0.826
recall	0.708	0.375	0.625
precision	0.447	0.231	0.392
auroc score	0.785	0.609	0.740
selection rate	0.217	0.181	0.206
count	175	72	247

Table 22: Metrics per Subgroup: Female Gender, Metavision

Metrics	unprivileged	privileged	overall
accuracy	0.832	0.816	0.826
recall	0.529	0.733	0.625
precision	0.346	0.440	0.392
auroc score	0.700	0.782	0.740
selection rate	0.174	0.255	0.206
count	149	98	247

Table 23: Metrics per Subgroup: Male Gender, Metavision

The comparison between subgroups reveals, that within the "metavision" dataset male patients exhibit significantly higher recall and precision compared to female patients. In the left table, which focuses on female patients as the privileged class with a count of 72 cases, their recall is notably low at 0.375. On the other hand, the right table displays the male patients as the privileged class, showcasing a much higher recall of 0.733. These findings indicate unfairness regarding female patients in the "metavision" database system, as was initially suspected.

Understanding the impact of missing values or the varying influence of features between these two database systems is crucial in order to gain deeper insights into the suitability of the databases within the MIMIC-III dataset. Thus, these substantial differences warrant further investigation.

7.2.3 Sensitive Features: Gender and Ethnicity

For this last analysis, the distinguishing features are a combination of gender and ethnicity. More precisely, class "1" encompasses the intersection of "female" and "white" patients. Based on the previous results this privileged class is supsected to show up most pronounced unfairness. Combined privileged classes, that are made up of multiple features, are also referred to as "patterns".

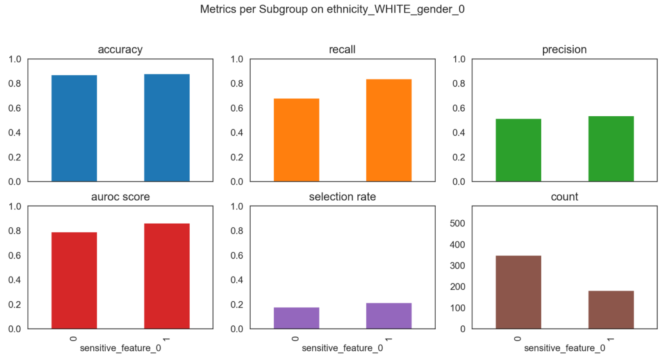


Figure 38: Metrics per Subgroup: Ethnicity and Female Gender

Metrics	unprivileged	privileged	overall
accuracy	0.871	0.879	0.874
recall	0.681	0.840	0.736
precision	0.516	0.538	0.525
auroc score	0.791	0.863	0.816
selection rate	0.178	0.214	0.190
count	349	182	531

Table 24: Metrics per Subgroup: Ethnicity and Female Gender

The count of this class is much lower, with 182 instances, than for the previous selections. Still, with 0.879, the accuracy for this privileged class is comparable to the "white" ethnicity group and only slightly lower than for the sole "female" group which was at 0.891. Furthermore, the metrics recall, precision, and AUROC score are each higher for this case, clearly supporting suspected differences in the prediction quality between classes.

Again, the fairness metrics in Table 25 can be compared between the combined subgroup “female” and “white” with the rest of the dataset.

Similar to the performance metrics, there is an evident increase in the disparities of recall, as well as Equalized Odds and Equal Opportunity. This highlights a more pronounced unfairness when examining the variables “gender” and “ethnicity”.

Metrics	Value
Accuracy Parity	-0.008
Recall Parity	-0.159
Precision Parity	-0.022
Demographic Parity	-0.037
Equalized Odds	-0.174
Equal Opportunity	-0.159
Generalized Entropy	-0.003

Table 25: Fairness Metrics: Ethnicity and Female Gender

However, Precision Parity and Demographic Parity exhibit a relatively lesser degree of imbalance, compared to the analysis where only white patients are considered. As a possible explanation, the smaller precision value of the privileged class can be attributed to the lower precision rate observed within the group consisting solely of female patients, which stands at 0.511. Nevertheless, in the context of this particular use case, the precision difference is not as crucial as the increase in the difference of recall. Lastly, it can be seen that the entropy values are relatively similar between both subgroups.

Overall, the combination of the privileged groups does further increase differences in recall and general model imbalance.

7.2.4 Fairness Analysis Summary

In summation, this comprehensive fairness analysis of the XGBOOST model did detect potential instances of unfairness.

The first selection based on ethnicity did indicate representational bias, as the fairness metrics differed between privileged and unprivileged classes. This problem was anticipated, as certain subgroups are underrepresented, as was seen in Chapter 4. This does not imply, that the real treatment and survival chances of patients from these subgroups are actually unequal. However, it does indicate that the XGBOOST model does not classify some patients with the same quality. Next, the fairness analysis based on gender led to unexpected results. Even though gender inequalities are common within medical research the analysis suggested better results for female patients. While this might be attributed to the explained missing values within the carevue dataset, a further investigation of relevant features and the decision process of the model is required. This affirms the importance of better explainability of prediction models. At last, a combined privileged class of “female” and “white” patients resulted in the highest unfairness, yet, with a difference in recall rates of 0.159.

In addition to this analysis, it is possible to compare different prediction models regarding their fairness in the supplemented frontend.

On the one hand, the manual selection of privileged classes implemented in this chapter provided valuable insights. However, it is important to acknowledge that it was not possible to analyze some potential subgroups, such as black women, due to the limited size of the dataset. In detail, after splitting the data into training and test sets, there are only 531 patients available. Among these, there are only 32 black women. Within this small subset, there were no predicted death cases, which led to no true positives. As previously mentioned, this renders the calculation of fairness metrics impossible. This subgroup would have been particularly relevant for research, due to their known high mortality rate in the context of stroke. For example, some studies state that black women may have 47% higher risk of stroke than white women [12]. A more focused analysis of this group, and other disadvantaged groups, can be very insightful.

This limitation poses a significant challenge when investigating specific groups within the MIMIC-III dataset. One potential solution could be to allocate a relatively larger portion of the data to the test set when splitting the data into training and test sets. However, this approach can quickly lead to reduced prediction quality. Another possible solution is to increase the overall number of patients in the dataset. This may necessitate a switch to a different use case, such as heart attacks or general mortality, to achieve a more substantial sample size. In future studies, including such subgroups would be advantageous to gain a more comprehensive understanding of fairness and performance across diverse populations.

Another shortcoming of manual detection is that it requires previous context-related knowledge. One needs to have medical knowledge or is reliant on other research to select potential subgroups.

Lastly, the inclusion of all available features leads to a very large amount of possible combinations of privileged classes. Comparing the performance metrics for all of these classes may not be feasible. Thus, not all potential patterns of privileged groups can be analyzed manually.

These reasons highlight the necessity for a process, that is less dependent on manual selection. Therefore, the exploration of an automated process to detect potential subgroups presents a promising alternative. This approach is tested and evaluated in the following chapter.

8 Subgroup Analysis

This final chapter presents an approach to compare subgroups for diverging prediction fairness and potential differences of feature relevance.

8.1 Automated Subgroup Detection

For this analysis step, the previously presented topics come together. The main idea is to use a clustering method to find subgroups inside the dataset, that show up a high difference from each other. Once a clustering algorithm is selected, the differentiation of these subgroups can be done automatically based on the concept of feature entropy. However, these subgroups might not have a clear medical or demographic explanation and are not always clearly explainable, as their differentiation might be based on multiple features. Thus the selection of relevant subgroups, especially those with a high mortality rate, still remains a manual task. The analysis of such subgroups can be insightful to evaluate the fairness and quality of a prediction model. If potential unfairness can be detected, previously hidden and unsuspected privileged groups may become apparent. This increased explainability can further support transparency and trust in the model.

A promising approach to automatically distinguish clusters is based on feature entropy. It was used for the development of the FairVis tool, which is intended to visualize the bias across these clusters [34]. The concept of feature entropy, which is implemented for this approach, has to be differentiated from the Generalized Entropy metric, from section 7.1.3.

Cabrera et al. derive the importance of each feature based on the feature entropy. If a feature in a subgroup is dominated by a certain value, e.g. all cases are “male”, then the value “male” dominates the feature “gender”. The feature would show up with low entropy and would thus be a dominant feature with high influence on this cluster [34]. They calculate the feature entropy S for subgroup, or cluster, k, with the features i and values v as follows:

$$S_{k,i} = -\sum_{v \in V_i} \frac{N_{k,v}}{N_k} log \frac{N_{k,v}}{N_k} \tag{13}$$

This formula specifies, that for each occurrence of a value, the influence upon the respective feature is calculated. The sum of these value influences multiplied by their logarithm makes up the entropy of each feature inside the respective cluster.

This concept was extended by Schäfer and Wiese with a normalization factor [20]. This is necessary to compensate for features that differ in the number of their available values.

$$H_{i,j} = -\frac{1}{log_2|Dom(A_j)|} \cdot \sum_{v \in Dom(A_j)} \frac{N_{i,j,v}}{N_i} \cdot log_2 \frac{N_{i,j,v}}{N_i} \tag{14}$$

The naming convention was slightly adjusted to include the clusters in the index. Here i is the index for the clusters, j is the index for features and v are the available values. Moreover, the feature entropy is now denoted by *H*. The variable *Dom(Aj)* is comparable to *Vi*. The crucial part of this second formula is the normalization factor in front of the summation. With this factor, the entropy range is shifted between 0 and 1 for each feature, independent of the number of available values. Thus feature entropy can be compared between all features across all clusters.

One might further decide if a ranking of features is used to find subgroups, or if a certain entropy threshold has to be crossed, for the feature to be considered dominant. However, for the stroke use case, the average of all available feature entropies is calculated within each respective cluster. This metric is also referred to as cluster entropy. Based on this, the diversity within each cluster can be approximated.

A ranking of the cluster entropy is then used to find clusters that show up with lower overall entropy. As clusters with lower entropy than the complete dataset indicate a more homogeneous group of patients, these clusters can be interpreted as actual subgroups, which are divergent from each other.

In addition, it is important to consider the mortality rate per cluster. If a cluster was dominated by a certain patient subgroup but had no death cases, the medical relevance might not be high. Only cases with a mortality rate at least as high as the complete dataset may be considered relevant.

8.2 Initial Subgroup Analysis Results

The following analysis is conducted based on the previously introduced feature selection of nine features.

As a preliminary step, for the kMeans clustering the silhouette scores are calculated and displayed in Figure 39. Ideally, a cluster count that maximizes the silhouette score should be selected. However, it is also advised not to choose a too-high cluster count, to ensure interpretability and to prevent excessively small cluster sizes.

Nevertheless, in this exemplary analysis, a relatively high cluster count of thirteen was chosen. While the silhouette score of 0.46 is not the highest achievable, it surpasses the scores obtained with a selection of ten or fewer clusters. Moreover, opting for even more clusters would constrain interpretability without significant improvement in the score. Additionally, the examination of cluster entropy in Table 26 reveals promising results for clusters eight and nine, further supporting the choice of thirteen clusters.

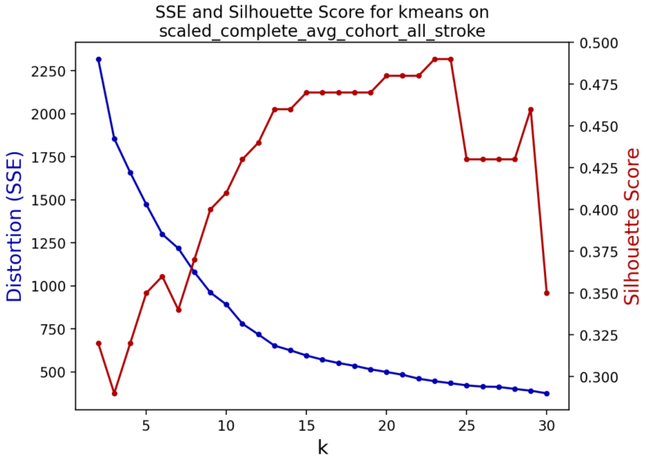


Figure 39: kMeans Silhouette Score

Table 26 presents the average feature entropy of the resulting clusters. Five clusters can be found, for which the average feature entropy is smaller than that of the complete dataset.

As previously explained, clusters with a higher mortality rate compared to the overall dataset, which is at 15%, are considered particularly relevant. In this regard, cluster number eight stands out, as it exhibits a mortality rate of 100%. This shows that all patients within cluster eight are actual death cases, making it a highly significant and distinct subgroup.

Clusters	**Count**	**death_in_hosp_rate**	**Average Entropy**
cluster_0	372	0.00	0.25
cluster_4	433	0.00	0.25
cluster_6	393	0.00	0.26
cluster_8	**135**	**1.00**	**0.27**
cluster_2	389	0.00	0.28
cluster_12	109	0.08	0.30
complete_set	**2655**	**0.15**	**0.30**
cluster_1	143	0.09	0.31
cluster_5	88	0.17	0.31
cluster_10	143	0.27	0.32
cluster_7	78	0.23	0.34
cluster_9	124	1.00	0.34
cluster_3	136	0.23	0.35
cluster_11	112	0.12	0.37

Table 26: Cluster Entropy Overview

The subsequent Table 27 provides a detailed examination of cluster eight. It is evident that this subgroup consists of 135 female, white patients exclusively. Moreover, all of these patients originate from the "CareVue" database, which is the reason for the absence of values for "O2 saturation pulseoxymetry" and "oasis". This distinctive pattern of missing values likely contributed to the clustering. These patients may also make up the subgroup discussed in the fairness analysis regarding gender inequality, in section 7.2.2. On a side note, it can be seen that the oversampling algorithm plays a role in this analysis, as the original "CareVue" system only contains 88 female, white patients.

Features	**Values**	**complete_set _count**	**cluster_8 _count**	**cluster_8 _entropy**
total_count	icustay_ids	2655	135	0
ethnicity	UNKNOWN/ NOT SPECIFIED	279	0	0
ethnicity	**WHITE**	**1846**	**135**	**0**
ethnicity	ASIAN	78	0	0
ethnicity	HISPANIC OR LATINO	112	0	0
ethnicity	BLACK	252	0	0
ethnicity	OTHER	88	0	0
gender	**0**	**1353**	**135**	**0**
gender	1	1302	0	0
dbsource	carevue	1415	135	0
dbsource	both	8	0	0
dbsource	metavision	1232	0	0
O2 saturation pulseoxymetry	(-0.001, 0.33]	3	0	0
O2 saturation pulseoxymetry	(0.33, 0.67]	55	0	0
O2 saturation pulseoxymetry	(0.67, 1.0]	1184	0	0
oasis	(-0.001, 0.33]	464	0	0
oasis	(0.33, 0.67]	1799	0	0
oasis	(0.67, 1.0]	372	0	0
death_in_hosp	0	2259	0	0
death_in_hosp	**1**	**396**	**135**	**0**
White Blood Cells	(-0.001, 0.33]	2274	104	0.43
White Blood Cells	(0.33, 0.67]	209	19	0.43
White Blood Cells	(0.67, 1.0]	11	1	0.43
Sodium (whole blood)	(-0.001, 0.33]	331	106	0.49
Sodium (whole blood)	(0.33, 0.67]	2203	16	0.49
Sodium (whole blood)	(0.67, 1.0]	36	5	0.49
gcs	(-0.001, 0.33]	70	111	0.51
gcs	(0.33, 0.67]	247	17	0.51
gcs	(0.67, 1.0]	2318	6	0.51
Heart Rate	(-0.001, 0.33]	1292	83	0.68
Heart Rate	(0.33, 0.67]	1322	48	0.68
Heart Rate	(0.67, 1.0]	31	3	0.68
Anion Gap	(-0.001, 0.33]	1031	79	0.75
Anion Gap	(0.33, 0.67]	1431	39	0.75
Anion Gap	(0.67, 1.0]	37	8	0.75

Table 27: Feature Relevance within Cluster Eight

Cluster nine reveals another noticeable subgroup with a mortality rate of 1. However, this group comprises 124 male patients. Initially, this cluster may not have been considered relevant due to its higher entropy compared to the original complete dataset. However, this higher entropy can be attributed to multiple continuous features such as “Heart Rate” or “Anion Gap”, which do not exhibit clear imbalances in their distributions.

This observation highlights a crucial aspect of this subgroup detection method. While subgroups with low entropy in their categorical features (e.g. gender, ethnicity, insurance status, marital status) offer more straightforward interpretations, the inclusion of continuous variables, like vital signs, allows for the discovery of feature patterns across all available features. In the stroke use case, the detection of such patterns, including vital signs, has been implemented. However, future researchers may explore the alternative perspective of considering entropy solely across categorical features to detect more interpretable subgroups.

As a final step, it was intended to calculate the fairness metrics based on cluster eight as the privileged group. This approach differs from the fairness metrics derived from the manual selection of features, as it enables the detection of unfairness patterns, independent of prior medical knowledge. By considering clusters as the privileged group, it becomes possible to assess fairness in a more data-driven manner, uncovering potential biases that may not have been evident through traditional feature-based analysis. This can be valuable in identifying hidden sources of bias within the model.

Regrettably, the limitations of the dataset for the stroke use case once again hinder the analysis of such specific groups. The reduced size and specificity of the subgroups obtained through clustering, in combination with the smaller proportion of test data, pose significant challenges. In addition with the limited number of positive cases, the calculation of fairness metrics becomes infeasible. The reliance on fairness metrics on the existence of true positive cases makes it impossible to conduct a comprehensive fairness analysis based on clusters in this particular setting.

A similar challenge arises when attempting to explore feature relevance using Shapley values. The comparison of feature influence across subgroups could potentially offer valuable insights. However, in the current selection, this investigation is not feasible due to the limited size of cluster eight.

In detail, cluster eight, initially consisting of 135 instances, was significantly reduced to only 22 instances after the test data split. With such a small number of instances, the calculation of feature relevance is not possible, and the resulting Shapley values are null, as illustrated in the following two figures.

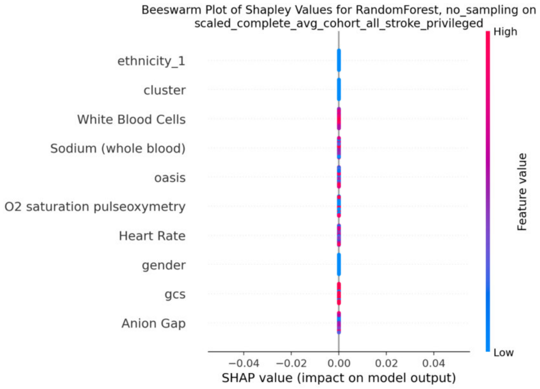


Figure 40: SHAP Beeswarm Plot, Cluster Eight

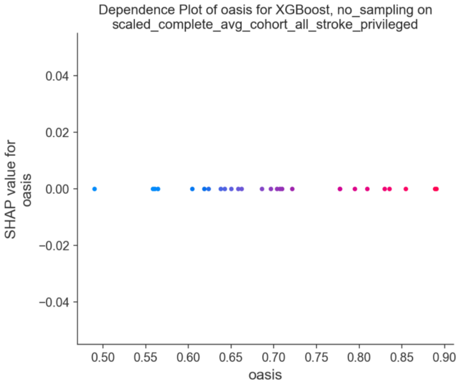


Figure 41: Dependence Plot on Oasis, Cluster Eight

Even with a higher test data rate of 30%, the cluster only contained 49 instances, which was still insufficient to conduct a meaningful investigation of fairness or feature relevance. Furthermore, increasing the test data rate to such high levels can compromise the overall prediction performance of the model.

Consequently, the intended analysis step based on cluster eight was not possible with the MIMIC-III dataset for the stroke use case. This again highlights the need to explore alternative use cases, such as general patient mortality, to determine if they provide a more suitable context for conducting this approach.

8.2.1 Optimized Subgroup Analysis Results

Various settings were explored in an attempt to achieve better results for this methodology, but unfortunately, this was not feasible for the selected stroke use case.

Most clusters became too small for a cluster count above eight, as was seen in the previous analysis. Thus an initial approach was to reduce the number of clusters to achieve larger clusters. However, this resulted in lower silhouette scores and made it difficult to identify meaningful subgroups. There exists a trade-off between having larger clusters that lack relevant subgroups and smaller clusters for which fairness measures and feature relevance cannot be reliably calculated.

Furthermore, another challenge arose when certain clusters exhibited a mortality rate of either 0 or 1. This was observed for a selection of six kMeans clusters, which initially seemed reasonable in terms of cluster count. However, for the relevant clusters, respectively, only a single class was represented. Although analyzing these clusters would be intriguing, the absence of multiple classes makes it impossible to calculate fairness measures or determine feature relevance.

For the use case at hand, this approach regrettably does not deliver insightful subgroups. Thus, these continued difficulties should be critically reflected. It must be questioned if the problem lies just in the limited dataset, or if the method, to manually search for relevant and analyzable clusters, may itself be misguided. For further improvements of this method, different steps can be considered.

One option is to filter the resulting clusters in regard to their interpretational value. For example, a minimum threshold for the cluster size could be implemented to remove too small cluster options.

Alternatively, it may be a promising step, to consider a further automated solution to directly find interpretable clusters. For this, the results for multiple cluster counts could be calculated in an iterative way and only those options that contain potentially useful clusters get displayed.

Ultimately, the actual applicability of this method needs to be tested on a different use case to determine whether adjustments are necessary or if the proposed steps can indeed yield meaningful results. Despite these unresolved questions, the concept of identifying relevant subgroups in this manner continues to hold promise.

In conclusion, although it was not possible to perform the final step of assessing fairness and feature importance across clusters in this study, the proposed method may prove to be useful for researchers aiming to identify relevant subgroups within a larger dataset. It offers the potential to uncover groups that might have been overlooked solely based on medical knowledge.

9 Conclusion

9.1 Summary of Results

The key finding of this thesis is the general applicability of the overall MIMIC-III dataset as a robust data source for healthcare research. A comprehensive descriptive analysis of the dataset was conducted and various clustering methods were employed to gain further insights. Especially, the thesis successfully replicated different prediction models and achieved results that were comparable to existing research.

Nevertheless, it is important to acknowledge the limitations and shortcomings that were identified in this thesis. Firstly, it was observed that the methods employed in this study did not yield improved prediction results compared to previous research findings. Furthermore, the focus on the stroke use case led to a dataset containing only 2,655 patients. This presented limitations in conducting detailed subgroup analysis. The small sample size restricted the ability to comprehensively examine and assess specific subgroups, thereby impeding a more comprehensive fairness analysis.

However, this thesis successfully presented various fairness metrics and revealed instances of unfairness. It became apparent, that the detection of unfairness alone is not sufficient, but in-depth investigations are necessary to understand and explain the underlying factors contributing to these results.

It further introduced manual and automated subgroup detection approaches. The importance of explanations for differences between subgroups, such as diverging feature relevance, was emphasized.

Overall, this thesis serves as a significant step in conducting research within the MIMIC-III database. The code provided alongside this thesis offers a valuable foundation for future researchers to build upon.

9.2 Limitations of this Thesis

As discussed in the previous chapters, the filtering process of the dataset resulted in a relatively small subset of stroke patients. To obtain more robust results, it would be beneficial to explore alternative use cases that can offer a larger and more diverse sample size.

One suggestion is to consider a use case with a larger dataset, such as 5,000 patients, where 1,000 patients remain in the 20% test data. By filtering for a subgroup representing 10% of the population, there would still be approximately 100 patients in the final test data subgroup. With a mortality rate of 10%, or potentially 20% with oversampling, there should be a sufficient number of true positive cases for further analysis.

Moreover, it is worth noting that the MIMIC-III dataset solely focuses on patients in the ICU. While the dataset provides valuable insights into diseases and conditions treated in the ICU, it does not cover the entire spectrum of medical conditions. Still, machine learning models can be very impactful for the whole field of medicine and further research might want to focus on more comprehensive hospital data.

It is also important to keep in mind that the geographical scope of the MIMIC-III dataset is strictly limited to patients treated in the United States.

9.3 Future Research: Alternative Directions

This section summarizes potential alternative avenues of research. Some of these aspects have already been mentioned throughout this thesis.

Select a larger Use Case
The impeding factor for this thesis was that the stroke use case with 2,655 patients is not sufficient for the analysis of specific subgroups. Thus, the step for future research with the highest priority is to select a suitable use case with a higher number of patients. One promising option can be general mortality within the complete MIMIC-III dataset, based on approximately 32,000 cases.

Implement Methods to Reduce Unfairness
The practical implementation of bias-mitigating methods was not within the scope of this thesis. This represents an important and extensive topic that warrants further investigation. It is recommended to first rework the feature selection and factorization process before the following adjustments are implemented.

As of yet, there are no known solutions, with which to automatically improve a model's fairness in an unsupervised way. Nevertheless, various approaches can be explored to manually intervene and construct a fairer model. One starting point is to modify the pre-processing stage to early-on mitigate model bias. In addition, in-processing methods offer opportunities to directly adjust the classification step. Lastly, post-processing algorithms can be employed to afterward refine the model's fairness [28]. For the stroke use case, mainly the following pre-processing techniques are considered promising solutions.

One common approach is to adjust attribute weights, to mitigate bias. By assigning different weights to the features, the influence of features that lead to potential unfairness can be reduced. At the same time, the features are not completely removed from the process and no information is lost. This method may be a promising option for future improvement.

Notably, the automated-subgroup clustering approach, presented in Chapter 8, may be regarded as a form of pre-processing adjustment. By discovering diverging fairness between subgroups one can then consider different methods to tackle potential inequality.

If the underlying reason for unfairness is representational bias, an increase in instances of the unprivileged subgroup is a common solution. However, it is not recommended to decrease the instances of the privileged groups, as this can reduce overall model performance.

Furthermore, conducting a thorough analysis of the underlying feature relevance is sensible to uncover the root causes of the disparate prediction quality. For example, it is essential to investigate if certain subgroups exhibit missing feature values or similar patterns, which may necessitate the refinement of data collection methods.

Finally, a more radical approach is to replace the existing model with an alternative model that is inherently less prone to bias.

In addition to the proposed methods, existing open-source tools are available that can assist in the analysis and improvement of model fairness. For example, the Responsible AI Toolbox[31], developed by Microsoft, reflects the growing recognition of the importance of fairness in AI applications. This toolbox offers a comprehensive and powerful solution for in-depth dataset analysis. Notably, the included "Fairness Dashboard" can aid in the fairness analysis and with the identification of potential steps for bias mitigation.

Investigate Explanations
A more detailed evaluation of explanations has not been implemented within this thesis. The comparison of feature importance across subgroups is only useful if these explanations are reliable. This remains an important area for future research.

Improve Prediction Quality
As of yet, the prediction results are not as good as comparable research, thus an increase in recall and accuracy are important points for improvement. Moreover, external validation on an alternative dataset is recommended.

Include alternative Features
There are several additional features in the MIMIC-III dataset that have not been incorporated into the analysis thus far. These features, such as transfers, services, microbiology events, and prescriptions, might provide valuable insights into the treatment process and the patient's conditions.

For instance, the number of transfers between different units could potentially serve as an indicator of the severity of the patient's condition or their need for specialized care. An alternative hypothesis is, that specific types of prescriptions might be indicators for relapse rates or treatment effectiveness. Moreover, the "note-events" table, which contains textual data such as clinical notes and reports, holds great potential for Natural Language Processing analysis. Integrating NLP techniques can offer new opportunities for predicting patient outcomes. However, these features mainly describe the treatment and behavior of the medical staff. As they are not directly related to the illness itself, they were not included in the analysis for now.

To facilitate the inclusion of these sources, it is possible to modify the SQL function "get_all_events_view". In general, more informed medical knowledge can further support feature engineering and lead to improved model performance.

[31] `https://github.com/microsoft/responsible-ai-toolbox`

Validate the Transferability of the Oasis Score
It is recommended to examine the Oasis score and its construction. The Oasis score, which exhibits the highest correlation with deaths, appears to be a valuable option for potential real-world integration in hospitals. However, it is important to validate its effectiveness and generalizability by testing it on external datasets beyond the MIMIC-III dataset from which it was developed. Conducting such external validation studies would provide valuable insights into the reliability and applicability of the Oasis score in diverse healthcare settings.

Investigate specific Patient Groups
Another interesting research approach worth considering is the analysis of patients who experience a stroke while they are already admitted to the ICU. This specific subgroup of patients could provide valuable insights into the changes in vital signs and other relevant factors leading up to a stroke. However, the MIMIC-III dataset may not have a sufficient number of such patients for a robust analysis of this particular scenario.

9.4 Future Research: Technical Adjustments

This section introduces different technical aspects that have the potential for further development. As mentioned before, the code developed for this thesis may offer a helpful foundation for research on the MIMIC-III dataset. There are several areas marked directly within the code as *"todo future research"*, which may be improved and expanded upon.

Setup
Some initial parts of the setup process should be reassessed. For instance, the mapping between the "CareVue" and "metavision" features does not seem ideal yet, as was seen with the missing values for the feature "O2 saturation pulseoxymetry". As was assumed within the fairness analysis, this pattern of missing values might have affected prediction quality for the female subgroup from the "CareVue" data system. Thus, an improvement in data quality is of great importance for prediction performance. Moreover, the feature selection is currently hardcoded. As this is not flexible, a table similar to the factorization table, which is described in the chapter regarding the use case setup, might offer a better solution. Finally, the distribution of hemorrhagic and ischemic stroke types raised some questions, and filtering based on ICD-9 codes need to be reconsidered.

Pre-Processing
Regarding the pre-processing step, a couple of steps can be refined. Firstly, the factorization of categorical features might benefit from an alternative approach, as relying on an external factorization table may be suboptimal when compared to one-hot encoding. Secondly, the outlier removal process can be reassessed to ensure its effectiveness. Thirdly, enhancing the overall analysis by incorporating time series data, rather than solely relying on average data per patient, could provide better prediction performance. This may require additional pre-processing, such as interpolation and imputation. Furthermore, optimization of the scaling method can support the prediction performance. Introduc-

ing the scaling method as a user-selectable parameter in the frontend would allow for customization based on specific requirements.

Correlations
The correlation analysis provided valuable insights, although certain features, like age, exhibited unexpectedly low p-values. It may be necessary to reevaluate the implemented correlation methods to obtain more reliable results.

Clustering
The clustering methods DBSCAN and SLINK are promising alternatives to the kMeans clustering approach. A deeper investigation into these methods and optimization of the respective parameters may lead to better results.

Classification
Similarly, for the classification models, further improvement of the XGBOOST classifier and the neural network model can be achieved by parameter optimization. Additionally, it may be worth considering the interpretation of the positive and negative classes in this analysis. While the current approach considers "death" as the positive case, it could be argued, that using "survival" as the positive class aligns more closely with common standards. However, it is important to acknowledge that there is a significant body of research where "death" also represents the positive class in mortality prediction studies. This matter remains open for discussion and warrants consideration based on the specific context.

Fairness Analysis
Next, expanding the fairness analysis with additional metrics or integrating existing, more comprehensive fairness dashboards can enhance the evaluation. Such tools might additionally offer methods to directly reduce unfairness inside the dataset. The introduced Responsible AI Toolbox or the integration of the "ASDF"-Dashboard may represent useful options for further development in this direction. Furthermore, incorporating Shapley values more extensively into the fairness analysis, as demonstrated in this official SHAP example[32], can be a valuable addition. By considering the Shapley values in the context of fairness metrics a more nuanced assessment of model fairness may be achieved.

Subgroup Analysis
For future research, it is recommended to prioritize the subgroup analysis within a larger use case setting. This approach has the potential to provide more comprehensive insights into the unfairness present in the dataset. Additionally, exploring different clustering algorithms to identify subgroups can be a promising improvement. As discussed in the respective chapter, it may be beneficial to calculate the feature entropy specifically for categorical features, as this can enhance the interpretation and distinction of clusters. At last, the overall approach might be reflected upon. A more automated solution to detect clusters, that are relevant and analyzable, could prove to be necessary.

[32] https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Explaining%20quantitative%20measures%20of%20fairness.html

Frontend
Finally, the frontend of this thesis provides a comprehensive dashboard that encompasses various topics and perspectives on the entire dataset. However, it can be beneficial to narrow down the focus to a specific topic, such as subgroup detection, and dedicate more in-depth analysis to that particular aspect. By scaling down the dashboard and placing emphasis on one main topic, the analysis can be further developed and explored in greater detail. This targeted approach would allow for a deeper understanding and more extensive exploration of the selected research area.

9.5 Final Conclusion

The primary objective of this thesis was to assess the effectiveness of predictive tools in the medical domain, recognizing the immense potential for the development of new digital tools in this field.

Indeed, despite working with a relatively limited dataset for the stroke use case, it was possible to develop predictive models. Although the current prediction results may not meet the requirements for real-world implementation, these methods demonstrate the potential for reliable predictions in the future. Thus, this work can serve as a foundation for development within the MIMIC-III dataset. The thesis further highlighted potential continuing avenues of research, that may be pursued by future data scientists.

However, it is crucial to remain mindful of the concerns expressed by organizations such as the European Union and the World Health Organization, as highlighted at the outset of this thesis. In line with these considerations, this thesis emphasizes the importance of explainability and fairness in predictive models. Once again, it must be highlighted that the implementation of unfair prediction models may turn out harmful and diminish trust.

Nevertheless, the benefits offered by these predictive tools are undeniable and their significant potential in the medical field is cause for optimism. Further research in the field of data science within healthcare holds the promise of making a positive impact and, ultimately, saving lives.

In the end, the findings presented in this thesis shall encourage future developers to critically reflect on their work and approach it with a heightened sense of responsibility.

List of Figures

List of Tables

References

[1] Kanadpriya Basu, Ritwik Sinha, Aihui Ong, and Treena Basu. Artificial intelligence: How is it changing medical sciences and its future? *Indian journal of dermatology*, 65(5):365–370, 2020.

[2] Sri Sunarti, Ferry Fadzlul Rahman, Muhammad Naufal, Muhammad Risky, Kresna Febriyanto, and Rusni Masnina. Artificial intelligence in healthcare: opportunities and risk for future. *Gaceta Sanitaria*, 35:S67–S70, 2021. The 1st International Conference on Safety and Public Health.

[3] Silja Voeneky, Philipp Kellmeyer, Oliver Mueller, and Wolfram Burgard. *The Cambridge Handbook of Responsible Artificial Intelligence*. Cambridge University Press, 2022.

[4] World Health Organization. *WHO guideline*. World Health Organization, Geneva, 2019.

[5] Michael Moor, Bastian Rieck, Max Horn, Catherine R. Jutzeler, and Karsten Borgwardt. Early prediction of sepsis in the icu using machine learning: A systematic review. *Frontiers in medicine*, 8:607952, 2021.

[6] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmark of deep learning models on large healthcare mimic datasets.

[7] Blanca Vazquez, Gibran Fuentes-Pineda, Fabian Garcia, Gabriela Borrayo, and Juan Prohias. Risk markers by sex for in-hospital mortality in patients with acute coronary syndrome: a machine learning approach.

[8] Laura A. Barrett, Seyedeh Neelufar Payrovnaziri, Jiang Bian, and Zhe He. Building computational models to predict one-year mortality in icu patients with acute myocardial infarction and post myocardial infarction syndrome. *AMIA Summits on Translational Science Proceedings*, 2019:407–416, 2019.

[9] Jiayang Wang, Xiaoshuo Huang, Lin Yang, and Jiao Li. National institutes of health stroke scale (nihss) annotations for the mimic-iii database.

[10] Xiao-Dan Li and Min-Min Li. A novel nomogram to predict mortality in patients with stroke: a survival analysis based on the mimic-iii clinical database. *BMC medical informatics and decision making*, 22(1):92, 2022.

[11] Peter Appelros, Birgitta Stegmayr, and Andreas Terént. Sex differences in stroke epidemiology: a systematic review. *Stroke*, 40(4):1082–1090, 2009.

[12] Kathryn M. Rexrode, Tracy E. Madsen, Amy Y. X. Yu, Cheryl Carcel, Judith H. Lichtman, and Eliza C. Miller. The impact of sex and gender on stroke. *Circulation research*, 130(4):512–528, 2022.

[13] Alistair Johnson, Tom Pollard, and Roger Mark. Mimic-iii clinical database, 2020.

[14] Alistair E W Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018.

[15] Michael Moor, Max Horn, Bastian Rieck, Damian Roqueiro, and Karsten Borgwardt. Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time warping, 2019.

[16] Alistair E. W. Johnson, Jerome Aboab, Jesse D. Raffa, Tom J. Pollard, Rodrigo O. Deliberato, Leo A. Celi, and David J. Stone. A comparative analysis of sepsis identification methods in an electronic database. *Critical care medicine*, 46(4):494–499, 2018.

[17] Rebecca Woodfield, Ian Grant, and Cathie L. M. Sudlow. Accuracy of electronic health record data for identifying stroke cases in large-scale epidemiological studies: A systematic review from the uk biobank stroke outcomes group. *PloS one*, 10(10):e0140533, 2015.

[18] Alessandra C. Goulart, Tiotrefis G. Fernandes, Itamar S. Santos, Airlane P. Alencar, Isabela M. Bensenor, and Paulo A. Lotufo. Predictors of long-term survival among first-ever ischemic and hemorrhagic stroke in a brazilian stroke cohort. *BMC neurology*, 13:51, 2013.

[19] Alistair E. W. Johnson, Andrew A. Kramer, and Gari D. Clifford. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Critical care medicine*, 41(7):1711–1718, 2013.

[20] Jero Schäfer and Lena Wiese. Clustering-based subgroup detection for automated fairness analysis. In Silvia Chiusano, Tania Cerquitelli, Robert Wrembel, Kjetil Nørvåg, Barbara Catania, Genoveva Vargas-Solar, and Ester Zumpano, editors, *New Trends in Database and Information Systems*, volume 1652 of *Communications in Computer and Information Science*, pages 45–55. Springer International Publishing, Cham, 2022.

[21] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.

[22] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, pages 21–34. Citeseer, 1997.

[23] Miroslav Kubat. *Machine learning and knowledge discovery in databases; Part 1*. Springer, 2021, Cham, 2021.

[24] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011.

[25] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[26] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[27] Donato Gemmati, Katia Varani, Barbara Bramanti, Roberta Piva, Gloria Bonaccorsi, Alessandro Trentini, Maria Cristina Manfrinato, Veronica Tisato, Alessandra Carè, and Tiziana Bellini. Bridging the gap: Everything that could have been avoided if we had applied gender medicine, pharmacogenetics and personalized medicine in the gender-omics and sex-omics era. *International journal of molecular sciences*, 21(1), 2019.

[28] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, New York, NY, USA, 2021. ACM.

[29] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics, 2021.

[30] Corinna Hertweck and Christoph Heitz. A systematic approach to group fairness in automated decision making. *8th Swiss Conference on Data Science*, 106:1–6, 2021.

[31] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[32] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[33] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness. In *Proceedings of the 24th ACM, SIGKDD, International Conference on Knowledge Discovery, Data Mining.* ACM, jul 2018.

[34] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. *CoRR*, abs/1904.05419, 2019.