Visualisierungsinstitut der Universität Stuttgart

Universität Stuttgart
Allmandring 19
D–70569 Stuttgart

Bachelorarbeit

System zur visuellen Analyse für Daten der kontinuierlichen Glukosemessung von Diabetes-Patienten

Muhammed Kaya

Studiengang:	Softwaretechnik
Prüfer/in:	Prof. Dr. Daniel Weiskopf
Betreuer/in:	Dipl.-Inf. Tanja Munz, M.Sc.
Beginn am:	1. März 2021
Beendet am:	1. September 2021

Kurzfassung

Die vorliegende Arbeit beschäftigt sich mit der visuellen Analyse von Daten der kontinuierlichen Glukosemessung. Hierzu wird ein Visualisierungssystem entwickelt, welches in der Lage ist, Muster und Ausreißer in den Daten mithilfe von verschiedenen Visualisierungen darzustellen und dazugehörige Statistiken zur Auswertung der Daten zu generieren. Erreicht wird dies mithilfe von unüberwachtem maschinellen Lernen in Kombination mit klassischen Zeitreihen-Visualisierungen. Zunächst werden theoretische Grundlagen gemeinsam mit aktuellen Visualisierungsansätzen aus verschiedenen verwandten Arbeiten vermittelt. Darauf basierend wird ein Visualisierungsansatz für die Analyse der Daten vorgestellt. Diese beinhalten Kastengrafiken, Heatmaps, Linien-, Säulen-, Kreis- und Streudiagramme. Schließlich wird dazu ein Visualisierungssystem realisiert, welches in Form von drei Fallstudien evaluiert wird. Dafür werden reale Daten verwendet, die aus der kontinuierlichen Glukosemessung verschiedener Patienten/-innen entstanden sind. Die Ergebnisse der Fallstudien konnten zeigen, dass sich der Einsatz von unüberwachten maschinellen Lernverfahren in Kombination mit klassischen Zeitreihen-Visualisierungen für die Erkennung von Mustern und Ausreißern in den Daten verwenden lassen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AGP Ambulantes Glukoseprofil (engl. *Ambulatory Glucose Profile*). 27

CGM Kontinuierliche Glukosemessung (engl. *Continous Glucose Monitoring*). 13

GMI Glukose-Management-Indikator (engl. *Glucose Management Indicator*). 22

GV Glykämische Variabilität (engl. *Glycemic Variability*). 22

MDS Multidimensionale Skalierung (engl. *Multidimensional Scaling*). 21

PCA Hauptkomponentenanalyse (engl. *Principial Component Analysis*). 20

TAR Zeit oberhalb des Zielbereichs (engl. *Time above Range*). 22

TBR Zeit unterhalb des Zielbereichs (engl. *Time below Range*). 22

TIR Zeit im Zielbereich (engl. *Time in Range*). 22

t-SNE t-Distributed Stochastic Neighbor Embedding. 21

UMAP Uniform Manifold Approximation and Projection. 21

1 Einleitung

1.1 Motivation

Diabetes mellitus - eine Stoffwechselerkrankung, die seit Jahren immer mehr Aufmerksamkeit erregt hat. Die Anzahl der an Diabetes erkrankten Personen nahm in den letzten Jahrzehnten erheblich zu. Laut der Weltgesundheitsorganisation (engl. *World Health Organization* (WHO)) [Wor21] betrug die Anzahl an Diabetiker/-innen im Jahr 1980 insgesamt 108 Millionen. Im Gegensatz dazu kann heutzutage aus den Statistiken der Weltgesundheitsorganisation ein starker Anstieg beobachtet werden. Diese verzeichnen eine Anzahl an 422 Millionen Diabetiker/-innen im Jahr 2014. Zudem hat diese Stoffwechselerkrankung des Öfteren einen erkennbaren Anteil an vielen Todesfällen gehabt. So waren nach den Statistiken der Weltgesundheitsorganisation im Jahr 2019 schätzungsweise 1,5 Millionen Todesfälle direkt durch Diabetes verursacht worden, während im Jahr 2012 insgesamt 2,2 Millionen Todesfälle aufgrund hoher Blutzuckerwerte verzeichnet wurden [Wor21].

Diese Statistiken zeigen, dass Diabetes eine ernstzunehmende Stoffwechselerkrankung ist. Deren Behandlung spielt demnach weltweit eine sehr wichtige Rolle. Es lässt sich dadurch ebenfalls erkennen, dass für eine entsprechende Therapie regelmäßige Arztbesuche unumgänglich sind. Zudem werden Betroffene mithilfe von Lanzetten zu regelmäßigen Blutzuckerkontrollen veranlasst. Während dadurch ein akkurater Wert über den Glukosegehalt im Blut gewonnen wird, hält sich die Anzahl der Werte in einem gewissen Rahmen. Diese können nur zu festen Zeiten, an denen eine Messung unternommen wird, jeweils einen Wert liefern, wodurch ein begrenzter Verlauf der Werte geschaffen wird. Dabei ist es gerade für Therapieanpassungen entscheidend, den Verlauf dieser Werte zu verstehen. Vielmehr ist die Erkennung von Mustern und Ausreißern zu spezifischen Zeitabschnitten wichtig, damit diese rechtzeitig behandelt werden können.

Inzwischen werden für die Blutzuckermessungen außerdem auch spezielle Systeme eingesetzt. Jene werden als Systeme zur kontinuierlichen Glukosemessung (engl. *Continous Glucose Monitoring* (CGM)) bezeichnet, wie zum Beispiel das FreeStyle Libre 2[1] oder das Dexcom G6[2]. CGM-Systeme sind in der Lage in festen Zeitabständen, wie beispielsweise fünf Minuten, einen Wert über den Glukosegehalt im Blut zu liefern. Mit den generierten CGM-Daten lassen sich anschließend Visualisierungen über den Verlauf der Werte für die Analyse schaffen. Hierzu existierende Analysesysteme können diese bereits als Liniendiagramm oder statistische Werte in aggregierter Form darstellen. Weitere Systeme versuchen daraus auch Muster und Trends herzuleiten.

[1] https://www.freestylelibre.de/libre/freestylelibre2.html
[2] https://www.dexcom.com/

1.2 Zielsetzung

Das Ziel dieser Arbeit ist die Entwicklung eines Visualisierungssystem, welches zur Analyse von Daten der kontinuierlichen Glukosemessung eingesetzt werden soll. Jenes System soll insbesondere zur Erkennung von Mustern und Ausreißern in den Daten dienen. Erreicht werden soll dies mithilfe von unüberwachtem maschinellen Lernen in Kombination mit klassischen Zeitreihen-Visualisierungen. Abschließend soll das Visualisierungssystem anhand von drei Fallstudien evaluiert werden.

1.3 Aufbau der Arbeit

Die Arbeit ist folgenderweise gegliedert:

Kapitel 2 - Grundlagen: In diesem Kapitel werden die theoretischen Grundlagen vermittelt, um ein grundlegendes Verständnis für die darauf folgenden Kapitel zu schaffen. Zu Beginn werden Zeitreihen mit möglichen Analyse-, Vorverarbeitungs- und Visualisierungsmethoden vorgestellt. Anschließend wird auf das Prinzip der Dimensionsreduktion eingegangen und diesbezüglich verschiedene Verfahren erläutert. Zum Schluss folgt eine Einführung zur Realisierung der Messungen und Auswertungen, welche bei den Systemen zur kontinuierlichen Glukosemessung verwendet werden.

Kapitel 3 - Verwandte Arbeiten: Dieses Kapitel befasst sich mit den verwandten Arbeiten. Hierzu werden aktuelle Visualisierungsansätze zur Zeitreihenanalyse mit unüberwachtem maschinellen Lernen und zu den CGM-Daten vorgestellt.

Kapitel 4 - Visualisierungsansatz: Im vierten Kapitel wird die verwendete Methodik zur Visualisierung der CGM-Daten vorgestellt. Es werden zunächst die funktionalen Anforderungen und ein möglicher Arbeitsvorgang des Visualisierungssystems beschrieben. Daraufhin wird die Vorgehensweise zur Vorverarbeitung dieser Daten erläutert. Anschließend werden zuerst detailliert auf die Visualisierungen eingegangen und danach die grafische Benutzeroberfläche des Visualisierungssystems vorgestellt, gefolgt von weiteren Interaktionsmöglichkeiten zu den einzelnen Visualisierungen.

Kapitel 5 - Implementation: Das fünfte Kapitel beinhaltet Informationen zum Back- und Frontend des Visualisierungssystems. Hierzu werden verwendete Bibliotheken zur Verarbeitung und Visualisierung der CGM-Daten vorgestellt.

Kapitel 6 - Fallstudie: Dieses Kapitel beschäftigt sich mit der Evaluation des entwickelten Visualisierungssystems. Dafür werden anhand von realen Daten, die aus der kontinuierlichen Glukosemessung verschiedener Patienten/-innen entstanden sind, die Fähigkeiten des Visualisierungssystems untersucht. Durchgeführt wird die Evaluation in Form von drei Fallstudien, welche jeweils auf drei verschiedenen CGM-Daten beruhen.

Kapitel 7 - Zusammenfassung und Ausblick: Das letzte Kapitel fasst die Ergebnisse der Arbeit zusammen und stellt Anknüpfungspunkte vor.

2 Grundlagen

In diesem Kapitel werden notwendige Grundlagen vermittelt, welche zum Verständnis der nachfolgenden Kapitel vorausgesetzt werden. Zunächst folgt eine allgemeine Einführung zu Zeitreihen. Darin werden Ansätze zur Analyse, Vorverarbeitung und Visualisierung der Zeitreihendaten vorgestellt. Anschließend wird auf das Prinzip der Dimensionsreduktion eingegangen. Hierzu wird zuerst der Zweck, und dann verschiedene Verfahren zu der Dimensionsreduktion erläutert. Zum Schluss werden Inhalte zu den Systemen der kontinuierlichen Glukosemessung vorgestellt. Diese beinhalten grundlegende Informationen zu den Messungen und Auswertungen der CGM-Daten.

2.1 Zeitreihen

Nach Shurkhovetskyy et al. [SAAF17] beschreiben Zeitreihen eine Reihe von chronologisch angeordneten Beobachtungen zu unterschiedlichen Zeitpunkten. Demnach bilden sie heutzutage eine wichtige Grundlage für die Analyse von Informationen und sind in vielen Bereichen wie Finanzwesen, Medizin und Wissenschaft vertreten. Weiterhin basieren die Beobachtungen einer Zeitreihe jeweils auf einem Zeitstempel und beschreiben bestimmte Eigenschaften, beispielsweise einen spezifischen Wert, eine Kombination von Werten oder auch bestimmte Variablen [SAAF17]. Zeitreihen, welche nur einen Wert pro Zeitstempel enthalten, werden auch als univariate Zeitreihen bezeichnet. Bei der Involvierung mehrerer Werte wird von einer multivariaten Zeitreihe gesprochen [SAAF17]. Weiterhin wird durch die Art der Messung zwischen einer kontinuierlichen und einer diskreten Zeitreihe unterschieden. Kontinuierliche Zeitreihen zeichnen sich hierbei durch eine über den gesamten Zeitraum anhaltende Messung aus, während bei diskreten Zeitreihen die Messungen nur zu spezifizierten Zeiten oder in gleichen Abständen erfolgen [VP20]. In der Tabelle 2.1 wird ein Beispiel zum Aufbau einer diskreten multivariaten Zeitreihe veranschaulicht. Darin werden pro Zeitstempel zwei Beobachtungen in Zeitabständen von fünf Minuten beschrieben.

Zeitstempel	Wert	Kategorie
2021-03-01T00:00:00.000Z	119	A
2021-03-01T00:05:00.000Z	120	B
2021-03-01T00:10:00.000Z	122	B
...	...	...
2021-08-31T23:45:00.000Z	141	C
2021-08-31T23:50:00.000Z	140	C
2021-08-31T23:55:00.000Z	139	B

Tabelle 2.1: Beispiel zum Aufbau einer diskreten multivariaten Zeitreihe mit zwei Beobachtungen (*Wert*, *Kategorie*) pro Zeitstempel in Zeitabständen von fünf Minuten.

2.1.1 Analyse

Die Zeitreihenanalyse beschreibt eine Disziplin, die sich mit der Erkennung von Mustern und Auffälligkeiten in Zeitreihen auseinandersetzt und darüber hinaus Prognosen über deren weitere Entwicklung generiert [VP20]. Muster können dabei in verschiedenen Formen auftreten. Zudem werden diese in der Zeitreihenanalyse auch als Komponenten der Zeitreihe bezeichnet und setzen sich aus Trend-, Saison-, zyklischer und irregulärer Komponente zusammen [Dod08]. Mit der Trendkomponente wird die langfristige Entwicklungstendenz der Zeitreihe beschrieben, während zyklische Schwankungen mit regelmäßigen Abständen durch die Saisonkomponente ausgedrückt werden. Die zyklische Komponente hingegen charakterisiert unregelmäßige zyklische Schwankungen. Folglich werden die restlichen Verhalten der irregulären Komponente zugeordnet [Dod08]. Weitere Auffälligkeiten werden durch Ausreißer (engl. *Outlier*) oder auch Anomalien beschrieben. Damit werden einzelne Werte bezeichnet, die eine starke Abweichung zu anderen Werten aufweisen, welche durch untypische Verhalten bei Datengenerierungsprozessen entstehen und oft nützliche Informationen über ungewöhnliche Merkmale, die zu neuen Erkenntnissen führen können, enthalten [Agg15].

Diese Arbeit bezieht sich auf den analytischen Aspekt und wird daher nicht auf die Prognose von Zeitreihen eingehen.

2.1.2 Vorverarbeitung

Bevor Zeitreihen analysiert werden können, müssen diese häufig auf die Analyse vorbereitet werden. Dabei spielt die Behebung von Fehlerwerten eine zentrale Rolle, denn unbehandelt besteht die Gefahr, solche als richtige Werte zu interpretieren [Agg15]. Für die Vorverarbeitung der Daten hat sich Aggarwal [Agg15] folgende Gedanken gemacht:

Entstehung von Fehlerwerten

Fehlerwerte entstehen entweder durch technisches Versagen oder durch menschliche Fehler. Dabei können folgende Ursachen der Grund sein:

- Hardware-Ausfälle oder ausgeschöpfte Akkus, die Sensoren bei der Messung beeinträchtigen.
- Falsche Messungen durch ungenaue Sensoren.
- Willkürliche Falschangaben durch Nutzer/-innen.
- Manuell erstellte Daten.
- Zu kostspielige Daten, die nicht erfasst werden.

Behandlung von Fehlerwerten

Um die Fehlerwerte in den Daten richtig handzuhaben, werden verschiedene Verfahren angewendet. Dafür ist die Verwendung der richtigen Methode für die Art eines Fehlerwertes entscheidend. Hierbei wird zwischen fehlendem und falschem Wert unterschieden.

Für die Behandlung von fehlenden Werten können folgende Methoden verwendet werden:

- Eliminierung der Daten.
- Schätzung oder Imputation.
- Den Analyseprozess so designen, dass mit fehlenden Werten gearbeitet werden kann.

Jedoch ist zu beachten, dass die erste Methode ungünstig ist, wenn viele fehlende Werte in den Daten existieren, und die zweite Methode zu weiteren Fehlern führen kann. In solchen Fällen haben Zeitreihen den Vorteil fehlende Werte linear oder mithilfe des Durchschnitts der umgebenen Werte zu interpolieren.

Für die Behandlung von falschen Werten können folgende Methoden Abhilfe schaffen:

- Inkonsistente Werte durch den Abgleich richtiger Werte korrigieren.
- Durch Fachwissen richtige Werte herleiten.
- Ausreißer mithilfe vom statistischen Verhalten der Daten erkennen.

2.1.3 Visualisierung

Visualisierungen sind in der Lage bessere Übersichten über die Zeitreihen zu verschaffen als vergleichsweise Tabellen (vgl. Tabelle 2.1). Zudem werden sie auch in der Zeitreihenanalyse verwendet. Für die Visualisierung von zeitlichen Daten existieren viele verschiedene Ansätze, die auf unterschiedlichen Arten von Nutzen sind [AMM+08] [TAMS17]. Nachfolgend werden dazu einige Visualisierungsansätze vorgestellt, welche unter anderem im entwickelten Visualisierungssystem zum Einsatz kommen.

Liniendiagramm

Zeitreihen werden klassischerweise in einem Liniendiagramm (engl. *Line Chart*) dargestellt. Dabei wird die horizontale Achse mit den Zeiten und die vertikale Achse mit den Werten versehen. Folglich wird eine Linie generiert, die eine univariate Zeitreihe darstellt (vgl. Abbildung 2.1a). Multivariate Zeitreihen bilden hingegen pro Beobachtung jeweils eine eigene Linie ab. Um in solchen Fällen die Linien besser voneinander zu unterscheiden, werden verschiedene Farben oder Linienarten verwendet [JME10].

Gestapeltes Säulendiagramm

Eine weitere Möglichkeit zur Visualisierung von Zeitreihen sind gestapelte Säulendiagramme (engl. *Stacked Bar Chart*). Diese sind in der Lage definierte Zeitabschnitte als Säulen zu repräsentieren und zusätzliche Eigenschaften, wie die Gesamtanzahl einzelner Kategorien, durch ihre Höhe auszudrücken [GE18]. In der Abbildung 2.1b wird hierzu ein Beispiel veranschaulicht. Die Tage werden in der horizontalen Achse und die Gesamtanzahl einzelner Kategorien in der vertikalen Achse abgebildet. Unterteilt werden diese anhand von drei Kategorien (*A*, *B*, *C*) innerhalb von 24 Stunden.

Streudiagramm

Streudiagramme (engl. *Scatterplot*) bilden eine weitere Möglichkeit zur Visualisierung von multivariaten Zeitreihen. Diese können Zeitreihen in höheren Dimensionen darstellen. Dabei basiert die Höhe der Dimension auf der Anzahl der Beobachtungen. Das liegt daran, dass die Werte eines Zeitstempels als Tupel für die Koordinierung benötigt werden. Demnach werden auch die Achsen mit den einzelnen Beobachtungen beschriftet. Die Tupel werden dabei typischerweise als Punkte visualisiert (vgl. Abbildung 2.1c). Für den zeitlichen Verlauf werden die Punkte durch Linien chronologisch miteinander verbunden und können zusätzlich mit Richtungspfeilen versehen werden. In jenem Zusammenhang wird auch von einem „Connected Scatterplot“ gesprochen [HKF16].

Kreisdiagramm

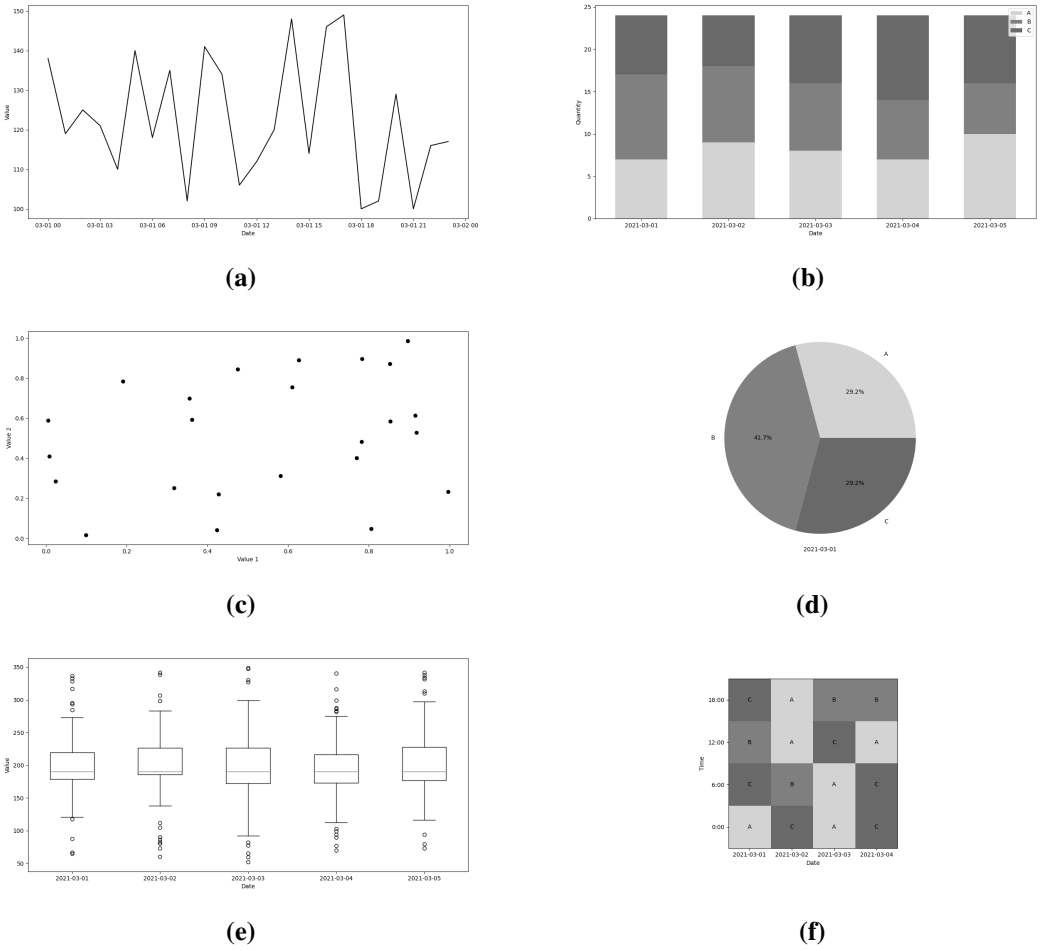
Kreisdiagramme (engl. *Pie Chart*) zeichnen sich durch die Visualisierung von prozentualen Verteilungen aus. Bei Zeitreihen können dadurch die Verteilungen spezifizierter Kategorien veranschaulicht werden. Damit lassen sich beispielsweise stündliche Stromverbräuche mehrerer Büros innerhalb von 24 Stunden darstellen [MEB+13]. Für die Veranschaulichung wird in der Abbildung 2.1d ein Kreisdiagramm dargestellt. Darin werden die prozentualen Verteilungen der Kategorien (*A*, *B*, *C*) über einen Tag beschrieben. Die Prozentwerte drücken hierbei die Dauer in den jeweiligen Kategorien aus.

Kastengrafik

Für die Visualisierung von Zeitreihen eignen sich auch Kastengrafiken (engl. *Boxplot*). Dabei handelt es sich um spezifizierte Zeitabschnitte in Form von Kasten, welche zudem in der Lage sind, statistische Werte wie Median, Minimum, Maximum, erstes und drittes Quartil zu beschreiben [VP20]. In der Abbildung 2.1e werden mehrere Kasten nebeneinander visualisiert. Die horizontale Achse beschreibt die Tage und die vertikale Achse die gemessenen Werte. Als Zeitraum werden in diesem Fall fünf Tage beschrieben. Mit den Größen der Kasten werden das erste und dritte Quartil repräsentiert. Die Linie innerhalb der Kasten stellt den Median dar. Die obersten und untersten horizontalen Linien werden als Antennen (engl. *Whisker*) bezeichnet. Werte, die außerhalb dieser Antennen liegen, werden als Ausreißer angesehen. Weiterhin werden anhand der Ausreißer auch das Minimum bzw. Maximum der Werte spezifiziert.

Heatmap

Bei den Heatmaps werden bestimmte Kategorien durch Farben hervorgehoben und somit mögliche Muster aufgedeckt [GE18]. Weiterhin können mit Heatmaps auch mehrere Zeitreihen dargestellt werden, wodurch sich insbesondere visuelle Vergleiche von Zeitreihen ermöglichen [PND20]. Ein Beispiel dazu kann der Abbildung 2.1f entnommen werden. Darin werden in der horizontalen Achse die Tage und in der vertikalen Achse die Stunden dargestellt. Diese werden anhand von drei Kategorien (*A*, *B*, *C*) spezifiziert. Somit ergibt sich ein Vergleich für die Kategorien der einzelnen Zeitabschnitte über verschiedene Tage.



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Abbildung 2.1: Beispiele zur Visualisierung von Zeitreihen: (a) Liniendiagramm: Darstellung eines Tages; (b) Gestapeltes Säulendiagramm: Darstellung von fünf Tagen mit jeweils drei Kategorien; (c) Streudiagramm: Darstellung von 24 Tagen als Punkte; (d) Kreisdiagramm: Darstellung eines Tages mit drei Kategorien; (e) Kastengrafik: Darstellung von fünf Tagen; (f) Heatmap: Darstellung von vier Tagen mit Zeitabständen von jeweils sechs Stunden.

2.2 Dimensionsreduktion

Zeitreihen beschreiben oft große Mengen an Informationen, die Probleme bei der Visualisierung und Analyse der Daten bereiten können. Zum einen können die Visualisierungen durch die vorhandene Bildschirmauflösung eingeschränkt werden, zum anderen jedoch können große Mengen an Daten höhere Rechenleistungen erfordern [SAAF17]. Besonders multivariate Zeitreihen stellen hierbei eine große Herausforderung dar. Dies ist unter anderem ihrer hohen Dimensionalität geschuldet. Falls multivariate Zeitreihen größere Dimensionen als vorhandene Datenproben besitzen, können bei einigen Algorithmen ungenaue Ergebnisse auftreten, da diese mehr Datenproben als Dimensionen erfordern [HCGD20]. Darüber hinaus ist dieses Phänomen auch als „Fluch der Dimensionalität“ bekannt [KM17]. Um hier entgegen zu wirken, können weitere Datenproben hinzugefügt oder die Anzahl der Beobachtungen bzw. die Dimensionen der Datenproben reduziert werden [HCGD20]. In solchen Fällen werden allerdings Dimensionsreduktionsverfahren bevorzugt, da der Gewinn neuer Datenproben nicht immer möglich ist [HCGD20]. Die Dimensionsreduktion bezeichnet einen Prozess, bei dem die Dimensionen der Daten reduziert werden und gleichzeitig relevante Informationen erhalten bleiben [HCGD20]. Es wird dabei zwischen der Feature Selection und der Feature Extraction unterschieden. Ein Feature bezeichnet in diesem Fall eine Beobachtung der Zeitreihe. Bei der Feature Selection wird die Dimensionsreduktion durch eine gezielte Auswahl einer Untermenge der Dimensionen erreicht [VAT19]. Allerdings besteht die Gefahr Features aufzunehmen, die möglicherweise ausgeschlossen werden sollten [VAT19]. Die Feature Extraction hingegen nimmt alle Features auf und versucht daraus die relevanten Informationen abzuleiten [VAT19].

Für die Dimensionsreduktion existieren mehrere Verfahren, welche sich in überwachtes und unüberwachtes maschinelles Lernen kategorisieren lassen. Beim überwachten maschinellen Lernen wird anhand von annotierten Datenproben trainiert, wodurch Zielvariablen vorgegeben werden, die der Algorithmus nur noch korrekt zuzuordnen braucht [SA13]. Dagegen werden beim unüberwachten maschinellen Lernen keine Datenproben annotiert, was dazu führt, dass durch gemeinsame Muster in den Daten die korrekten Zielvariablen spezifiziert werden müssen [SA13].

Der Fokus dieser Arbeit limitiert sich dabei nur auf Verfahren, die auf der Feature Extraction mit unüberwachtem maschinellen Lernen beruhen. Hierzu werden nachfolgend vier Dimensionsreduktionsverfahren vorgestellt. Der Unterschied liegt dabei in ihrer Vorgehensweise und den daraus entstehenden Resultaten.

2.2.1 Hauptkomponentenanalyse (PCA)

Die Hauptkomponentenanalyse (engl. *Principial Component Analysis* (PCA)) ist ein lineares Dimensionsreduktionsverfahren, welches von Pearson [Pea01] im Jahr 1901 eingeführt und von Hotelling [Hot33] in 1933 weiterentwickelt wurde. Das Ziel der Hauptkomponentenanalyse ist es, die relevanten Informationen aus den Dimensionen der Daten zu extrahieren, und diese dann anhand von Hauptkomponenten in kleineren Dimensionen zu beschreiben. Um dies zu erreichen, werden nacheinander die Hauptkomponenten durch eine Hauptachsentransformation gebildet. Dabei wird, beginnend mit der ersten Hauptkomponente, die größte Varianz der gemessenen Daten, in abnehmender Folge beschrieben. Somit werden neue Koordinaten für die gemessenen Daten gewonnen, wodurch die Un-/Ähnlichkeiten der Daten zueinander ersichtlich werden.

2.2.2 Multidimensionale Skalierung (MDS)

Die Multidimensionale Skalierung (engl. *Multidimensional Scaling* (MDS)) [Kru64] bildet ein nicht-lineares Dimensionsreduktionsverfahren, dessen Ansatz auf Distanzen basiert. Hierzu werden zunächst die Distanzen der gemessenen Daten zueinander benötigt, welche beispielsweise mit der Euklidischen Distanz ausgerechnet werden können. Anhand dieser Distanzen wird bei dem Verfahren versucht, die Daten möglichst genau in einer kleineren Dimension anzuordnen. Anschließend lassen sich durch die Abstände der Daten die Un-/Ähnlichkeiten dieser zueinander ausdrücken. Größere Distanzen veranschaulichen dabei die Unähnlichkeiten.

2.2.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Bei dem t-SNE Verfahren von van der Maaten und Hinton [MH08] handelt es sich um ein nicht-lineares Dimensionsreduktionsverfahren, mit dem Ziel, hochdimensionale Daten auf zwei oder drei Dimensionen zu reduzieren. Das Verfahren versucht dabei mithilfe von Distanzen der Daten die Wahrscheinlichkeiten für die Ähnlichkeiten zueinander zu ermitteln. Als Resultat werden dabei Ansammlungen von Punkten geschaffen, bei denen ähnliche Daten möglichst nah aneinander und unähnliche Daten möglichst fern voneinander liegen.

2.2.4 Uniform Manifold Approximation and Projection (UMAP)

UMAP wurde im Jahr 2018 von McInnes et al. [MHM20] als ein nicht-lineares Dimensionsreduktionsverfahren eingeführt. Sie verfolgt ein ähnliches Ziel wie das t-SNE Verfahren und soll laut McInnes et al. bessere Geschwindigkeiten und Datenstrukturierungen anbieten. Das Verfahren basiert auf der exponentiellen Wahrscheinlichkeitsverteilung und kann beliebige Distanzen verwenden.

2.3 System zur kontinuierlichen Glukosemessung

Nach Harreiter und Roden [HR19] bezeichnet Diabetes mellitus eine Gruppe von Stoffwechselerkrankungen, bei der eine Erhöhung des Blutzuckers (Hyperglykämie) vorliegt. In schweren Fällen können Symptome wie starker Durst, häufiges Wasserlassen, Müdigkeit und Leistungsabfall auftreten. Chronische Fälle können zu Langzeitfolgen führen, die Schäden an verschieden Organen und Geweben verursachen können. Weiterhin wird Diabetes mellitus in mehrere Arten klassifiziert. Zu den bekanntesten Fällen zählen Typ-1- und Typ-2-Diabetes. Die Typ-1-Variante wird durch ein Insulinmangel ausgezeichnet, während bei der Typ-2-Variante eine Insulinresistenz vorliegt, die zu höheren Blutzuckerwerten führt [HR19].

Daten, die mithilfe von CGM-Systemen generiert werden, sind wie Zeitreihendaten aufgebaut. Dabei kann sich der Gewinn dieser Daten von Gerät zu Gerät unterscheiden. Auch werden für diese Daten zusätzliche Werte ermittelt, die nicht als Metrik in gewöhnlichen Zeitreihen auftreten.

2.3.1 Messung

Der Glukosegehalt im Blut wird anhand von Blutzuckerwerten gemessen, welche in der Regel in Milligramm pro Deziliter (mg/dl) oder in Millimol pro Liter (mmol/l) angegeben werden. In dieser Arbeit erfolgen die Angaben in mg/dl, da solche CGM-Daten vorliegen.

Für die Messung wird das CGM-System am Körper der zu behandelnden Person befestigt und dadurch Messwerte in festen Zeitabständen gewonnen [KAD17]. Des Weiteren sind CGM-Systeme darauf spezialisiert den Glukosegehalt in der Gewebeflüssigkeit (interstitielle Flüssigkeit (ISF)) des Unterhautfettgewebes zu messen, anstatt direkt aus dem Blut [KAD17].

Nach Freckmann [Fre20] unterscheiden sich CGM-Systeme durch Real-Time Continuous Glucose Monitoring (rtCGM) und Intermittently Scanned Continuous Glucose Monitoring (iscCGM). Die rtCGM-Systeme nehmen Messungen zum aktuellen Glukosegehalt in Zeitabständen von fünf Minuten vor, welche direkt angezeigt werden. Bei iscCGM-Systemen hingegen werden alle 15 Minuten ein Wert gespeichert, die spätestens alle acht Stunden eingescannt werden müssen, um diese nicht zu verlieren. Die Übertragung der Ergebnisse erfolgt durch einen Transmitter, welcher sich am Sensor befindet und die Ergebnisse an einen Empfänger/Smartphone zusendet [Fre20].

2.3.2 Auswertung

Battelino et al. [BDB+19] empfehlen für die Auswertung von CGM-Daten standardisierte Verfahren zu verwenden. Zunächst wird für eine optimale Analyse empfohlen, Daten zu verwenden, bei denen über 70 Prozent der Messwerte für mindestens 14 Tage vorliegen. Weiterhin wird ein gemeinsamer durchschnittlicher Glukosewert aus den vorliegenden Tagen angegeben. Als Nächstes wird die Verwendung des Glukose-Management-Indikators (engl. *Glucose Management Indicator* (GMI)) und der glykämischen Variabilität (engl. *Glycemic Variability* (GV)) empfohlen. Bei dem GMI handelt es sich um einen prozentualen Wert, welcher zur Schätzung des HbA1c-Wertes (Hämoglobin A1c) dienen soll [BBC+18]. Dieser wird mit $[3{,}31 + 0{,}02392 \times [\text{Durchschnittswert}]]$ berechnet [BBC+18]. Mit der GV wird die Amplitude, Häufigkeit und Dauer der Schwankungen beschrieben [DNB+17]. Dieser wird wie der Variationskoeffizient mit *[Standardabweichung/Durchschnittswert]* berechnet. Hierfür wird ein Zielwert von $\leq$ 36 Prozent festgelegt. Ferner wird auch eine Einteilung für die Glukosewerte beschrieben, die in der Tabelle 2.2 mit den vorgegebenen Zielwerten veranschaulicht werden. Darin werden fünf Glukosebereiche spezifiziert, welche mit Zeit unterhalb des Zielbereichs (engl. *Time below Range* (TBR)), Zeit im Zielbereich (engl. *Time in Range* (TIR)) und Zeit oberhalb des Zielbereichs (engl. *Time above Range* (TAR)) gekennzeichnet werden. Zusätzlich werden für die Dauer in den jeweiligen Glukosebereichen prozentuale Zielwerte vergeben.

Glukosebereich		Wertebereich	Zielwert
Sehr hoher Bereich	(TAR)	>250 mg/dl	<5%
Hoher Bereich	(TAR)	181-250 mg/dl	<25%
Zielbereich	(TIR)	70-180 mg/dl	>70%
Niedriger Bereich	(TBR)	54-69 mg/dl	<4%
Sehr niedriger Bereich	(TBR)	<54 mg/dl	<1%

Tabelle 2.2: Einteilung der Glukosewerte [BDB+19].

Als primäres Ziel beschreiben Battelino et al. [BDB+19] die Anzahl der Werte im Zielbereich (TIR) zu erhöhen und in den niedrigen Bereichen (TBR) zu senken. Jede fünf prozentige Erhöhung im Zielbereich (TIR) soll dabei für deutliche klinische Vorteile sorgen, und jede zehn prozentige Erhöhung eine 0,5 prozentige Senkung des HbA1c-Wertes bewirken (vgl. Tabelle 2.3) [BDB+19].

Zielbereich (TIR)	HbA1c (Hämoglobin A1c)
20%	9,4%
30%	8,9%
40%	8,4%
50%	7,9%
60%	7,4%
70%	7,0%
80%	6,5%
90%	6,0%

Tabelle 2.3: Geschätzter HbA1c-Wert für den Zielbereich (TIR): Basierend auf Typ-1-Diabetiker/-innen [BDB+19].

3 Verwandte Arbeiten

In diesem Kapitel werden verwandte Arbeiten zu dieser Bachelorarbeit vorgestellt, welche uns bei der Entwicklung des Visualisierungssystems unterstützen. Der Fokus wird dabei insbesondere auf die verwendeten Visualisierungsansätze gelegt. Zunächst werden Visualisierungsansätze zu der Zeitreihenanalyse mit unüberwachtem maschinellen Lernen und anschließend zu den Daten der kontinuierlichen Glukosemessung vorgestellt.

3.1 Zeitreihenanalyse mit unüberwachtem maschinellen Lernen

Für die Analyse und Visualisierung von Zeitreihendaten existieren viele Ansätze, bei denen unüberwachtes maschinelles Lernen angewendet wird. Im Folgenden werden dazu einige Visualisierungsansätze vorgestellt, die durch die Anwendung verschiedener Dimensionsreduktionsverfahren ermöglicht werden.

Bach et al. [BSH+16] stellen einen Visualisierungsansatz für die Analyse von Zeitreihendaten vor. Dieser Ansatz basiert auf der Selbstähnlichkeit der Daten und setzt voraus, dass sich diese in zeitliche Einheiten einteilen lassen. Weiterhin werden sinnvolle Werte benötigt, um Ähnlichkeiten und Unterschiede durch Metriken hervorzuheben. Als Beispiele eignen sich Daten wie Videoaufnahmen, sich entwickelnde Texte, geographische Ereignisse und Gehirnkonnektivitäten. Für die Visualisierung wird zunächst die MDS auf die Daten angewendet. Die daraus resultierenden Ergebnisse werden anschließend als Punkte in einem Streudiagramm veranschaulicht. Für den zeitlichen Überblick werden die Punkte chronologisch durch Bézierkurven miteinander verbunden. Darüber hinaus werden die Punkte durch Farben mit verschiedenen Helligkeiten versehen, wobei dunklere Punkte als spätere Zeiten interpretiert werden. Außerdem lassen sich durch deren Größe die ungefähren Werte herleiten. Des Weiteren können durch die Lage der Punkte und deren Verbindungen zueinander bestimmte Muster erkannt werden, die beispielsweise Ansammlungen, große Übergänge oder Ausreißer aufzeigen. Naheliegende Punkte kennzeichnen dabei die Ähnlichkeit und distanzierte Punkte die Unähnlichkeit zueinander.

Mit einem ähnlichen Visualisierungsansatz haben sich Hinterreiter et al. [HSS+20] beschäftigt. Dieser Ansatz ermöglicht die Musteranalyse anhand von Entscheidungsvorgängen, die durch Menschen oder Algorithmen ausgeführt werden. Dafür werden Entscheidungsvorgänge verwendet, welche auf Domänen wie Logik-Puzzle, Strategiespielen und Prozessoptimierungen basieren. Diese gehen von einem Startzustand aus und enthalten verschiedene Zustände, die anhand der Entscheidungen besucht werden bis ein Endzustand erreicht wird. Aufgrund der vielen Entscheidungsmöglichkeiten, die zu hohen Dimensionen führen, wird hierfür eine Dimensionsreduktion angewandt, um diese auf einen zweidimensionalen Bereich zu führen. Der zu diesem Ansatz entwickelte Prototyp „ProjectionPathExplorer“ unterstützt dabei als Dimensionsreduktionsverfahren t-SNE und UMAP. Für die Visualisierung werden die Zustände, die als Resultat aus der Dimensionsreduktion hervorgehen, als

Punkte repräsentiert und ähnlich wie in den „TimeCurves“ [BSH+16] mit Bézierkurven miteinander verbunden. Zusätzlich können die Punkte durch ihre Formen, Farben und Größen, kategorische oder quantitative Metadaten kodieren, welche sich auch durch Farben der Bézierkurven ausdrücken lassen. Mittels der Positionierung des Mauszeigers auf einen Punkt werden Tooltips zu diesem angezeigt. Es bietet sich zudem die Möglichkeit Informationen durch Filterungen der Daten zu veranschaulichen. Die Erkennung von Mustern erfolgt anhand der Positionen der Punkte und den dazugehörigen Richtungen der Kurven. In ihrer Fallstudie werden Entscheidungsvorgänge anhand von Zauberwürfel, Schachspielen und neuronalem Netzwerk-Training entnommen. Insgesamt hat sich der Visualisierungsansatz als nützlich erwiesen und ist auf weitere Domänen erweiterbar.

Der Ansatz von Ali et al. [AJXW19] ermöglicht es mithilfe der Dimensionsreduktion univariate und multivariate Zeitreihen auf Muster und Ausreißer zu untersuchen. Dafür werden zuerst die Zeitreihendaten vorverarbeitet. Hierzu werden die Werte der Zeitreihen auf das Intervall *[0,1]* normalisiert. Anschließend folgt der „Sliding Window Ansatz“, mit dem univariate und multivariate Zeitreihen auf die Dimensionsreduktion vorbereitet werden. Der „Sliding Window Ansatz“ wird für die Glättung der Zeitreihendaten benötigt, sodass bei univariaten Zeitreihen höhere Dimensionen entstehen, wodurch eine Dimensionsreduktion ermöglicht wird. Als Dimensionsreduktionsverfahren werden PCA, t-SNE und UMAP angewendet. Visualisiert wird auf einem Streudiagramm, bei dem die dargestellten Punkte miteinander verbunden werden. Ausstechende Punkte und Verbindungen deuten hierbei auf mögliche Muster und Ausreißer in den Daten hin. Jener Ansatz wurde in Form zweier Fallstudien mit dem dazu entwickelten System evaluiert. Die Ergebnisse haben gezeigt, dass dieser Visualisierungsansatz effektiv zur Erkennung von Mustern und Ausreißern beitragen kann. Insbesondere konnten durch die Verwendung von UMAP gute Ergebnisse erzielt werden.

Der Visualisierungsansatz von van den Elzen et al. [EHBW16] untersucht und analysiert dynamische Netzwerke. Dabei werden die Momentaufnahmen und Entwicklungen der Netzwerke visualisiert. Folglich ermöglicht sich die Analyse von stabilen Zuständen, wiederkehrenden Zuständen und Ausreißern. Zunächst wird der „Sliding Window Ansatz“, bei dem die Daten, ähnlich wie bei den „TimeCluster“ [AJXW19], für die Dimensionsreduktion geglättet werden, verwendet. Als Dimensionsreduktionsverfahren werden PCA, MDS und t-SNE eingesetzt. Die Zustände werden hierbei als Punkte dargestellt. Naheliegende Punkte kennzeichnen die Ähnlichkeit dieser Zustände. Deren Evolution wird durch Linien repräsentiert, welche die Punkte miteinander verbinden. Die Farben der Punkte werden dabei, ähnlich wie in den „TimeCurves“ [BSH+16], zur zeitlichen Orientierung eingesetzt. Diese werden zudem als Indikator für die Musteranalyse verwendet, womit Informationen über die Dauer des Netzwerkes in einem stabilen Zustand erfasst werden können.

Bernard et al. [BWS+12] stellen einen Visualisierungsansatz zur Analyse von multivariaten Zeitreihen vor. Dafür werden zunächst Dimensionsreduktionsverfahren wie PCA, MDS und Selbstorganisierende Karte (engl. *Self-Organizing Map* (SOM)) auf die Daten angewendet, damit diese in einem zweidimensionalen Raum dargestellt werden können. Nach dem temporären Verbinden der daraus resultierenden angrenzenden Punkte werden diese als „TimeSeriesPaths“ bezeichnet. Um eine bessere Analyse zu ermöglichen, werden die entsprechenden Zeitreihen weiter durch aggregierte Daten, Cluster Visualisierungen und Mehrfachansichten ergänzt. Zusätzlich bieten sich Interaktionsmöglichkeiten wie Auswählen, Gruppieren und Hervorheben dieser Punkte an. Weiterhin wird die Visualisierung durch Tooltips und Zeitschieberegler unterstützt. Der Visualisierungsansatz wurde mithilfe einer Fallstudie bewertet. Dazu wurden Wetterdaten auf Muster und Ausreißer untersucht. Insgesamt konnte festgestellt werden, dass sich der Visualisierungsansatz für die Analyse von multivariaten Zeitreihen eignet.

In unserer Arbeit werden wie Zeitreihen aufgebaute CGM-Daten analysiert. Die vorgestellten Visualisierungsansätze für Zeitreihendaten eignen sich demnach auch für unsere Visualisierungen. Es werden auch Vorverarbeitungsmöglichkeiten demonstriert, mit denen sich univariate Zeitreihen in multivariate Zeitreihen umformen lassen. Das hat zur Folge, dass sich die Dimensionsreduktion auch auf die CGM-Daten anwenden lässt. Dadurch ergibt sich die Möglichkeit einer Analyse auf Muster und Ausreißer.

3.2 Visualisierung von Daten der kontinuierlichen Glukosemessung

Danne et al. [DNB+17] empfehlen im Falle einer Visualisierung, sich an das Ambulante Glukoseprofil (engl. *Ambulatory Glucose Profile* (AGP)) zu halten. Dabei handelt es sich um eine standardisiertes Verfahren zur Auswertung von CGM-Daten, dessen Ursprung von Mazze et al. [MLL+87] stammt und heute vom International Diabetes Center (IDC) als AGP-Bericht [Int21] benutzt wird. Dieser beschreibt einen einseitigen Bericht, in dem Statistiken und Graphen zum gesamten Zeitraum der CGM-Daten aufgeführt werden.

In der Abbildung 3.1 wird der obere Bereich eines AGP-Berichts dargestellt. Darin werden die im Abschnitt 2.3.2 beschriebenen Statistiken zur Auswertung der CGM-Daten veranschaulicht. Weiterhin enthalten diese auch die einzelnen Wertebereiche der Glukosewerte. Diese werden hier eingesetzt, um die Dauer in den Glukosebereichen in Prozent und verschiedenen Zeiteinheiten zu beschreiben.

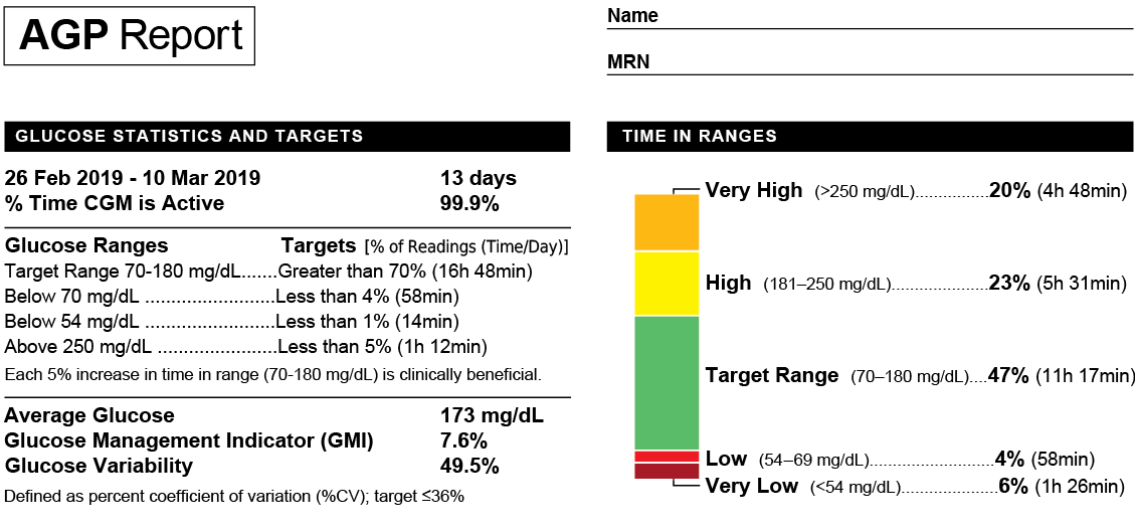


Abbildung 3.1: Ambulantes Glukoseprofil - Oberer Bereich: (Links) Statistiken zu den CGM-Daten. (Rechts) Glukosebereich mit Angaben in Prozent und Zeit zur Auswertung der CGM-Daten [Int21].

Der mittlere Bereich des AGP-Berichts wird in der Abbildung 3.2 dargestellt. In dieser wird ein Glukoseverlauf über 24 Stunden visualisiert, der durch die Kombination der vorliegenden Tage entstanden ist. Die grünen Linien werden hier eingesetzt um den Zielbereich (TIR) zu repräsentieren. Die schwarze Linie stellt den Median dar, welcher angibt, dass sich die Hälfte der Werte über und

unter diesem befinden [Int21]. Der dunkelblaue Bereich wird aus dem ersten und dritten Quartil gebildet und drückt damit 50 Prozent der Werte aus, während der hellblaue Bereich 90 Prozent ausdrückt [Int21].

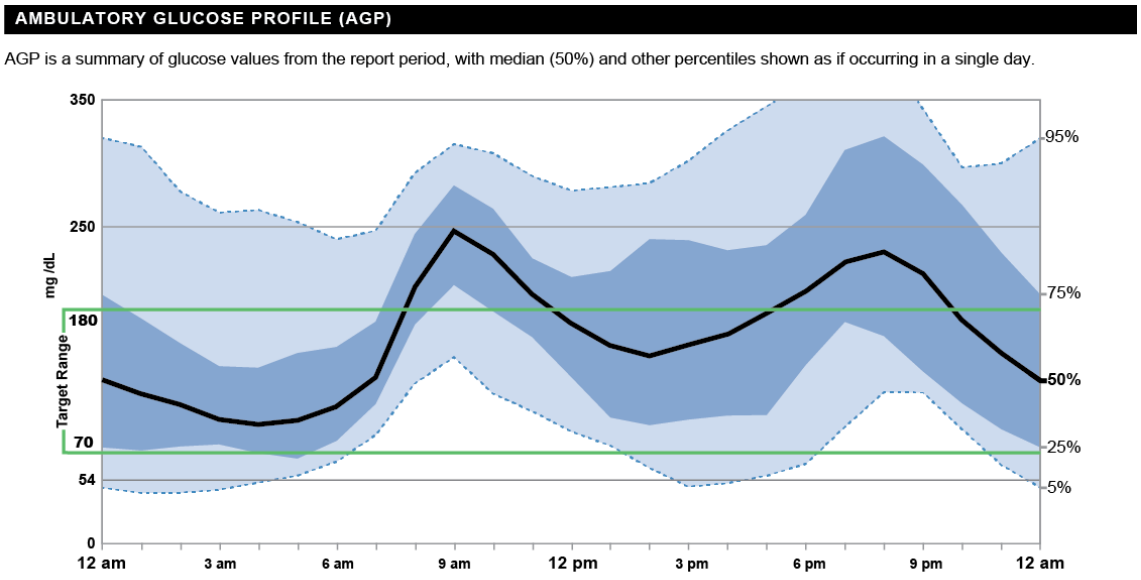


Abbildung 3.2: Ambulantes Glukoseprofil - Mittlerer Bereich: Gemeinsamer 24 Stunden Glukoseverlauf der vorliegenden Tage. (Schwarz) Median. (Dunkelblau) 50 Prozent der Werte. (Hellblau) 90 Prozent der Werte. (Grün) Zielbereich (TIR) [Int21].

In der Abbildung 3.3 wird der untere Bereich des AGP-Berichts abgebildet. Dieser visualisiert die vorliegenden Tage in einzelnen kleinen Liniendiagrammen. Dabei werden mit den Farben mögliche Nebenwirkungen visualisiert. Die gelben Bereiche deuten auf mögliche erschwerte Heilungen von Infektionen und die roten Bereiche auf mögliche Schwächegefühle hin [Int21].

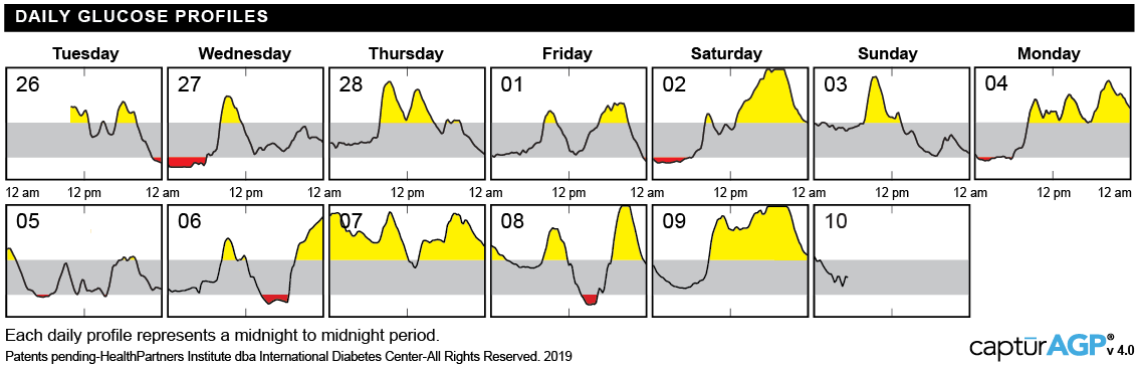


Abbildung 3.3: Ambulantes Glukoseprofil - Unterer Bereich: Darstellung der einzelnen Tage in seperaten Liniendiagrammen. (Gelb) Mögliche erschwerte Heilungen von Infektion. (Rot) Mögliche Schwächegefühle [Int21].

4 Visualisierungsansatz

In diesem Kapitel wird der verwendete Visualisierungsansatz präsentiert. Zunächst werden dafür die funktionalen Anforderungen und ein möglicher Arbeitsablauf des Visualisierungssystems vorgestellt. Hinterher wird auf die Vorverarbeitung der CGM-Daten eingegangen. Anschließend erfolgt eine detaillierte Erklärung zu den Visualisierungen. Zum Schluss werden die grafische Benutzeroberfläche des Visualisierungssystems und die Interaktionsmöglichkeiten zu den Visualisierungen präsentiert.

4.1 Funktionale Anforderungen

- Das System soll die Möglichkeit besitzen, verschiedene Zeitabschnitte (Mittagszeit/Abendzeit)/Stunden/Wochentage/Wochen/Monate/... auf Muster und Ausreißer zu vergleichen.
- Das System soll insbesondere zur Unterscheidung von ähnlichen und verschiedenen Werten in den Zeitabschnitten eingesetzt werden.
- Das System soll Änderungswerte auf Muster und Ausreißer untersuchen können.
- Das System soll durch Visualisierungen Rückschlüsse auf Therapieanpassungen ermöglichen.
- Das System soll die Zeiten, an denen der/die Patient/Patientin am besten eingestellt oder ähnlich zu einem anderen Zeitpunkt ist, ersichtlich darstellen.

4.2 Arbeitsablauf

In der Abbildung 4.1 wird ein möglicher Arbeitsablauf des Visualisierungssystems vorgestellt. Zunächst wird die CGM-Datei in das Visualisierungssystem geladen. Anschließend werden Daten- und Systemeinstellungen vorgenommen. Als Nächstes folgt die Auswertung der zugehörigen Statistiken. Hinterher werden die Visualisierungen auf interessante Muster und Ausreißer untersucht. Darauf basierend erfolgt die Auswahl und Analyse von interessanten Datenpunkten. Die einzelnen Schritte lassen sich dabei mit neuen Anpassungen wiederholen.

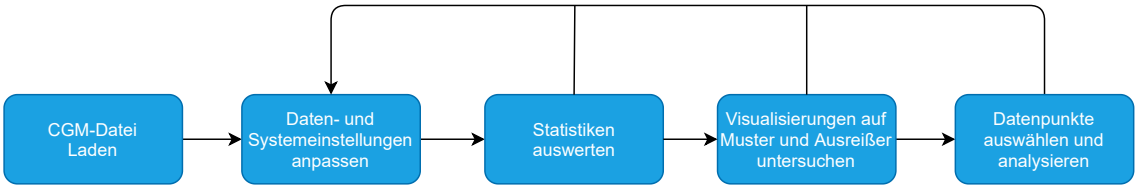


Abbildung 4.1: Ein möglicher Arbeitsablauf des Visualisierungssystems.

4.3 Vorverarbeitung

Die Vorverarbeitung der CGM-Daten setzt sich aus der Fehlerbehandlung und der Vorbereitung auf die Dimensionsreduktion zusammen. Um sich ein Bild über die vorliegenden CGM-Daten machen zu können, werden in den Tabellen 4.1a und 4.1b CGM-Dateien mit möglichen Beispielwerten dargestellt. Die Tabelle 4.1a veranschaulicht eine ideale CGM-Datei bei denen alle Zeitstempel im selben ISO 8601[1] Format in Zeitabständen von fünf Minuten mit Messwerten vorliegen, während in der Tabelle 4.1b sowohl fehlende als auch doppelte Messwerte sowie Zeitstempel mit unterschiedlichen Formaten enthalten sind.

Zeitstempel	Wert
2019-04-27T23:55:00.000Z	145
2019-04-27T23:50:00.000Z	144
2019-04-27T23:45:00.000Z	143
...	...
2018-12-12T00:10:00.000Z	114
2018-12-12T00:05:00.000Z	115
2018-12-12T00:00:00.000Z	113

(a)

Zeitstempel	Wert
04/27/2019 23:56:09	145
2019-04-27T23:56:09.000Z	145
04/27/2019 23:46:09	143
...	...
2018-12-12T00:06:35.000Z	115
12/12/2018 00:01:34	null
2018-12-12T00:01:34.000Z	null

(b)

Tabelle 4.1: Beispiel einer CGM-Datei: (a) Zeitstempel in selben Formaten; (b) Zeitstempel in unterschiedlichen Formaten und fehlende und doppelte Zeitstempel sowie Messwerte.

4.3.1 Fehlerbehandlung

Damit eine fehlerfreie Anzeige der CGM-Daten ermöglicht wird, werden diese zu Beginn auf Korrektheit überprüft und angepasst. Dazu werden zunächst die Formate der Zeitstempel inspiziert und auf ein einheitliches Format gebracht. Dies dient zum einen für spätere Filterungsverfahren, und zum anderen für die Beseitigung doppelter Zeitstempel. Hinterher werden die einzelnen Einträge nach den Zeitstempel aufsteigend sortiert. Anschließend folgt die Überprüfung der Werte auf mögliche Messfehler. Dabei werden gefundene Fehlerwerte vorerst eliminiert und durch ungültige Werte ersetzt. Hiermit wird ein möglicher Verlust des Zeitstempels sichergestellt. Danach wird überprüft, ob jeder Tag im vorliegenden Zeitraum 288 Zeitstempel in Zeitabständen von fünf Minuten besitzt. Fehlende Zeitstempel werden dabei vorerst mit einem zugehörigen ungültigen Wert ergänzt. Weiterhin werden alle Zeitstempel auf 0 bis 23:55 Uhr in Zeitabständen von fünf Minuten zugeordnet. In der Tabelle 4.1b werden dadurch beispielsweise 00:06:34 zu 00:05:00 Uhr und 23:56:09 zu 23:55:00 Uhr geändert. Damit sollen Vergleiche beliebiger Zeitabschnitte zu den selben Zeitpunkten ermöglicht werden. Nach Abschluss lassen sich die ungültigen Werte mithilfe der umgebenden Werte linear interpolieren. In der Tabelle 4.2a werden die Daten nach der Fehlerbehandlung veranschaulicht. Zu erkennen ist, dass in der ersten Zeile noch ein ungültiger Wert enthalten ist. Dies liegt daran, dass kein Zeitstempel zu einem vorherigen Zeitpunkt existiert. Daher ist in diesem Fall keine lineare Interpolation für diesen fehlenden Wert möglich.

[1] https://www.iso.org/iso-8601-date-and-time-format.html

4.3.2 Vorbereitung auf die Dimensionsreduktion

Um die Dimensionsreduktion auf einer univariaten Zeitreihe auszuführen, muss die Zeitreihe zunächst in eine multivariate Zeitreihe bzw. hochdimensionale Matrix überführt werden. Dafür wird ein ähnlicher Ansatz wie der „Sliding Window Ansatz“ [AJXW19] verwendet. Aufgrund der gleichen Zeitabstände, die bei den CGM-Daten vorliegen, lassen sich die Zeitreihendaten in verschiedene Zeitabschnitte einteilen. Für die Einteilung in Stunden werden dabei zwölf aufeinanderfolgende Werte aus den Zeitreihendaten entnommen und jeweils in der selben Zeile der neuen Matrix eingefügt. Durch das Wiederholen dieses Vorgangs entstehen mehrere Zeilen, die jeweils eine Stunde pro Zeile repräsentieren. Für größere Zeitabschnitte werden entsprechend mehr Spalten für mehr Werte pro Zeile benötigt. Bei Tagen entspricht das 288 Werten, bei Wochen 2016 Werten und bei Monaten 8064 bis 8928 Werten. Die monatliche Variierung der Werte liegt an den unterschiedlichen Größen der Monate. Daher können einige Monate mehr Werte pro Zeile enthalten als andere. Hierzu kann im Nachhinein ein Filterungsverfahren eingesetzt werden, um diese auf dieselbe Dimensionalität zu skalieren. Nach der Ausführung aller Schritte wird schließlich eine hochdimensionale Matrix geschaffen, auf der sich die Dimensionsreduktion ausführen lässt. Das Resultat der Vorverarbeitung wird in der Tabelle 4.2b veranschaulicht. Darin werden die einzelnen Tage mit ihren zugehörigen 288 Werten jeweils in einer Zeile dargestellt. Für die Dimensionsreduktion werden nur Zeilen verwendet, die mit Zahlenwerten gefüllt sind. Zeilen, die demnach ungültige Werte enthalten, werden ausgeschlossen oder mithilfe von Filterungsverfahren einbezogen.

Zeitstempel	Wert
2018-12-12 00:00:00	null
2018-12-12 00:05:00	115
2018-12-12 00:10:00	114
...	...
2019-04-27 23:45:00	143
2019-04-27 23:50:00	144
2019-04-27 23:55:00	145

(a)

	00:00:00	00:05:00	00:10:00	...	23:45:00	23:50:00	23:55:00
2018-12-12	null	115	114	...	109	110	109
2018-12-13	109	111	112	...	128	127	130
2018-12-14	132	133	133	...	115	113	116
...	...	...	...	...	...	...	...
2019-04-25	140	137	138	...	127	128	130
2019-04-26	131	132	131	...	135	133	132
2019-04-27	130	128	127	...	143	144	145

(b)

Tabelle 4.2: Vorverarbeitung der CGM-Daten: (a) Nach der Fehlerbehandlung: Fehlende Zeitstempel und Messwerte ergänzt, sortiert und interpoliert; (b) Nach der Vorbereitung auf die Dimensionsreduktion: Tage in Zeilen und Uhrzeiten in Spalten unterteilt.

4.4 Visualisierungen

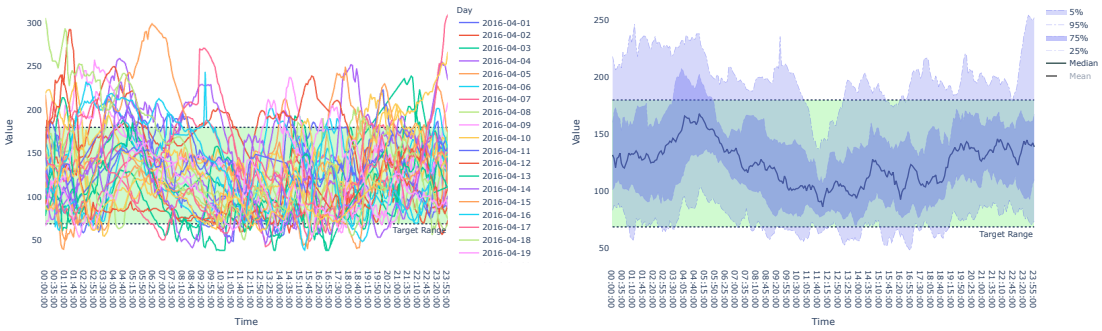
Für die Visualisierung der CGM-Daten werden Kastengrafiken, Heatmaps, Linien-, Säulen-, Kreis- und Streudiagramme verwendet. Im Nachfolgenden werden deren Einsatz in unserem Visualisierungssystem näher vorgestellt. Dabei werden des Öfteren die gleichen Farben in den Visualisierungen auftreten. Diese sind für die Visualisierung der einzelnen Glukosebereiche gedacht. Die Tabelle 4.3 stellt diesbezüglich die Zuordnung der einzelnen Farben mit den zugehörigen Wertebereichen dar.

Glukosebereich		Wertebereich	Farbe
Sehr hoher Bereich	(TAR)	>250	Rot
Hoher Bereich	(TAR)	180-250	Orange
Zielbereich	(TIR)	70-180	Grün
Niedriger Bereich	(TBR)	54-70	Hellblau
Sehr niedriger Bereich	(TBR)	<54	Blau

Tabelle 4.3: Farben zur Visualisierung der einzelnen Glukosebereiche.

4.4.1 Liniendiagramm

Das Liniendiagramm liefert einen Überblick über den zeitlichen Verlauf der gemessenen Werte. Für die gleichzeitige Darstellung und Unterscheidung mehrerer Linien werden verschiedene Farben (vgl. Abbildung 4.2a) verwendet. Wie im AGP-Bericht [Int21] wird der definierte Zielbereich (TIR) für den Wertebereich von 70 bis 180 grün abgebildet. Nach dem AGP-Bericht besteht auch die Möglichkeit, mehrere Linien durch statistische Werte zu visualisieren, die sich aus den Kombinationen der Tage ergeben. Hierzu können neben Tagen auch Stunden, Wochen und Monate visualisiert werden. In der Abbildung 4.2b wird dieser Visualisierungsansatz mit Tagen verdeutlicht. Dabei wird der Median durch die dunkelblaue Linie ausgedrückt. 50 Prozent der Werte werden durch den blauen und 90 Prozent der Werte durch den hellblauen Bereich visualisiert.



(a) **(b)**

Abbildung 4.2: Liniendiagramm des Visualisierungssystems: (a) Darstellung multivariater Zeitreihen; (b) Darstellung multivariater Zeitreihen mithilfe statistischer Werte.

4.4.2 Gestapeltes Säulendiagramm

Das gestapelte Säulendiagramm ermöglicht die Darstellung einzelner Zeitabschnitte als Säulen. Die einzelnen Stapel beschreiben dabei, ähnlich wie im AGP-Bericht [Int21], die Anzahl der Werte in den jeweiligen Glukosebereichen. Diese werden hierbei allerdings eingesetzt, um einzelne Zeitabschnitte wie Stunden, Tage, Wochen oder Monate zu beschreiben, anstatt den gesamten Zeitraum. Dieser Visualisierungsansatz wird in der Abbildung 4.3 mit 30 Tagen zur Schau gestellt.

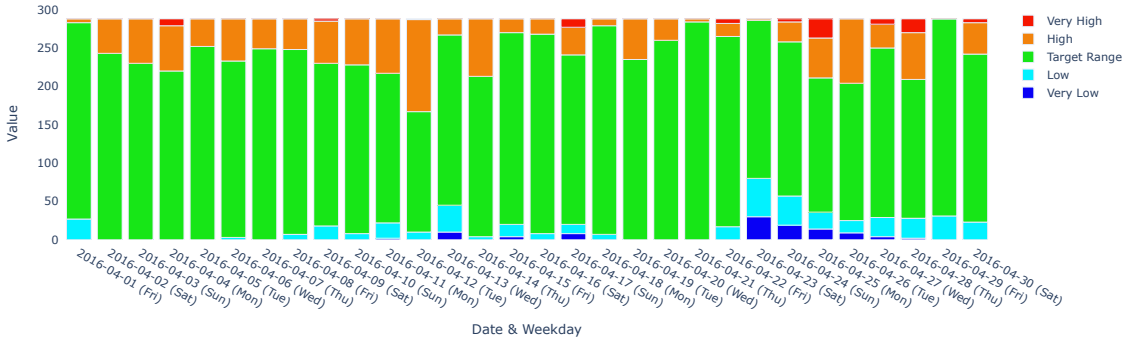


Abbildung 4.3: Gestapeltes Säulendiagramm des Visualisierungssystems: Darstellung von 30 Tagen mit der Anzahl ihrer Glukosewerte in den jeweiligen Glukosebereichen.

4.4.3 Streudiagramm

Im Streudiagramm (vgl. Abbildung 4.4) wird das Resultat der Dimensionsreduktion visualisiert. Darin werden die einzelnen Zeitabschnitte als Punkte dargestellt. Die Größe der Punkte repräsentiert dabei jeweils den Durchschnittswert des jeweiligen Zeitabschnitts. Größere Punkte deuten dabei auf höhere Werte. Um einen zeitlichen Überblick zu erhalten, werden die Punkte, ähnlich wie im Ansatz von Bach et al. [BSH+16], durch Farben mit verschiedenen Helligkeiten versehen und in chronologischer Reihenfolge mit Kurven verbunden. Hellere Farben kennzeichnen hierbei frühere und dunklere Farben spätere Zeiten. Weiterhin wird zur Orientierung der Start- und Endpunkt mit einem blauen Rand versehen. Als Kurven werden Spline-Kurven verwendet.

Für die Analyse der CGM-Daten werden dem Streudiagramm fünf weitere Elemente hinzugefügt, mit denen die Glukosebereiche repräsentiert werden. Dadurch sollen einzelne Zeitabschnitte besser zu den Glukosebereichen zugeordnet werden, da die Achsen nach der Dimensionsreduktion nicht mehr aussagefähig sind. Zur besseren Unterscheidung von den Zeitabschnitten werden für die Glukosebereiche in der Darstellung Kreuze statt Punkte verwendet. Diese werden jeweils mit den Farben der Glukosebereiche versehen. Die Position der Kreuze wird durch die Dimensionsreduktion festgelegt. Hierzu werden die Glukosebereiche ähnlich wie die Zeitabschnitte behandelt und als weitere Zeilen mit in die Matrix der Dimensionsreduktion (vgl. Tabelle 4.2b) aufgenommen. Als Werte der einzelnen Spalten werden dabei jeweils die Mittelwerte der Glukosebereiche (vgl. Tabelle 4.3) vergeben. Nach der Ausführung eines Verfahrens zur Dimensionsreduktion werden letztlich die Koordinaten der Kreuze ermittelt.

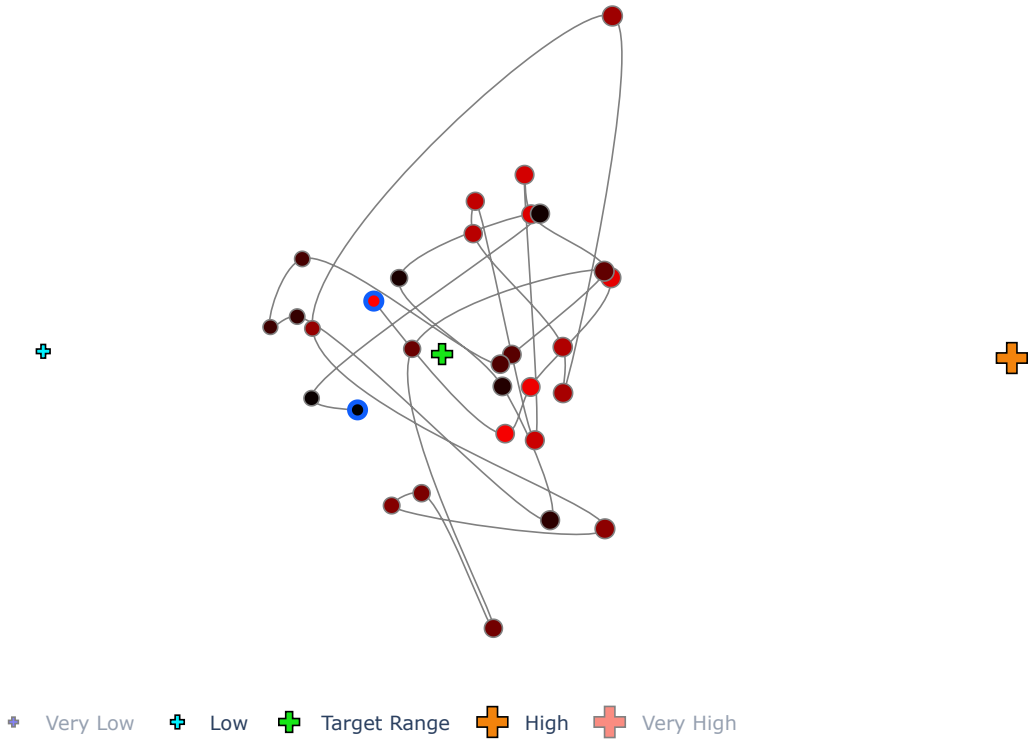


Abbildung 4.4: Streudiagramm des Visualisierungssystems: Punkte stellen einzelne Zeitabschnitte dar. Die Farben (rot nach schwarz) repräsentieren den Zeitverlauf. Blaue Ränder geben Start- und Endpunkt an. Kreuze stellen Glukosebereiche in ihren Farben dar.

4.4.4 Kreisdiagramm

Mit dem Kreisdiagramm (vgl. Abbildung 4.5) werden die prozentualen Verteilungen der vorliegenden Werte in den jeweiligen Glukosebereichen beschrieben.

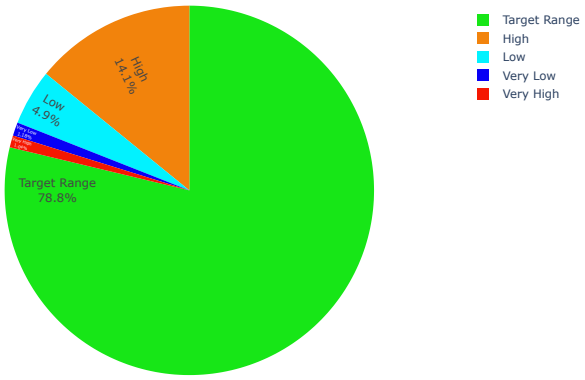


Abbildung 4.5: Kreisdiagramm des Visualisierungssystems: Prozentuale Verteilungen über die Anzahl der Messwerte in den jeweiligen Glukosebereichen.

4.4.5 Kastengrafik

Kastengrafiken werden für die Visualisierung einzelner Zeitabschnitte eingesetzt. Dabei können Stunden, Tage, Wochen oder Monate als einzelne Kasten dargestellt werden und zugehörige statistische Werte beschreiben. Weiterhin wird den Kasten eine gestrichelte Linie beigefügt, durch den der Verlauf der Medianwerte veranschaulicht wird. Für die Analyse der CGM-Daten wird ähnlich wie im Liniendiagramm der definierte Zielbereich (TIR) dargestellt. Mit den Farben der Kasten werden Zeitabschnitte ausgedrückt, bei denen über 70 Prozent der Messwerte im Zielbereich (TIR) liegen (grün) bzw. nicht darin liegen (grau). Die Abbildung 4.6 zeigt hierzu ein Beispiel, welcher 30 Tage als Kasten darstellt.

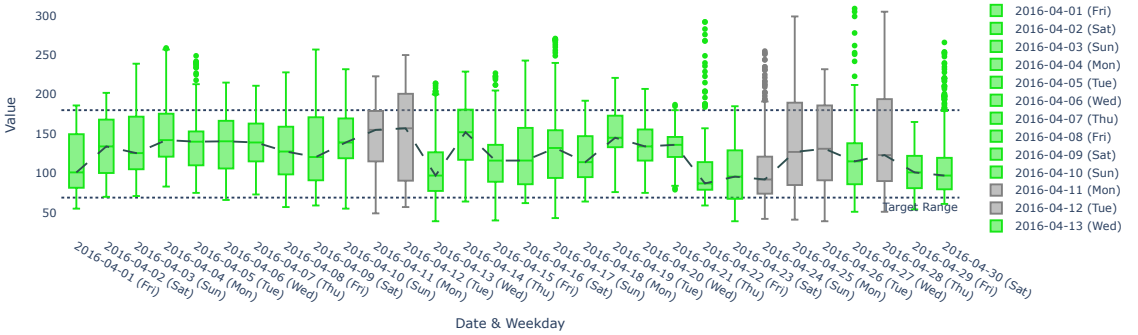


Abbildung 4.6: Kastengrafik des Visualisierungssystems: Es werden 30 Tage als Kasten dargestellt. Zielbereich (TIR) über 70 Prozent (grün) und unter 70 Prozent (grau).

4.4.6 Heatmap

Heatmaps werden ähnlich wie gestapelte Säulendiagramme eingesetzt. Der Unterschied hierbei ist, dass die vertikale Achse je nach ausgewähltem Zeitabschnitt mit Uhrzeiten oder Tagen versehen wird. Dadurch werden die einzelnen Glukosewerte über den gesamten zeitlichen Abschnitt ersichtlich. Dies soll bei der Analyse zur Durchführung eines Vergleiches über Uhrzeiten oder Tage in den ausgewählten Zeitabschnitten dienen. In der Abbildung 4.7 wird dieser Ansatz mit 30 Tagen ausgeführt, worin Messwerte zu unterschiedlichen Uhrzeiten dargestellt werden.

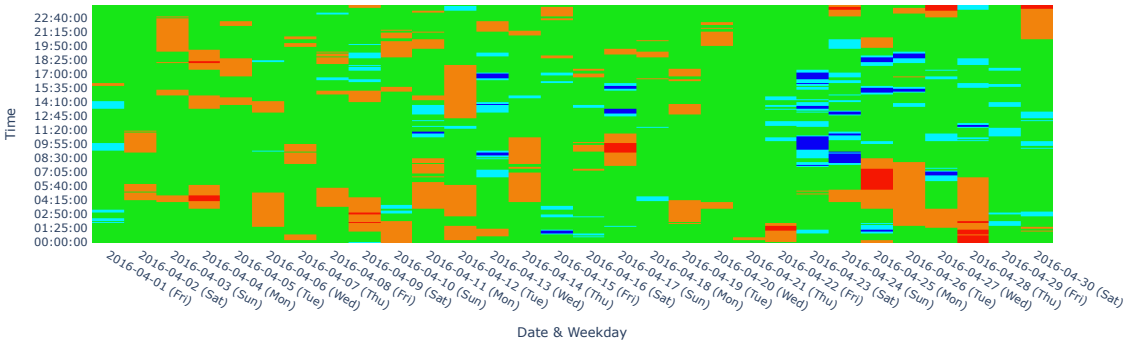


Abbildung 4.7: Heatmap des Visualisierungssystems: Darstellung von 30 Tagen. Die Farben repräsentieren die Glukosebereiche zu verschiedenen Uhrzeiten.

4.5 Grafische Benutzeroberfläche

Das Visualisierungssystem wird als Single-Page-Webanwendung (engl. *Single-Page Application* (SPA)) ausgeführt und setzt sich aus fünf verschiedenen Bereichen zusammen. Diese Bereiche bestehen aus dem Konfigurations-, Informations-, Übersichts-, Dimensionsreduktions- und Selektionsbereich (vgl. Abbildung 4.8). Im Nachfolgenden werden die einzelnen Bereiche näher vorgestellt. Dabei werden auch die zugehörigen Funktionen beschrieben.

Abbildung 4.8: Grafische Benutzeroberfläche des Visualisierungssystems: (a) Konfigurationsbereich; (b) Informationsbereich; (c) Übersichtsbereich; (d) Dimensionsreduktionsbereich; (e) Selektionsbereich.

4.5.1 Konfigurationsbereich

Den Anfang stellt der Konfigurationsbereich (vgl. Abbildung 4.9) des Visualisierungssystems dar. Dieser Bereich beinhaltet allgemeine Einstellungen zur Anpassung der CGM-Daten.

Abbildung 4.9: Konfigurationsbereich des Visualisierungssystems: (a) Datei laden; (b) Datumsfenster; (c) Wertebereich der Glukosebereiche; (d) Farben der Glukosebereiche; (e) Interpolationsgrenze; (f) Zielwert für den Zielbereich (TIR).

Mit der ersten Einstellung (vgl. Abbildung 4.9a) können die CGM-Daten in das System geladen werden. Dieser ist mit CGM-Dateien im CSV-Format kompatibel und setzt voraus, dass sie wie univariate Zeitreihen aufgebaut sind. Während dem Startvorgang des Visualisierungssystems wird im lokalen Projektordner nach CSV-Dateien gesucht und bei Kompatibilität im Dropdown-Listenfeld angezeigt. Als Standardeinstellung wird die erste gefundene Datei geladen. Diese kann nachfolgend im Dropdown-Listenfeld angepasst werden.

Nachdem die CSV-Datei geladen wurde, wird im *Datumsfenster* (vgl. Abbildung 4.9b) der entsprechende Zeitraum der vorliegenden Daten angezeigt. Bei der Anzeige ist zu beachten, dass das Enddatum nicht zu den Messdaten dazugehört, sondern ausgeschlossen wird, damit einzelne Tage ebenfalls untersucht werden können. Durch die Interaktion mit der Maus lässt sich der Zeitraum nachfolgend anpassen. Dabei wird ein kleines Fenster mit einem Kalender aufgerufen, mit dem sich Start- und Enddatum des Zeitraums bestimmen lassen.

Weiterhin können die Wertebereiche der einzelnen Glukosebereiche genauer spezifiziert (vgl. Abbildung 4.9c) werden. Hierzu wird ein Wertebereich von 0 bis 500 vorgegeben. Als Standardeinstellung werden die Wertebereiche nach den Vorgaben des AGP-Berichts [Int21] vergeben. Dabei werden, anders als im AGP-Bericht, aufgrund der Dezimalzahlen, die bei der linearen Interpolation entstehen, Intervalle zur Spezifizierung der Wertebereiche verwendet. Diese sind wie folgt eingeteilt: *[0, 54[*, *[54, 70[*, *[70, 180[*, *[180, 250[* und *[250, 500]*. Bei Bedarf lässt sich auch die untere Grenze des sehr niedrigen Bereichs bzw. die obere Grenze des sehr hohen Bereichs festlegen.

Die fünf Farben der Glukosebereiche können ebenfalls angepasst (vgl. Abbildung 4.9d) werden. Es stehen dafür verschiedene konfigurierte Farben zur Auswahl. Neben der beliebigen Anpassung sollen diese außerdem auch Menschen, denen manche Farben eventuell nicht ersichtlich sind, bei der Visualisierung unterstützen.

Für die Behandlung der Fehlerwerte wird eine Einstellung zur linearen Interpolation der Werte (vgl. Abbildung 4.9e) vorgegeben. Damit lässt sich eine Grenze für die Anzahl der zu interpolierenden Messwerte festlegen. Zur Einfachheit werden diese als Zeiteinheiten angegeben. Standardmäßig werden bis zu drei Stunden linear interpoliert. Auch kann entschieden werden, ob keine oder

alle fehlenden Werte interpoliert werden sollen. Die Auswahl keiner Interpolation stellt dabei rohe Messdaten dar. Dies kann unter anderem bei der Dimensionsreduktion dazu führen, dass unvollständige Zeitabschnitte aufgrund kleinerer Dimensionen herausgefiltert werden. Je nach beobachtetem Zeitraum empfiehlt es sich deshalb größere Interpolationsgrenzen auszuwählen, damit auch die Dimensionsreduktion auf ganze Wochen oder Monate angewendet werden kann.

Als letzte Einstellung (vgl. Abbildung 4.9f) lässt sich der prozentuale Zielwert für die Anzahl der Messwerte im Zielbereich (TIR) bestimmen. Als Basiswert werden nach den Vorgaben von Battelino et al. [BDB+19] 70 Prozent angenommen. Demnach wirkt sich die Einstellung auch auf die Farben der Kastengrafiken (vgl. Abbildung 4.4.5) aus.

4.5.2 Informationsbereich

Im Informationsbereich (vgl. Abbildung 4.10) werden für den ausgewählten Zeitraum Informationen zu den CGM-Daten angezeigt. Dafür wird eine Informationstabelle mit einem Kreisdiagramm bereitgestellt.

Area	Very Low 0-54	Low 54-70	Target Range 70-180	High 180-250	Very High 250-500	Total
Measurements	124 (0.71%)	692 (3.94%)	13984 (79.65%)	2465 (14.04%)	291 (1.66%)	17556 (98.32%)
Time in Ranges	10:20:00	2 days, 9:40:00	48 days, 13:20:00	8 days, 13:25:00	1 day, 0:15:00	60 days, 23:00:00
GMI	4.44%	4.83%	6.2%	8.21%	10.11%	6.48%
GV	9.52%	6.64%	24.53%	9.03%	11.02%	35.66%
Min	39.0	54.0	70.0	180.0	250.0	39.0
Max	53.0	69.0	179.81	249.0	381.0	381.0
Mean	47.29	63.45	120.63	204.77	284.12	132.38
Median	49	64	120	201.5	275	126
Standard Deviation	4.5	4.21	29.59	18.49	31.31	47.21

Abbildung 4.10: Informationsbereich des Visualisierungssystems: (a) Informationstabelle; (b) Kreisdiagramm.

Mithilfe der Informationstabelle (vgl. Abbildung 4.10a) werden die Messwerte der geladenen CGM-Daten für den gesamten Zeitraum in aggregierter Form veranschaulicht. Dazu werden verschiedene statistische Werte für den gesamten Bereich und für die einzelnen Glukosebereiche ermittelt. Zunächst werden darin die Anzahl der Messungen als Zahlen und Prozente dargestellt. Mit dem Prozentwert des gesamten Bereichs wird die Anzahl der Messwerte im Verhältnis zum ausgewählten Zeitraum beschrieben. Ein prozentualer Wert von unter 100 Prozent würde dementsprechend auf fehlende Messwerte deuten. Der Prozentwert der einzelnen Glukosebereiche basiert auf den vorhandenen Messwerten. Weiterhin wird die Dauer über den Aufenthalt in den jeweiligen Bereichen angegeben, gefolgt von dem GMI und der GV. Ebenfalls enthalten sind statistische Werte wie Median, Standardabweichung, Durchschnitts-, Höchst- und Mindestwert.

Das Kreisdiagramm (vgl. Abbildung 4.10b) wird hierbei eingesetzt, um einen visuellen Überblick über die Anzahl der Messwerte in den Glukosebereichen zu verschaffen.

4.5.3 Übersichtsbereich

Der Übersichtsbereich (vgl. Abbildung 4.11) dient zur Visualisierung der CGM-Daten.

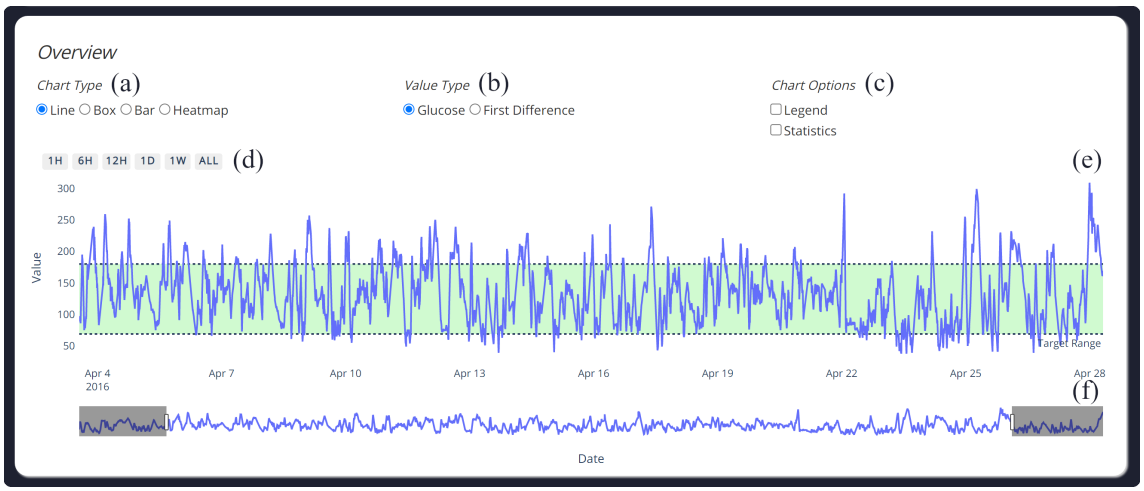


Abbildung 4.11: Übersichtsbereich des Visualisierungssystems: (a) Diagrammtyp; (b) Werttyp; (c) Diagrammoptionen; (d) Buttons zum Liniendiagramm; (e) Liniendiagramm; (f) Bereichsschieberegler zum Liniendiagramm.

Hierzu wird der in der Konfiguration festgelegte Zeitraum mit verschiedenen Diagrammen veranschaulicht. Als Standardeinstellung wird das Liniendiagramm (vgl. Abbildung 4.11e) verwendet. Dieser kann nachfolgend zu einem anderen Diagramm gewechselt (vgl. Abbildung 4.11a) werden. Zur weiteren Auswahl stehen Säulendiagramm, Kastengrafik und Heatmap.

Auch bietet sich für das Liniendiagramm die Einstellung zur Anzeige von aktuellen oder Änderungswerten zum vorherigen Wert (vgl. Abbildung 4.11b) an. Zugleich können mithilfe der kleinen Buttons (vgl. Abbildung 4.11d) und der unteren Leiste (vgl. Abbildung 4.11f) kleinere Zeiträume betrachtet werden. Die Buttons beinhalten verschiedene Zeiteinheiten, welche, ausgehend von der rechten Seite der unteren Leiste, die Zeitreihe anpassen lässt. Die Anzahl der Buttons basiert dabei auf dem gesamten Zeitraum, der sich durch die Anzahl der Messwerte ergibt. Die untere Leiste lässt sich beliebig von beiden Seiten bedienen und kann zusätzlich zur Navigation verwendet werden.

Um eine bessere Übersicht zu generieren, lässt sich auch die Legende (vgl. Abbildung 4.11c) ein- und ausblenden. Hierbei kann auch für das Liniendiagramm entschieden werden, ob alle Zeitabschnitte im gesamten Zeitraum oder nur die zugehörigen statistischen Werte, die sich aus der Kombination der Zeitabschnitte ergeben, angezeigt werden sollen.

4.5.4 Dimensionsreduktionsbereich

In dem Dimensionsreduktionsbereich (vgl. Abbildung 4.12) sollen die Messwerte, mithilfe von Dimensionsreduktionsverfahren auf Muster und Ausreißer untersucht werden.

Durch die erste Einstellung (vgl. Abbildung 4.12a) lassen sich die Werte der Matrix für die Dimensionsreduktion bestimmen. Diese sollen dazu dienen, Muster und Ausreißer anhand von verschiedenen Werten zu finden. Zur Auswahl stehen *Aktuelle Werte*, *Änderungswerte* und *Glukosebereichwerte*.

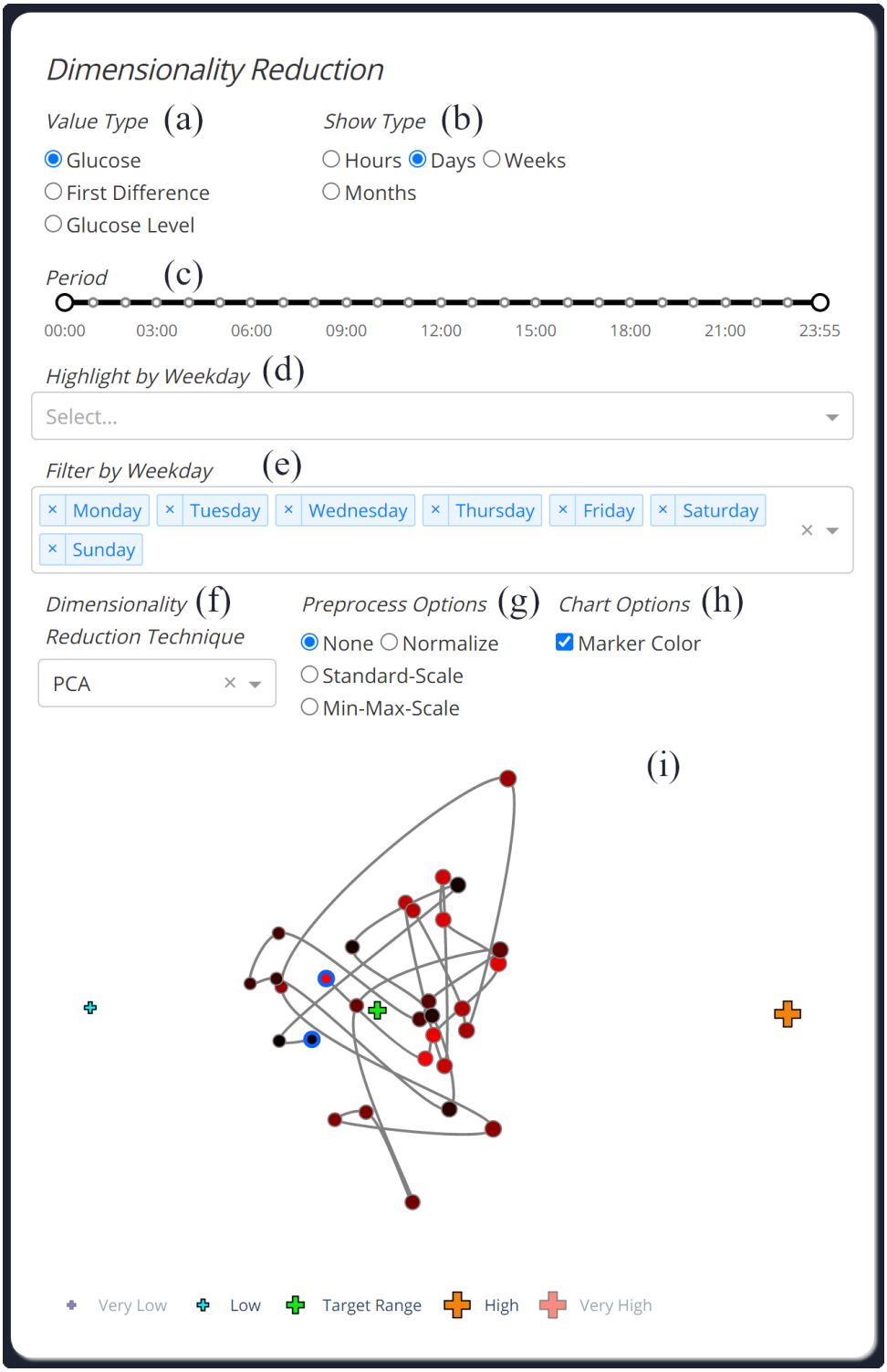


Abbildung 4.12: Dimensionsreduktionsbereich des Visualisierungssystems: (a) Werttyp; (b) Anzeigetyp; (c) Zeitraum; (d) Hervorhebung; (e) Filterung; (f) Dimensionsreduktion; (g) Vorverarbeitung; (h) Diagrammoptionen; (i) Streudiagramm.

Standardmäßig werden hier *Aktuelle Werte* ausgewählt. *Aktuelle Werte* beschreiben die Messwerte der CGM-Daten, während *Änderungswerte* die Differenzen zu den vorherigen Messwerten darstellen. *Glukosebereichwerte* teilen die aktuellen Messwerte in die fünf Glukosebereiche ein. Dabei werden die Messwerte durch Zahlen von eins bis fünf spezifiziert. Aufsteigend von dem niedrigsten Bereich mit eins und dem höchsten Bereich mit fünf.

Für die Visualisierung kann bestimmt werden, welchen Zeitabschnitt die einzelnen Punkte im Streudiagramm (vgl. Abbildung 4.12i) darstellen sollen. Zur Auswahl stehen *Stunden*, *Tage*, *Wochen* oder *Monate* (vgl. Abbildung 4.12b). *Tage* werden hierbei als Standardeinstellung genommen. Für *Stunden* lässt sich zusätzlich die Anzahl der Stunden pro Punkt bestimmen. Dabei können eine, zwei, drei, vier, sechs, acht oder zwölf Stunde/-n ausgewählt werden. Damit soll der Vergleich verschiedener Zeitabschnitte über den Tag ermöglicht werden.

Weiterhin stehen auch Filterungsverfahren (vgl. Abbildung 4.12c) zur Verfügung, welche den beobachteten Zeitabschnitt verkleinern können. Für *Tage* entspricht das 0:00 bis 23:55 Uhr und für *Monate* ein Tag bis 31 Tage. Dadurch sollen spezifizierte Zeitabschnitte über verschiedene Tage oder Monate miteinander verglichen werden. Als Standardeinstellung wird für *Tage* 0:00 bis 23:55 Uhr und für *Monate* die ersten 28 Tage ausgewählt.

Bei der Darstellung von *Stunden*, *Tage* oder *Monate* können einzelne Punkte hervorgehoben (vgl. Abbildung 4.12d) werden. Für *Stunden* stehen demnach 24 Stunden, für *Tage* sieben Wochentage und für *Monate* zwölf Monate zur Auswahl. Somit können Verhalten zu bestimmten Uhrzeiten, Wochentagen oder Monaten aufgedeckt werden.

Neben der Hervorhebung lassen sich auch einzelne Wochentage oder Monate aus dem Graph herausfiltern (vgl. Abbildung 4.12e). Dadurch werden explizite Darstellungen von Wochentagen oder Monaten ermöglicht. Für die Filterung müssen diese aus der Anzeige entfernt werden.

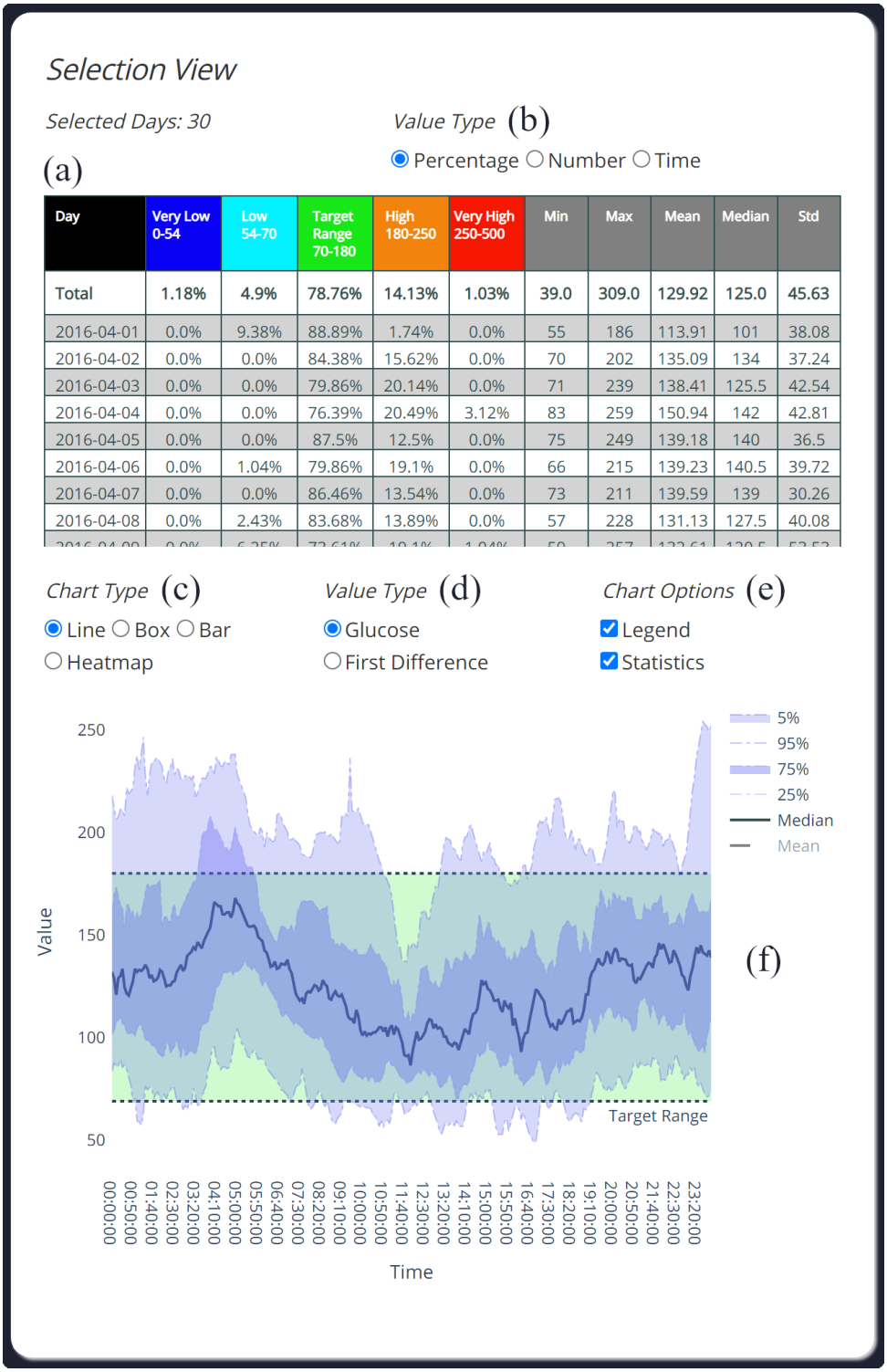
Zur Dimensionsreduktion werden PCA, MDS, t-SNE und UMAP bereitgestellt, welche sich im Dropdown-Listenfeld (vgl. Abbildung 4.12f) beliebig auswählen lassen. Als Standardeinstellung wird PCA verwendet.

Ferner werden auch Vorverarbeitungsmöglichkeiten (vgl. Abbildung 4.12g) für die Dimensionsreduktionsverfahren bereitgestellt. Diese werden hier eingesetzt, damit mit kleineren Werten gearbeitet, und dadurch möglicherweise bessere Rechenleistungen erzielt werden können. Außerdem können dadurch auch andere Darstellungen geschaffen werden. Folgende Vorverarbeitungsmöglichkeiten stehen zur Auswahl: *Keine Vorverarbeitung*, *Normalisierung*, *Standardskalierung* oder *Min-Max-Skalierung*. Je nach Auswahl treten häufig verschiedene Resultate auf, wobei die *Standardskalierung* und *Min-Max-Skalierung* vielmehr zu identischen Resultaten tendieren. Standardmäßig ist *Keine Vorverarbeitung* ausgewählt.

Mit der letzten Funktion (vgl. Abbildung 4.12h) kann entschieden werden, ob die Farben der Punkte den Zeitverlauf beschreiben sollen (rot nach schwarz) oder nicht (nur schwarz).

4.5.5 Selektionsbereich

Der Selektionsbereich (vgl. Abbildung 4.13) ermöglicht den Vergleich von ausgewählten Zeitabschnitten.



Day	Very Low 0-54	Low 54-70	Target Range 70-180	High 180-250	Very High 250-500	Min	Max	Mean	Median	Std
Total	1.18%	4.9%	78.76%	14.13%	1.03%	39.0	309.0	129.92	125.0	45.63
2016-04-01	0.0%	9.38%	88.89%	1.74%	0.0%	55	186	113.91	101	38.08
2016-04-02	0.0%	0.0%	84.38%	15.62%	0.0%	70	202	135.09	134	37.24
2016-04-03	0.0%	0.0%	79.86%	20.14%	0.0%	71	239	138.41	125.5	42.54
2016-04-04	0.0%	0.0%	76.39%	20.49%	3.12%	83	259	150.94	142	42.81
2016-04-05	0.0%	0.0%	87.5%	12.5%	0.0%	75	249	139.18	140	36.5
2016-04-06	0.0%	1.04%	79.86%	19.1%	0.0%	66	215	139.23	140.5	39.72
2016-04-07	0.0%	0.0%	86.46%	13.54%	0.0%	73	211	139.59	139	30.26
2016-04-08	0.0%	2.43%	83.68%	13.89%	0.0%	57	228	131.13	127.5	40.08

Abbildung 4.13: Selektionsbereich des Visualisierungssystems: (a) Informationstabelle; (b) Werttyp für die Informationstabelle; (c) Diagrammtyp; (d) Werttyp für das Liniendiagramm; (e) Diagrammoptionen; (f) Liniendiagramm.

Im oberen Bereich wird eine separate Informationstabelle (vgl. Abbildung 4.13a) für nur selektierte Zeitabschnitte bereitgestellt. Darin werden ähnliche Werte wie die aus der Tabelle des Informationsbereichs (vgl. Abschnitt 4.10a) beschrieben. Aus Platzgründen wird dabei die Anzahl der Messungen des jeweiligen Bereiches alleinstehend in Prozent, Zahlen oder Zeiten angegeben. Als Standardeinstellung werden die Werte in Prozent angezeigt, welche sich jedoch im Nachhinein mit der Funktion oberhalb der Tabelle (vgl. Abbildung 4.13b) anpassen lassen.

Im unteren Bereich werden die selben Diagramme (vgl. Abbildung 4.13f) mit denselben Einstellungen wie aus dem Übersichtsbereich (vgl. Abschnitt 4.5.3) verwendet. Sie können oberhalb des Diagramms angepasst (vgl. Abbildung 4.13c, 4.13d und 4.13e) werden.

4.6 Interaktionen

Neben den Interaktionen der grafischen Benutzeroberfläche existieren auch weitere Interaktionsmöglichkeiten, die durch die Graphen bereitgestellt werden. Hierzu waren neben Tooltips auch Funktionen wie Selektieren, Zoomen und Navigieren durch die verwendete interaktive Visualisierungsbibliothek (vgl. Kapitel 5.1.3) vorgegeben. Für unseren Visualisierungsansatz wurden diese weiter aufgegriffen und angepasst.

4.6.1 Selektion

Beim Streudiagramm besteht die Möglichkeit einzelne oder auch mehrere Punkte mithilfe der Maus auszuwählen. Dies kann durch Mausklick, Rechteck- oder Lasso-Auswahl erfolgen. Für die Wahl mehrerer Punkte wird dabei die Shift-Taste gedrückt gehalten. Es gibt ebenfalls die Möglichkeit zum Abwählen, indem auf einer leeren Stelle ein Doppelklick mit der Maus ausgeführt wird.

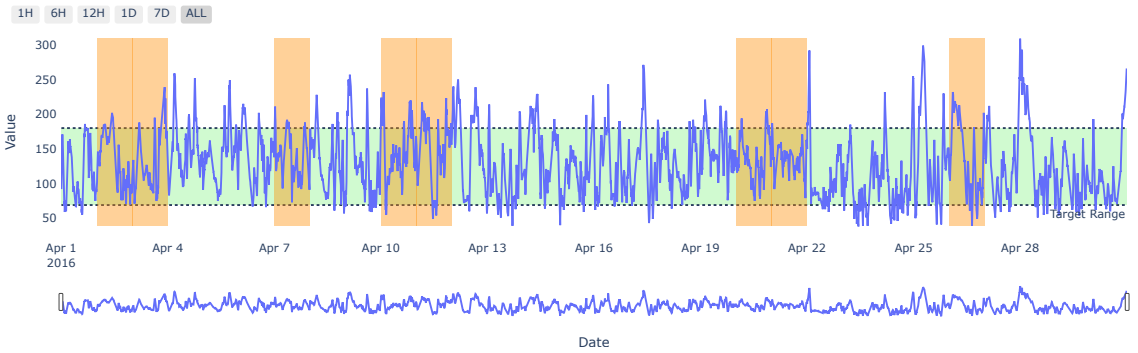
Eine Selektion wird auch in den Visualisierungen des Übersichtsbereichs angeboten. Durch einen Mausklick auf einen Zeitabschnitt wird der jeweilige Punkt im Streudiagramm ausgewählt. Die Auswahlmöglichkeiten beschränken sich in diesen Graphen allerdings nur auf einen Mausklick, da sie nicht von allen Diagrammen unterstützt werden. Ein größeres Problem stellt der schreibgeschützte Speicher dar. Dieser ist zwar in der Lage einen Mausklick zu speichern, kann diesen jedoch nicht wieder entfernen, ohne die Daten erneut laden zu müssen.

Die ausgewählten Zeitabschnitte in den Visualisierungen werden zudem im Selektionsbereich (vgl. Abbildung 4.5.5) angezeigt.

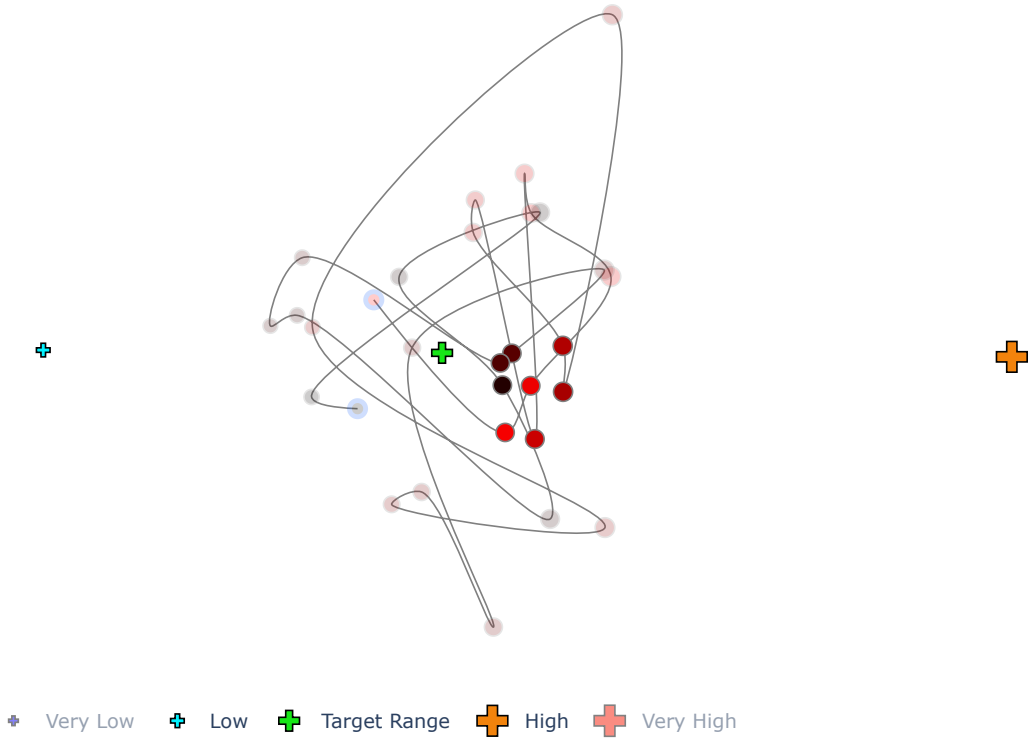
4.6.2 Hervorhebung

Selektierte Punkte im Streudiagramm werden im Liniendiagramm des Übersichtbereiches (vgl. Abbildung 4.5.3) hervorgehoben. Hierzu werden auf den ausgewählten Zeitabschnitten Rechtecke verwendet (vgl. Abbildung 4.14a).

Gleichzeitig heben sich durch die Selektion in den Visualisierungen des Übersichtsbereichs die entsprechenden Punkte im Streudiagramm hervor (vgl. Abbildung 4.14b).



(a)



(b)

Abbildung 4.14: Hervorhebung der Zeitabschnitte: (a) Im Liniendiagramm werden die Zeitabschnitte mithilfe von Rechtecken hervorgehoben; (b) Im Streudiagramm werden nur die Punkte der selektierten Zeitabschnitte hervorgehoben und die restlichen Punkte transparent gemacht.

5 Implementierung

Dieses Kapitel dient zur Veranschaulichung von Informationen zum Back- und Frontend des Visualisierungssystems. Hierzu werden verwendete Bibliotheken zur Verarbeitung und Visualisierung der CGM-Daten vorgestellt.

5.1 Backend

Für die Entwicklung des Visualisierungssystems haben wir uns für Python als Programmiersprache entschieden. Diesbezüglich werden umfangreiche Bibliotheken zur Visualisierung und zum unüberwachten maschinellen Lernen angeboten. Im Nachfolgenden werden die wichtigsten Bibliotheken aufgezählt, die im Visualisierungssystem zum Einsatz kommen.

5.1.1 Datenverarbeitung

NumPy[1] ist eine Open Source Bibliothek, die effiziente Funktionen für die Verarbeitung von Daten bereitstellt. Die Funktionen basieren auf ihrer eigenen Datenstruktur „ndarray“ und erlauben beliebige Dimensionen von Daten ähnlich wie Listen darzustellen.

Bei Pandas[2] handelt es sich um eine Open Source Bibliothek, mit der tabellarische Daten manipuliert werden können. Dazu gehört auch die Verarbeitung von Zeitreihendaten. Hierfür wird ihre Datenstruktur „DataFrame“ verwendet, die unter anderem auf der Datenstruktur von NumPy aufbaut. Des Weiteren stellt Pandas auch Funktionen für die Interpolation fehlender Werte bereit.

5.1.2 Dimensionsreduktion

Für die Dimensionsreduktion wird auf die Open Source Bibliothek Scikit-learn[3] zugegriffen. Dieser verfügt über Funktionen, welche zur Softwareentwicklung für maschinelles Lernen beitragen. Darunter enthalten sind auch Dimensionsreduktionsverfahren wie PCA, MDS und t-SNE. Außerdem finden sich auch Vorverarbeitungsmöglichkeiten wie die Normalisierung und Standardisierung der Daten in der Bibliothek wieder.

Für die Verwendung von UMAP wird auf die UMAP-learn[4] Bibliothek zugegriffen.

[1] https://numpy.org
[2] https://pandas.pydata.org
[3] https://scikit-learn.org
[4] https://umap-learn.readthedocs.io/

5.1.3 Visualisierung

Plotly[5] ist ein Unternehmen, welches Softwarewerkzeuge zur Analyse und Visualisierung von Daten bereitstellt. Darunter befinden sich auch Open Source Bibliotheken für die interaktive Visualisierung von Daten. Diese unterstützen eine vielfältige Auswahl an Diagrammen zu Themenbereichen wie beispielsweise Finanz, Wissenschaft, Statistik und maschinelles Lernen.

Für die Visualisierung war es uns bei der Auswahl einer Bibliothek wichtig, die im Abschnitt 2.1.3 erwähnten Visualisierungen umsetzen zu können. Das Angebot von Plotly war diesbezüglich geeignet. Insbesondere die Möglichkeit mit den generierten Grafiken zu interagieren, war ein entscheidender Punkt für die Auswahl dieser Bibliothek.

5.2 Frontend

Als Frontend wurde das Dash[6] Framework ausgewählt. Dash ist ein Open Source Framework von dem Unternehmen Plotly, dessen Zweck zur Entwicklung von Analyse- und Visualisierungssoftware dient. Die Software wird dabei als Webanwendung entwickelt und unterstützt Sprachen wie Python, HTML, JavaScript und CSS. Des Weiteren besteht eine einfache Integrierbarkeit der Visualisierungsbibliothek von Plotly.

Wegen der Kompatibilität mit der Visualisierungsbibliothek wurde schließlich Dash als Frontend ausgewählt. Zudem konnte uns die Open Source Variante von Dash durch ihre angebotenen Funktionen überzeugen.

[5] https://plotly.com
[6] https://plotly.com/dash

6 Fallstudie

Dieses Kapitel dient zur Evaluation des entwickelten Visualisierungssystems. Um die Fähigkeiten des Visualisierungssystems zu demonstrieren, werden reale Daten aus den CGM-Systemen verwendet. Zuerst werden die CGM-Daten mithilfe des Visualisierungssystems dargestellt und anschließend analysiert. Die Evaluation erfolgt dabei in Form von drei Fallstudien. Diese haben sich in den verwandten Arbeiten (vgl. Kapitel 3) als nützlich erwiesen und lassen sich auch für unseren Visualisierungsansatz einsetzen.

In den Fallstudien werden neben der Untersuchung der CGM-Daten auch die funktionalen Anforderungen des Visualisierungssystems in Abschnitt 4.1 unter Beweis gestellt. Weiterhin werden die Fallstudien nach dem vorgestellten Arbeitsvorgang in Abschnitt 4.2 durchgeführt.

6.1 Datenerhebung

Die verwendeten Daten wurden von der Nightscout Foundation[1] zur Verfügung gestellt. Dabei handelt es sich um ein Open Source Projekt, das weitere Open Source Projekte für Menschen mit Typ-1-Diabetes unterstützt. Hierzu stellen sie von Nutzern freiwillig hochgeladene CGM-Daten für den persönlichen Gebrauch in einer Cloud zur Verfügung. Von dort aus können die CGM-Daten, beispielsweise in einem Browser, weiter betrachtet werden.

6.2 Analyse

Dieser Abschnitt beschreibt die Fallstudien zur Evaluierung des Visualisierungssystems. Er besteht aus drei verschiedenen CGM-Daten, die unterschiedlich lange Zeiträume umfassen.

Für die Fallstudien werden, falls nicht weiter angemerkt wird, die Standardeinstellungen aus dem Konfigurationsbereich (vgl. Abschnitt 4.5.1) übernommen. Demnach werden nur bis zu drei Stunden linear interpoliert. Glukosebereiche und Zielwerte werden auch nicht weiter angepasst. Des Weiteren werden folgende Farben in der Abbildung 6.1 zur Hervorhebung der Wochentage verwendet.

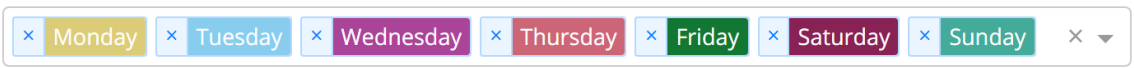


Abbildung 6.1: Farben zur Hervorhebung der Wochentage.

[1] https://www.nightscoutfoundation.org

6.2.1 Fallstudie 1

In der ersten Fallstudie fokussieren wir uns auf CGM-Daten, die in einem Zeitraum von 15 Tagen entstanden sind.

Hierbei wollen wir zunächst den Zeitraum mithilfe der generierten Statistiken untersuchen, welche in der Abbildung 6.2 veranschaulicht werden.

Area	Very Low 0-54	Low 54-70	Target Range 70-180	High 180-250	Very High 250-500	Total
Measurements	93 (2.15%)	288 (6.67%)	3261 (75.49%)	593 (13.73%)	85 (1.97%)	4320 (100.0%)
Time in Ranges	7:45:00	1 day, 0:00:00	11 days, 7:45:00	2 days, 1:25:00	7:05:00	15 days, 0:00:00
GMI	4.45%	4.8%	6.3%	8.23%	9.71%	6.49%
GV	10.21%	7.34%	22.36%	9.57%	5.46%	35.8%
Min	38.0	54.0	70.0	180.0	250.0	38.0
Max	53.5	69.67	179.5	249.0	321.0	321.0
Mean	47.71	62.29	125.08	205.75	267.72	133.11
Median	49	63	122	202	264	127
Standard Deviation	4.87	4.57	27.97	19.69	14.62	47.66

Abbildung 6.2: Fallstudie 1 - Statistiken zu den CGM-Daten.

Für das Erste lässt sich in den Statistiken erkennen, dass alle Messwerte vorliegen. Im Gesamtbereich liegt ein GMI von 6,49 Prozent vor. Zusätzlich erreicht die GV mit 35,80 Prozent knapp den vorgegeben Zielwert von unter 36 Prozent. Durch den Medianwert von 127 und die Standardabweichung von 47,66 lässt sich bereits vermuten, dass sich die meisten Werte im Zielbereich (TIR) befinden. Diese entsprechen 75,49 Prozent der Messwerte, wodurch auch der Zielwert von über 70 Prozent erreicht wird. Hingegen erreichen die niedrigen Bereiche (TBR) ihre vorgegebenen Zielwerte nicht, während bei den höheren Bereichen (TAR) die prozentualen Werte passen.

Im Liniendiagramm (vgl. Abbildung 6.3) werden die statistischen Werte der Tage dargestellt. Bei der Betrachtung lässt sich darin eine Variation der Werte zwischen ungefähr 10 und 23 Uhr erkennen, welches auch durch die Heatmap in der Abbildung 6.4 bestätigt wird. Insbesondere der *Montag, 23. März 2020*, *Mittwoch, 25. März 2020* und *Dienstag, 31. März 2020* besitzen in diesem Zeitabschnitt sehr hohe Werte. Zu erkennen ist auch, wie die Tage in der ersten Hälfte eher zu höheren Werten und bei der zweiten Hälfte eher zu guten, aber auch zu geringen Werten tendieren.

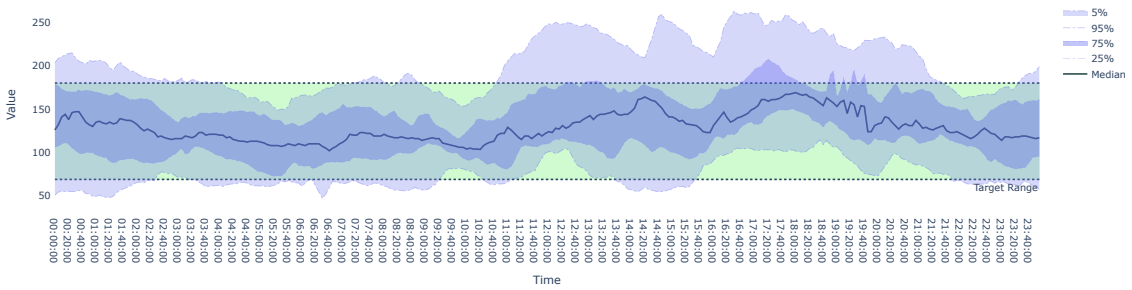


Abbildung 6.3: Fallstudie 1 - Liniendiagramm: Darstellung von statistischen Werten basierend auf 15 Tagen.

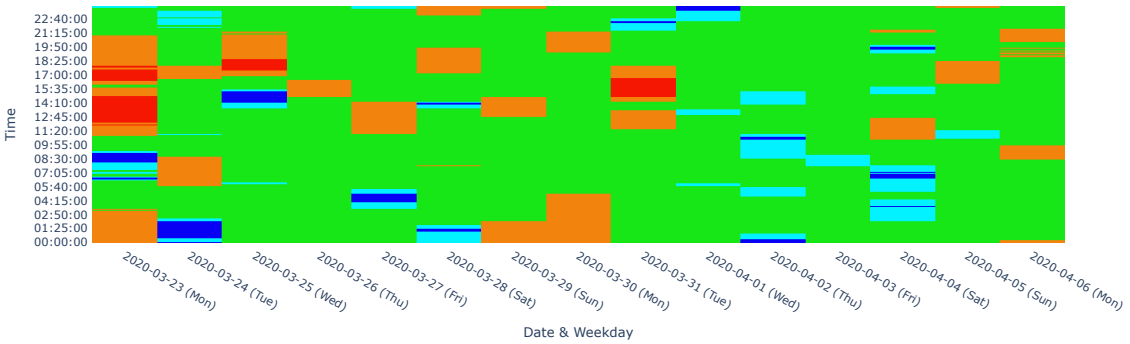


Abbildung 6.4: Fallstudie 1 - Heatmap: Darstellung von 15 Tagen mit Farben der einzelnen Glukosebereiche zu verschiedenen Uhrzeiten.

Als Nächstes wollen wir den gesamten Zeitabschnitt mithilfe der Dimensionsreduktion untersuchen. PCA (vgl. Abbildung 6.5) stellt *Montag, 23. März 2020* und *Dienstag, 31. März 2020* als Ausreißer fest und positioniert diese dementsprechend weit oben. Außerdem ist darin auch bereits ein Muster an den Wochentagen zu erkennen, nämlich dass Montag und Sonntag recht nah beieinander liegen. Anhand der Kreuze wird erkenntlich, dass sich die meisten Tage im Zielbereich (TIR) befinden. Dies wird auch durch die ähnlichen Größen der Punkte visualisiert.

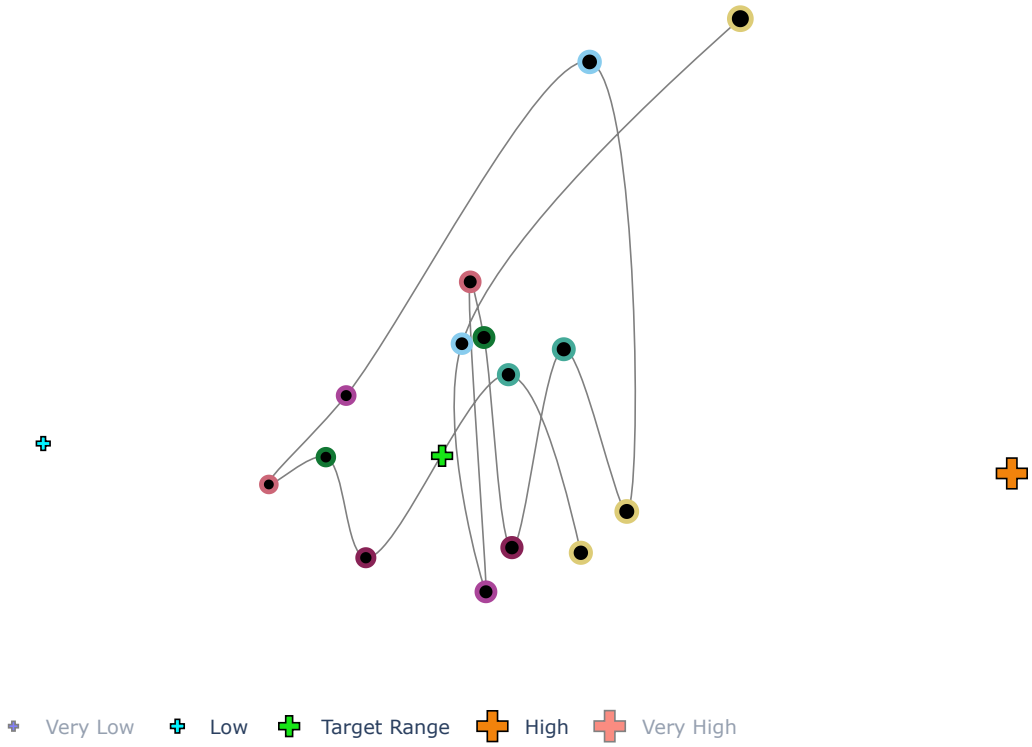
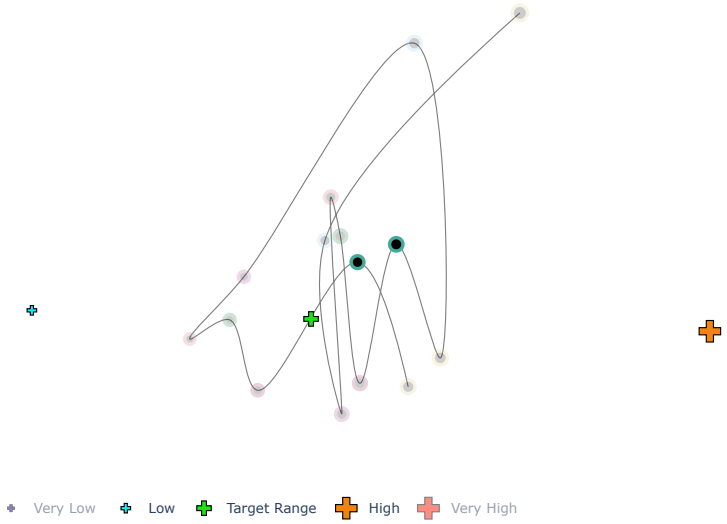
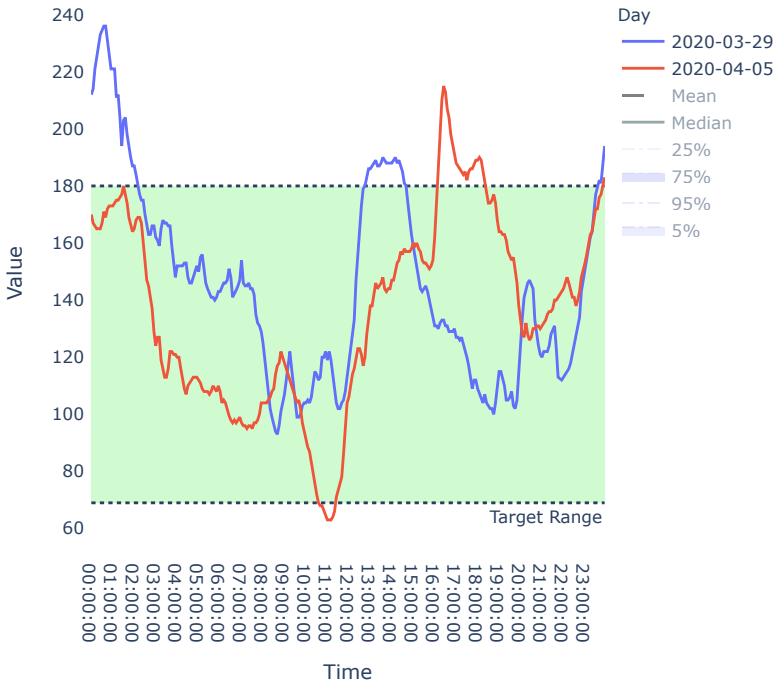


Abbildung 6.5: Fallstudie 1 - Streudiagramm PCA: Darstellung von 15 Tagen als Punkte. Es werden dabei die Wochentage hervorgehoben.

Bei der Untersuchung der Sonntage (vgl. Abbildung 6.6a) wird folgendes Muster in der Abbildung 6.6b festgestellt. Dabei lässt sich ein Verlauf erkennen, welcher von Mitternacht bis Mittag abnimmt und danach wieder aufsteigt.



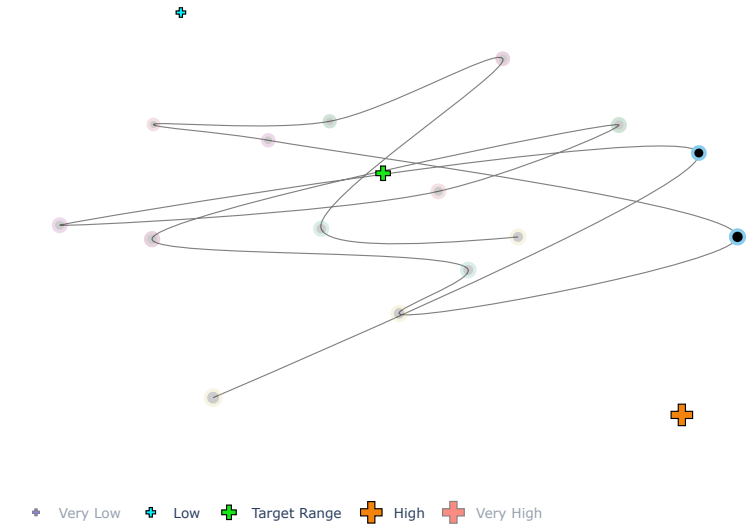
(a)



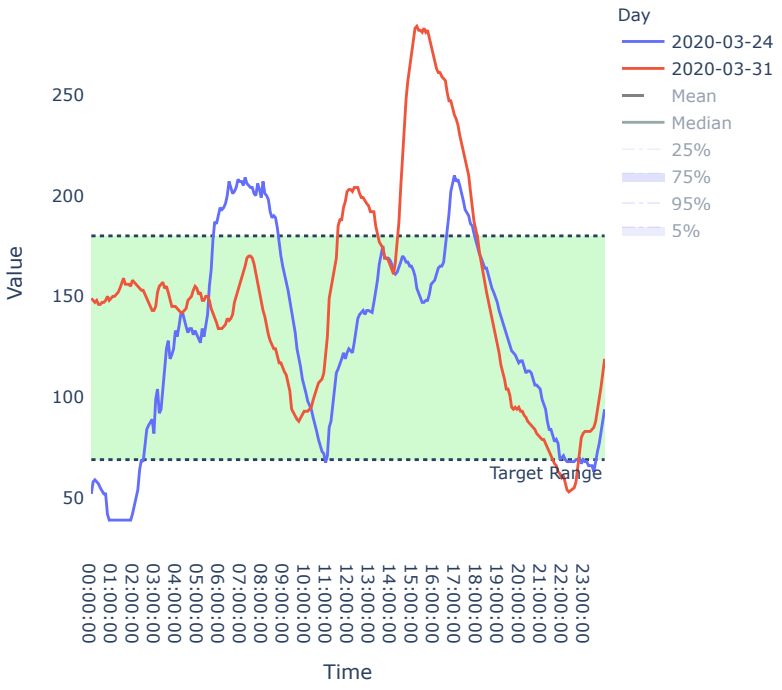
(b)

Abbildung 6.6: Fallstudie 1 - Muster PCA: (a) Hervorgehobene Sonntage im Streudiagramm; (b) Hervorgehobene Sonntage im Liniendiagramm.

MDS (vgl. Abbildung 6.7a) kommt zu ähnlichen Ergebnissen, legt aber die Dienstage recht nah zueinander. Bei der Untersuchung der Dienstage lässt sich ebenfalls ein Muster erkennen, welches in der Abbildung 6.7b veranschaulicht wird. In dieser kann eine Unterscheidung der Linien bis 4 Uhr festgestellt werden, danach nehmen sie einen ähnlichen Verlauf an.



(a)



(b)

Abbildung 6.7: Fallstudie 1 - Muster MDS: (a) Hervorgehobene Dienstage im Streudiagramm; (b) Hervorgehobene Dienstage im Liniendiagramm.

Durch die Anwendung von UMAP werden bei der Auswahl einiger Ansammlungen weitere Muster erkannt, die in der Abbildung 6.8a und 6.8b dargestellt werden. Demnach lassen sich, wie am Anfang der Untersuchung bereits erkannt wurde, auch Variationen zwischen 10 und 23 Uhr vorfinden. Ein Vergleich kann hierzu dem Liniendiagramm in der Abbildung 6.8c und 6.8d entnommen werden, welcher die Tage hervorhebt.

Abbildung 6.8: Fallstudie 1 - Muster UMAP: Hervorgehobene Tage von (a) werden in (c) dargestellt; Hervorgehobene Tage von (b) werden in (d) dargestellt.

6.2.2 Fallstudie 2

In der zweiten Fallstudie untersuchen wir CGM-Daten, die eine Messdauer von 43 Tagen, 17 Stunden und 50 Minuten aufweisen.

Zu dieser werden folgende Statistiken in der Abbildung 6.9 erstellt.

Area	Very Low 0-54	Low 54-70	Target Range 70-180	High 180-250	Very High 250-500	Total
Measurements	157 (1.25%)	312 (2.48%)	7863 (62.41%)	2714 (21.54%)	1552 (12.32%)	12598 (99.42%)
Time in Ranges	13:05:00	1 day, 2:00:00	27 days, 7:15:00	9 days, 10:10:00	5 days, 9:20:00	43 days, 17:50:00
GMI	4.41%	4.82%	6.24%	8.36%	10.52%	7.16%
GV	10.91%	7.15%	23.54%	9.38%	14.31%	44.51%
Min	39.0	54.0	70.0	180.0	250.0	39.0
Max	53.57	69.5	179.0	249.67	400.0	400.0
Mean	45.83	62.93	122.34	211.32	301.43	161.15
Median	47	63	120	210	292	145
Standard Deviation	5	4.5	28.8	19.82	43.14	71.72

Abbildung 6.9: Fallstudie 2 - Statistiken zu den CGM-Daten.

In diesem Fall sind 99,42 Prozent der Messwerte vorhanden. Der Gesamtbereich veranschaulicht einen etwas höheren GMI mit 7,16 Prozent, was auch am Durchschnittswert mit 161,15 erkenntlich wird. Ebenfalls sehr hoch ist auch die GV mit 44,51 Prozent, welche den Zielwert von unter 36 Prozent weit überschreitet. Daher sind in diesem Fall mehrere Schwankungen zu erwarten. Auch liegen sehr niedrige, aber auch sehr hohe Werte vor. Nur der niedrige Bereich und sehr hohe Bereich erreichen ihre Zielwerte.

Das Liniendiagramm stellt in der Abbildung 6.10 die statistischen Werte der Tage dar. Dabei ist zu erkennen, dass sich diese eher im Zielbereich (TIR) und im höheren Bereich (TAR) aufhalten. Von Mitternacht bis Mittag nehmen die Werte ab, steigen jedoch wieder auf und variieren anschließend sehr.

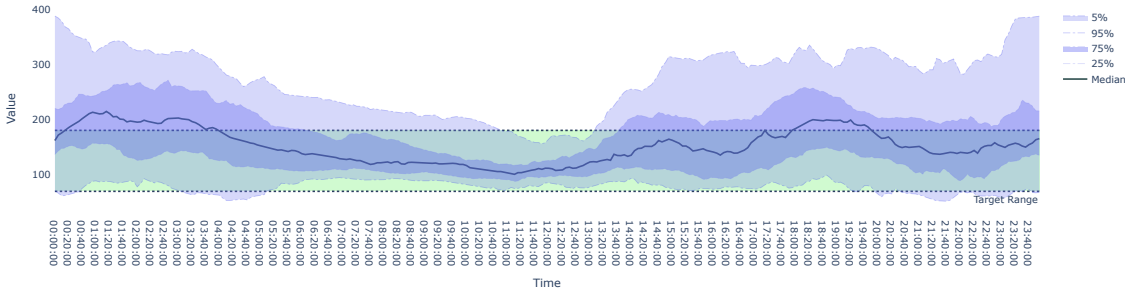


Abbildung 6.10: Fallstudie 2 - Liniendiagramm: Darstellung von statistischen Werten basierend auf 44 Tagen.

In der Kastengrafik (vgl. Abbildung 6.11) werden die Tage als Kasten veranschaulicht. Darin sind mehrere Tage zu erkennen, bei denen weniger als 70 Prozent des Zielwerts im Zielbereich (TIR) erreicht werden. Diese Tage werden vergleichsweise auch viel größer dargestellt als Tage, bei denen der Zielwert erreicht wird. Außerdem werden bei den Tagen, die den Zielwert erreichen, auch mehrere Ausreißer festgestellt.

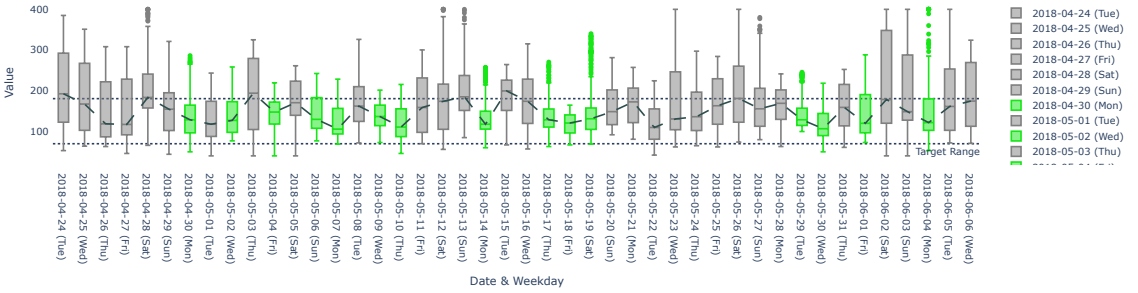


Abbildung 6.11: Fallstudie 2 - Kastengrafik: Darstellung von 44 Tagen als Kasten. (Grün) 15 Tage haben den Zielwert erreicht. (Grau) 29 Tage haben den Zielwert nicht erreicht.

Für die weitere Analyse wenden wir t-SNE auf die Tage an. Dazu wird folgendes Resultat in dem Streudiagramm (vgl. Abbildung 6.12) generiert. In diesem lassen sich die einzelnen Wochentage mit ihren zugeordneten Farben übersichtlich erkennen. Dabei halten sich einige der selben Wochentage in gegenseitiger Nähe auf und deuten dementsprechend schon auf ein ähnliches Verhalten hin. Außerdem lassen sich durch die Verteilung der Glukosebereiche die Werte einiger Wochentage nur durch ihrer Größe der Punkte einschätzen.

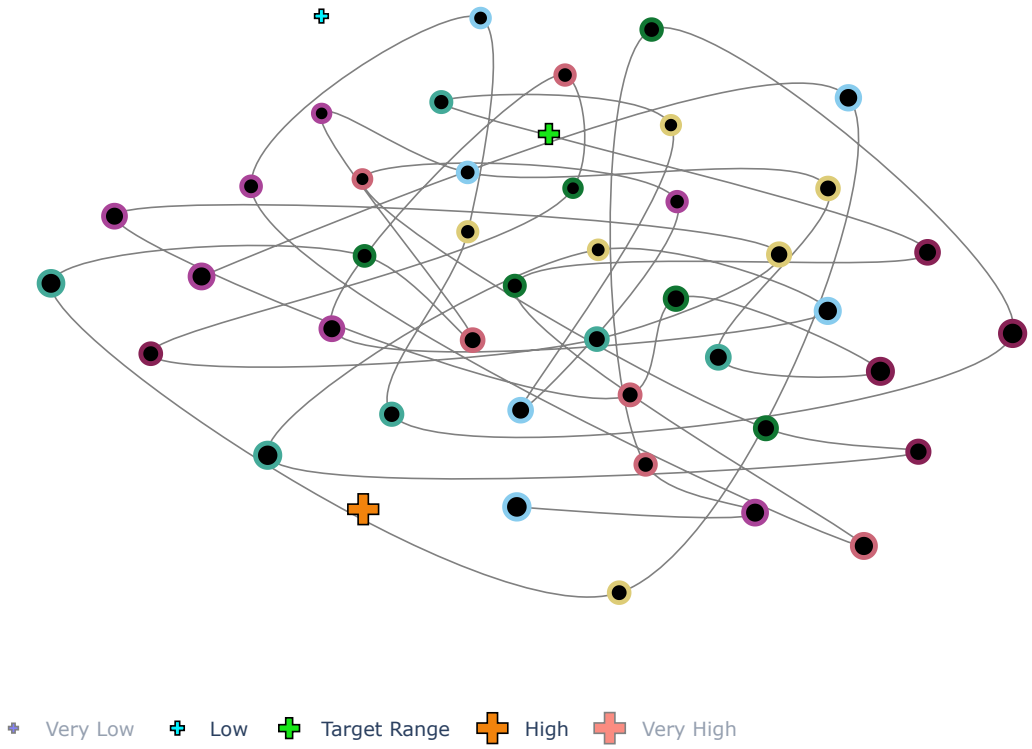


Abbildung 6.12: Fallstudie 2 - Streudiagramm t-SNE: Darstellung von 42 Tagen als Punkte. Es werden dabei die Wochentage hervorgehoben.

Bei der Untersuchung zweier Tage (vgl. Abbildung 6.13a) werden folgende Muster in der Abbildung 6.13c erkannt. Dabei lässt sich bei diesen Tagen ein ähnliches Verhalten wie im Liniendiagramm mit den statistischen Werten (vgl. Abbildung 6.10) erkennen. Hierzu werden in der Abbildung 6.13b beide Tage mit der Dimensionsreduktion auf deren Änderungswerte dargestellt. Diese werden in der Abbildung 6.13d mit den Ähnlichkeiten ihrer Änderungswerte visualisiert.

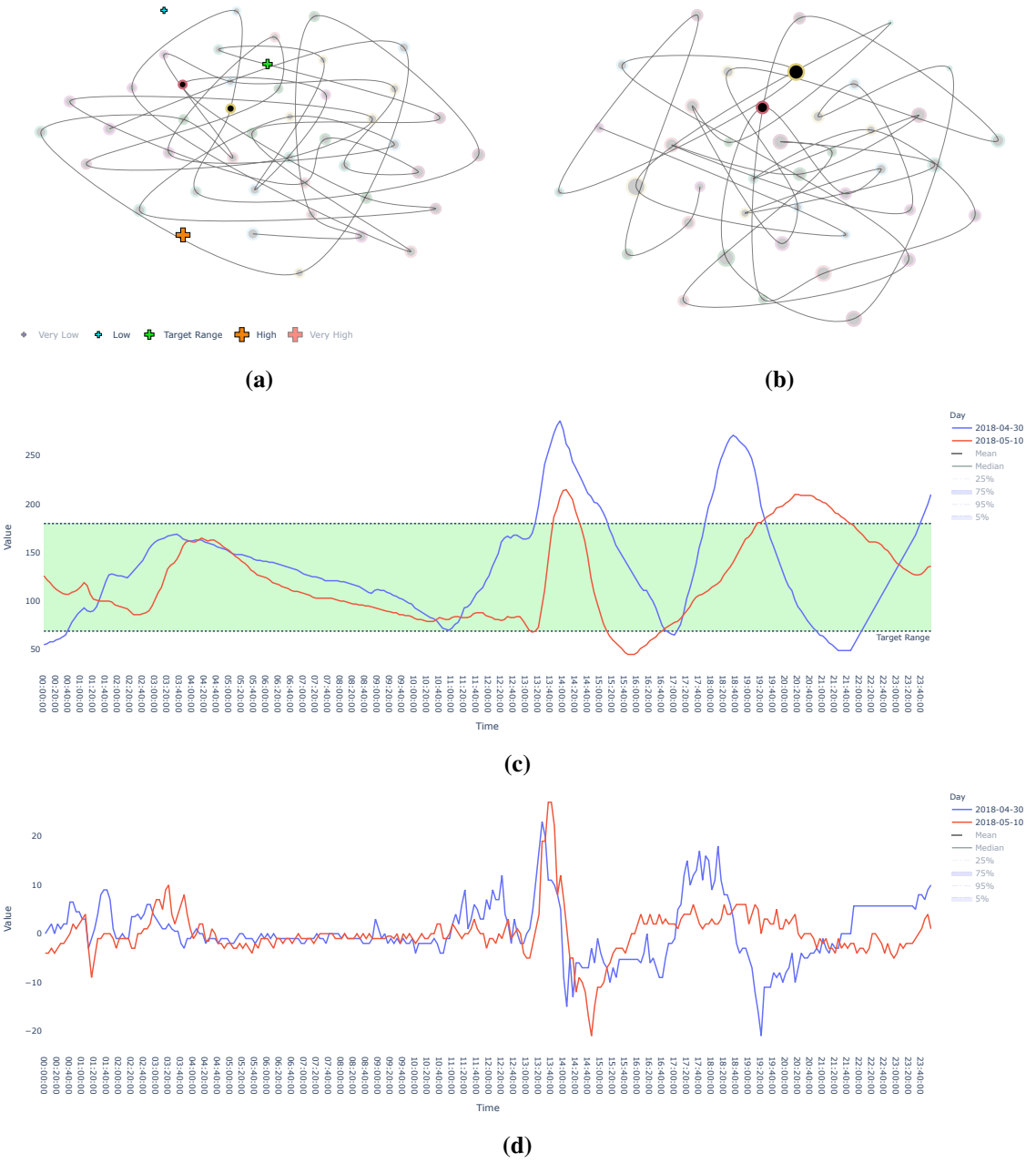


Abbildung 6.13: Fallstudie 2 - t-SNE: Hervorgehobene Tage von (a) werden in (c) dargestellt; Hervorgehobene Tage von (b) werden in (d) dargestellt.

6.2.3 Fallstudie 3

In der letzten Fallstudie liegt der Fokus auf CGM-Daten, die eine Messdauer von 336 Tagen, 6 Stunden und 35 Minuten aufweisen.

Aufgrund des großen Zeitraumes interpolieren wir bis zu zwei Tagen. Dazu werden folgende Statistiken in der Abbildung 6.14 generiert.

Area	Very Low 0-54	Low 54-70	Target Range 70-180	High 180-250	Very High 250-500	Total
Measurements	5518 (5.7%)	11938 (12.33%)	77725 (80.26%)	1574 (1.63%)	92 (0.09%)	96847 (99.78%)
Time in Ranges	19 days, 3:50:00	41 days, 10:50:00	269 days, 21:05:00	5 days, 11:10:00	7:40:00	336 days, 6:35:00
GMI	4.43%	4.8%	5.83%	8.05%	9.91%	5.67%
GV	9.13%	7.1%	22.96%	8.27%	7.13%	32.48%
Min	39.0	54.0	70.0	180.0	250.0	39.0
Max	53.92	69.8	179.9	249.0	322.0	322.0
Mean	46.98	62.15	105.48	198.01	275.78	98.47
Median	47.2	63	101.67	193	274	95
Standard Deviation	4.29	4.41	24.22	16.37	19.65	31.98

Abbildung 6.14: Fallstudie 3 - Statistiken zu den CGM-Daten.

Durch die lineare Interpolation liegen 99,78 Prozent der Messwerte vor. Die GV erreicht mit 32,48 Prozent den vorgegeben Zielwert von unter 36 Prozent. Weiterhin wird der GMI trotz des großen Zeitraumes mit 5,67 Prozent sehr gering gehalten. Dies lässt sich auch am geringen Durchschnittswert erkennen. Eine weitere Besonderheit stellt die niedrige Standardabweichung mit 31,98 und der Median mit 95 dar. Diesen nach sollten sich die meisten Werte im Zielbereich (TIR) befinden. Darüber hinaus erreichen alle Glukosebereiche ihre vorgegebenen Zielwerte, bis auf die Werte in den niedrigen Bereichen (TBR), da diese sehr hoch ausfallen.

Das Liniendiagramm in der Abbildung 6.15 stellt die statistischen Werte der Wochen dar. Wie vermutet, lässt sich darin eine geringe Variation der Werte erkennen. Nur am Freitag kommt es zu einem kurzen Anstieg. Größtenteils befinden sich die Werte im Zielbereich (TIR).

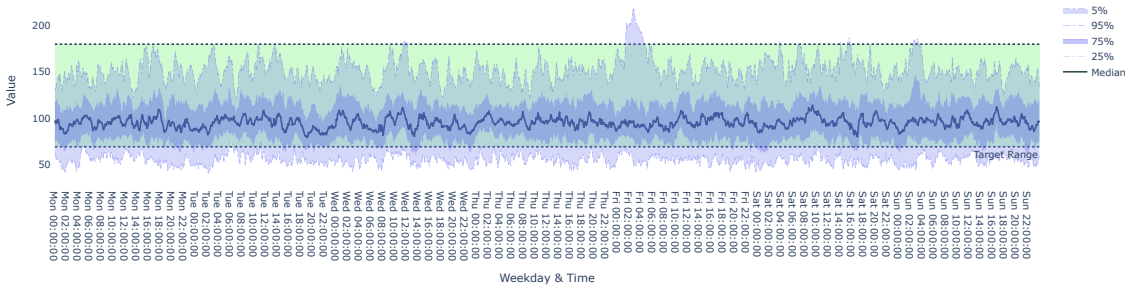


Abbildung 6.15: Fallstudie 3 - Liniendiagramm: Darstellung von statistischen Werten basierend auf 49 Wochen.

Im gestapelten Säulendiagramm (vgl. Abbildung 6.16) werden die niedrigen Glukosewerte sehr erkenntlich visualisiert. Dabei werden auch die Ähnlichkeiten der Wochen zueinander deutlich gemacht. Demnach besitzen die erste und die letzte Woche weniger Werte als die anderen Wochen.

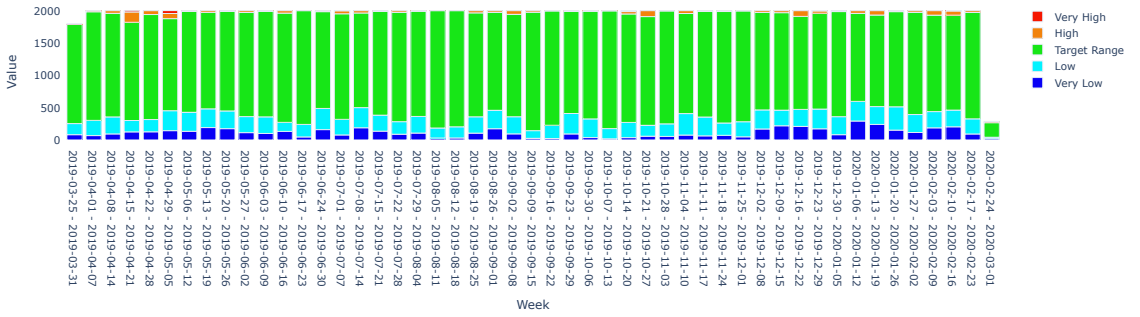


Abbildung 6.16: Fallstudie 3 - Gestapeltes Säulendiagramm: Darstellung von 49 Wochen als gestapelte Säulen.

Für die Dimensionsreduktion wenden wir UMAP auf die Wochen an. Dazu wird in der Abbildung 6.17 das Resultat in einem Streudiagramm dargestellt. Zunächst lassen sich darin drei Ansammlungen erkennen, worin sowohl Punkte mit ähnlichen Größen, als auch einige Punkte als Ausreißer enthalten sind. Weiterhin befinden sich die einzelnen Glukosebereiche in gemeinsamer Nähe, was die Analyse der anderen Ansammlungen erschwert.

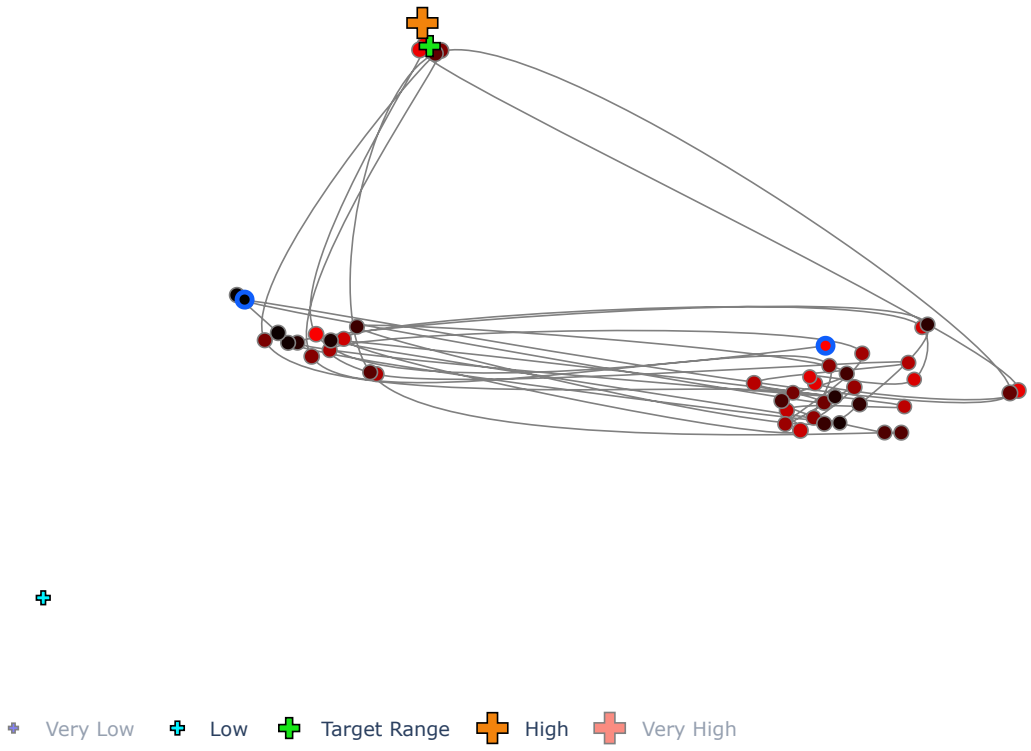


Abbildung 6.17: Fallstudie 3 - Streudiagramm UMAP: Darstellung von 47 Wochen als Punkte. Es werden drei Ansammlungen mit einigen Ausreißern veranschaulicht.

Wir wollen die Ausreißer auf der rechten Seite des Streudiagramms (vgl. Abbildung 6.18a) genauer untersuchen. Bei der Auswahl dieser Wochen lassen sich nahezu identische Muster aufweisen, die in der Abbildung 6.18c veranschaulicht werden. Bei diesen handelt es sich um sehr weit auseinander liegende Wochen, die aber ähnliche Verhalten aufweisen. Insbesondere am Freitag treten bei beiden sehr hohe Werte auf. Zwei weitere Wochen lassen sich in der Nähe vorfinden (vgl. Abbildung 6.18b). Diese stellen auch ein ähnliches Verhalten in der Abbildung 6.18d dar. Jedoch handelt es sich diesmal um zwei aufeinanderfolgende Wochen.

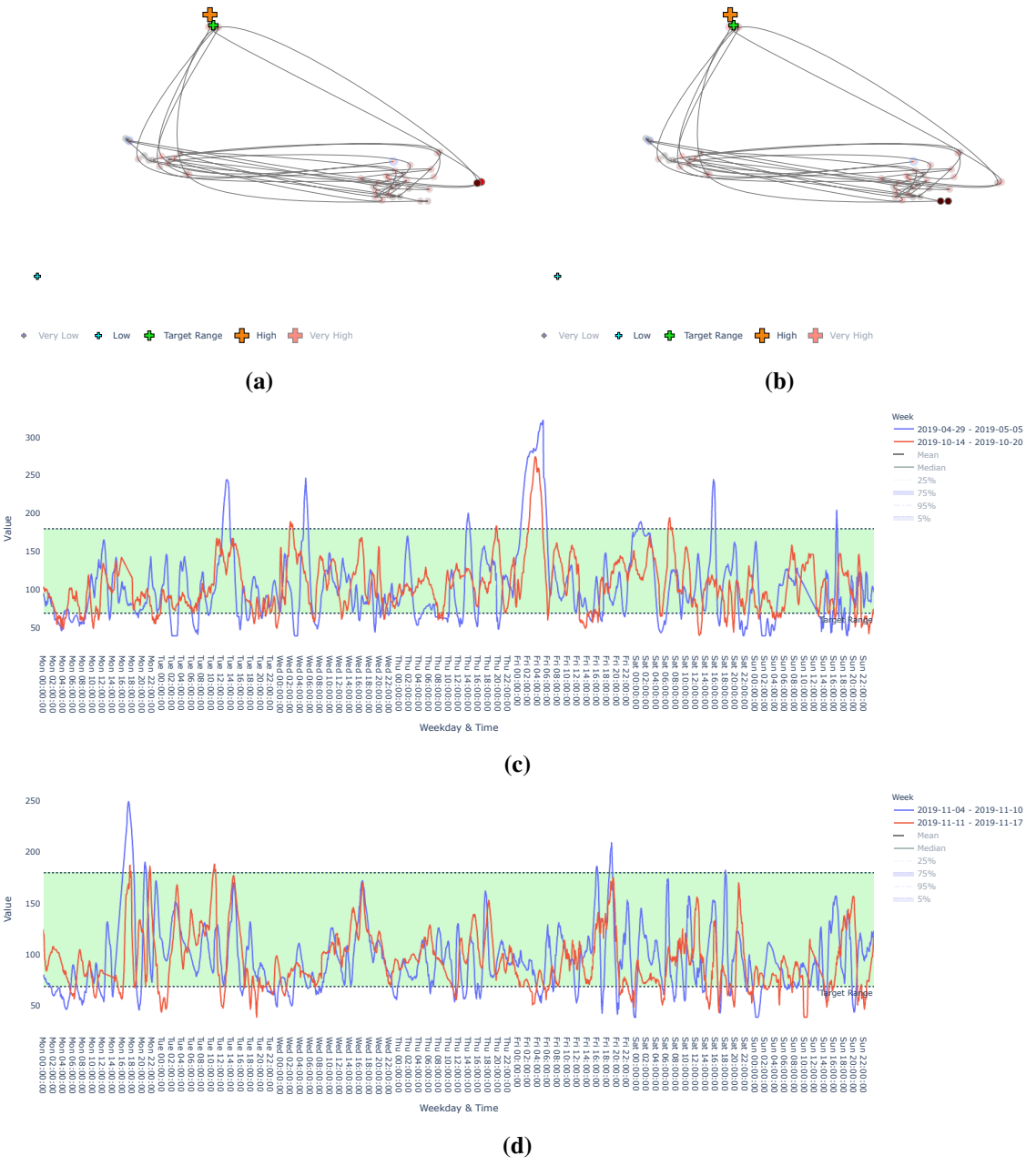


Abbildung 6.18: Fallstudie 3 - UMAP: Hervorgehobene Wochen von (a) werden in (c) dargestellt; Hervorgehobene Wochen von (b) werden in (d) dargestellt.

6.3 Ergebnisse

Die Fallstudien haben gezeigt, dass sich die CGM-Daten mithilfe des Visualisierungssystems untersuchen lassen.

Durch statistische Werte konnten zunächst Auswertungen zu den CGM-Daten gemacht werden. Dabei wurden Werte beobachtet, die im Nachhinein durch Visualisierungen wie Kastengrafik, Heatmap, Linien- und Säulendiagramm bestätigt wurden.

Weiterhin konnten mithilfe der Dimensionsreduktion verschiedene Zeitabschnitte in einem Streudiagramm dargestellt werden. Mittels PCA und MDS konnten Muster und Ausreißer in den Wochentagen erkannt werden, die anschließend in einem Liniendiagramm veranschaulicht wurden. Für die Analyse von Wochentagen wurde ebenfalls t-SNE eingesetzt. Dadurch konnten Wochentage übersichtlicher dargestellt werden und auf mögliche Muster deuten. Außerdem konnten mithilfe von t-SNE auch Muster zu den Änderungswerten zweier Tage gefunden werden. UMAP hat sich besonders für die Erkennung von Mustern geeignet. Durch die entstehenden Ansammlungen konnten schnell identische Zeitabschnitte gefunden werden. Es wurden dabei mehrere Wochen gefunden, die nahezu identische Verhalten aufwiesen. Bei einem handelte es sich um zwei aufeinanderfolgende Wochen, und bei dem anderen um zwei Wochen, die Monate entfernt voneinander lagen.

Mit diesen Ergebnissen sind wir davon überzeugt, dass sich unüberwachte maschinelle Lernverfahren neben klassischen Zeitreihen-Visualisierungen für die Suche nach Mustern und Ausreißern in den CGM-Daten einsetzen lassen.

7 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein System zur visuellen Analyse für Daten der kontinuierlichen Glukosemessung von Diabetes-Patienten/-innen entwickelt. Hierzu wurden unüberwachte maschinelle Lernverfahren in Kombination mit klassischen Zeitreihen-Visualisierungen verwendet.

Zu Beginn wurden Grundlagen über die Thematik dieser Arbeit zusammen mit verwandten Arbeiten erarbeitet. Darauf basierend wurde mithilfe der gewonnen Informationen ein Visualisierungsansatz zur Analyse der Daten vorgestellt. In diesem wurden zunächst verschiedene Visualisierungen im Detail erklärt. Daraufhin wurde die Benutzeroberfläche des Visualisierungssystems mit den zugehörigen Interaktionen präsentiert. Schließlich wurden die Fähigkeiten des Visualisierungssystems anhand von drei Fallstudien evaluiert.

Die Ergebnisse der Fallstudien haben gezeigt, dass sich unüberwachte maschinelle Lernverfahren in Kombination mit Zeitreihen-Visualisierungen für die Analyse von Daten der kontinuierlichen Glukosemessung einsetzen lassen. Zunächst konnten durch die Darstellung statistischer Werte Rückschlüsse zu den Daten der kontinuierlichen Glukosemessung gemacht werden. Weiterhin konnten durch verschiedene Verfahren zur Dimensionsreduktion Muster und Ausreißer zu verschiedenen Zeitabschnitten gefunden werden. Diese konnten schließlich mithilfe von verschiedenen Visualisierungen dargestellt werden.

Ausblick

In der Zukunft könnte das Visualisierungssystem weiter ausgebaut werden. Hierzu werden im Folgenden einige Ideen vorgestellt.

Zunächst könnte das System mit weiteren Dateien oder Messdaten erweitert werden. Diese könnten neben den Messwerten noch andere Variablen beschreiben. Einige Dateien der Nightscout Foundation beschreiben auch Aktivitäten, welche neben den Messvorgängen durchgeführt wurden. Jene könnten als weitere Indiz in der Erkennung von Mustern und Ausreißern benutzt werden. Beispielsweise könnten bestimmte Muster nur bei der Ausführung einer Aktivität auftreten. Diese würden sich eventuell als weitere Informationen für die Therapieanpassung eignen.

Für den Umgang mit den fehlenden Werten ist auch eine Erweiterung möglich. Aktuell werden die fehlenden Werte mithilfe der umgebenden Werte linear interpoliert. Dadurch treten bei der Dimensionsreduktion mit PCA sehr nah aneinander liegende Punkte auf, die ein lineares Verhalten aufweisen. Hierfür könnten maschinelle Lernverfahren helfen, um Werte aus bestimmten Verhalten herzuleiten. Beispielsweise anhand von mehreren Stunden, bei denen jedes Mal dieselben Muster auftreten. Es hätte insbesondere auf größere Zeitabschnitte eine Auswirkung. Wochentage würden nicht mehr als Ausreißer hervorstechen, sondern sich bei ähnlichen Wochentagen ansammeln. Somit könnten andere Ausreißer ersichtlicher werden.

Darüber hinaus könnten auch weitere Dimensionsreduktionsverfahren und Visualisierungen implementiert werden, wodurch sich eventuell neue Möglichkeiten für die Erkennung von Mustern oder Ausreißern ergeben.

Literaturverzeichnis

[Agg15] C. C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015, S. 17, 34–37. ISBN: 978-3-319-14142-8. DOI: 10.1007/978-3-319-14142-8 (zitiert auf S. 16).

[AJXW19] M. Ali, M. W. Jones, X. Xie, M. Williams. „TimeCluster: Dimension Reduction applied to Temporal Data for Visual Analytics“. In: *The Visual Computer* 35.6 (Juni 2019), S. 1013–1026. ISSN: 1432-2315. DOI: 10.1007/s00371-019-01673-y. URL: https://doi.org/10.1007/s00371-019-01673-y (zitiert auf S. 26, 31).

[AMM+08] W. Aigner, S. Miksch, W. Müller, H. Schumann, C. Tominski. „Visual Methods for Analyzing Time-Oriented Data“. In: *IEEE Transactions on Visualization and Computer Graphics* 14 (Jan. 2008), S. 47–60. DOI: 10.1109/TVCG.2007.70415 (zitiert auf S. 17).

[BBC+18] R. M. Bergenstal, R. W. Beck, K. L. Close, G. Grunberger, D. B. Sacks, A. Kowalski, et al. „Glucose Management Indicator (GMI): A New Term for Estimating A1C From Continuous Glucose Monitoring“. In: *Diabetes Care* 41.11 (Sep. 2018), S. 2275–2280. DOI: 10.2337/dc18-1581. URL: https://doi.org/10.2337/dc18-1581 (zitiert auf S. 22).

[BDB+19] T. Battelino, T. Danne, R. M. Bergenstal, S. A. Amiel, R. Beck, T. Biester, et al. „Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range“. In: *Diabetes Care* 42.8 (Juni 2019), S. 1593–1603. DOI: 10.2337/dci19-0028. URL: https://doi.org/10.2337/dci19-0028 (zitiert auf S. 22, 23, 38).

[BSH+16] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, P. Dragicevic. „Time Curves: Folding Time to Visualize Patterns of Temporal Evolution in Data“. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (Jan. 2016). DOI: 10.1109/TVCG.2015.2467851. URL: https://hal.inria.fr/hal-01205821 (zitiert auf S. 25, 26, 33).

[BWS+12] J. Bernard, N. Wilhelm, M. Scherer, T. May, T. Schreck. „TimeSeriesPaths: Projection-Based Explorative Analysis of Multivarate Time Series Data“. In: Bd. 20. Juni 2012 (zitiert auf S. 26).

[DNB+17] T. Danne, R. Nimri, T. Battelino, R. M. Bergenstal, K. L. Close, J. H. DeVries, et al. „International Consensus on Use of Continuous Glucose Monitoring“. In: *Diabetes Care* 40.12 (2017), S. 1631–1640. ISSN: 0149-5992. DOI: 10.2337/dc17-1600. eprint: https://care.diabetesjournals.org/content/40/12/1631.full.pdf. URL: https://care.diabetesjournals.org/content/40/12/1631 (zitiert auf S. 22, 27).

[Dod08] Y. Dodge. „Time Series“. In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, 2008, S. 536–539. ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_401. URL: https://doi.org/10.1007/978-0-387-32833-1_401 (zitiert auf S. 16).

[EHBW16] S. van den Elzen, D. Holten, J. Blaas, J. J. van Wijk. „Reducing Snapshots to Points: A Visual Analytics Approach to Dynamic Network Exploration“. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), S. 1–10. DOI: 10.1109/TVCG.2015.2468078 (zitiert auf S. 26).

[Fre20] G. Freckmann. „Basics and use of continuous glucose monitoring (CGM) in diabetes therapy“. In: *Journal of Laboratory Medicine* 44.2 (2020), S. 71–79. DOI: doi:10.1515/labmed-2019-0189. URL: https://doi.org/10.1515/labmed-2019-0189 (zitiert auf S. 22).

[GE18] T. Gschwandtner, O. Erhart. „Know Your Enemy: Identifying Quality Problems of Time Series Data“. In: *2018 IEEE Pacific Visualization Symposium (PacificVis)*. 2018, S. 205–214. DOI: 10.1109/PacificVis.2018.00034 (zitiert auf S. 18, 19).

[HCGD20] L. Hajderanj, D. Chen, E. Grisan, S. Dudley. „Single- and Multi-Distribution Dimensionality Reduction Approaches for a Better Data Structure Capturing“. In: *IEEE Access* 8 (2020), S. 207141–207155. DOI: 10.1109/ACCESS.2020.3038460 (zitiert auf S. 20).

[HKF16] S. Haroz, R. Kosara, S. L. Franconeri. „The Connected Scatterplot for Presenting Paired Time Series“. In: *IEEE Transactions on Visualization and Computer Graphics* 22.9 (2016), S. 2174–2186. DOI: 10.1109/TVCG.2015.2502587 (zitiert auf S. 18).

[Hot33] H. Hotelling. „Analysis of a complex of statistical variables into principal components.“ In: *Journal of Educational Psychology* 24 (1933), S. 498–520 (zitiert auf S. 20).

[HR19] J. Harreiter, M. Roden. „Diabetes mellitus – Definition, Klassifikation, Diagnose, Screening und Prävention (Update 2019)“. In: *Wiener klinische Wochenschrift* 131.1 (Mai 2019), S. 6–15. ISSN: 1613-7671. DOI: 10.1007/s00508-019-1450-4. URL: https://doi.org/10.1007/s00508-019-1450-4 (zitiert auf S. 21).

[HSS+20] A. Hinterreiter, C. Steinparz, M. Schöfl, H. Stitz, M. Streit. *ProjectionPathExplorer: Exploring Visual Patterns in Projected Decision-Making Paths*. 2020. arXiv: 2001.08372 [cs.AI] (zitiert auf S. 25).

[Int21] International Diabetes Center. *AGP - Ambulatory Glucose Profile: AGP Reports*. Zuletzt besucht: 31.08.2021. 2021. URL: http://www.agpreport.org/agp/agpreports (zitiert auf S. 27, 28, 32, 33, 37).

[JME10] W. Javed, B. McDonnel, N. Elmqvist. „Graphical Perception of Multiple Time Series“. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), S. 927–934. DOI: 10.1109/TVCG.2010.162 (zitiert auf S. 17).

[KAD17] D. C. Klonoff, D. Ahn, A. Drincic. „Continuous glucose monitoring: A review of the technology and clinical use“. In: *Diabetes Research and Clinical Practice* 133 (2017), S. 178–192. ISSN: 0168-8227. DOI: https://doi.org/10.1016/j.diabres.2017.08.005. URL: https://www.sciencedirect.com/science/article/pii/S0168822717304217 (zitiert auf S. 22).

[KM17] E. Keogh, A. Mueen. „Curse of Dimensionality“. In: *Encyclopedia of Machine Learning and Data Mining*. Hrsg. von C. Sammut, G. I. Webb. Boston, MA: Springer US, 2017, S. 314–315. ISBN: 978-1-4899-7687-1. DOI: 10.1007/978-1-4899-7687-1_192. URL: https://doi.org/10.1007/978-1-4899-7687-1_192 (zitiert auf S. 20).

[Kru64] J. B. Kruskal. „Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis“. In: *Psychometrika* 29.1 (März 1964), S. 1–27. ISSN: 1860-0980. DOI: 10.1007/BF02289565. URL: https://doi.org/10.1007/BF02289565 (zitiert auf S. 21).

[MEB+13] M. Masoodian, B. Endrass, R. Bühling, P. Ermolin, E. André. „Time-Pie visualization: Providing Contextual Information for Energy Consumption Data“. In: *2013 17th International Conference on Information Visualisation*. 2013, S. 102–107. DOI: 10.1109/IV.2013.12 (zitiert auf S. 18).

[MH08] L. van der Maaten, G. Hinton. „Viualizing data using t-SNE“. In: *Journal of Machine Learning Research* 9 (Nov. 2008), S. 2579–2605 (zitiert auf S. 21).

[MHM20] L. McInnes, J. Healy, J. Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML] (zitiert auf S. 21).

[MLL+87] R. Mazze, D. Lucido, O. Langer, K. Hartmann, D. Rodbard. „Ambulatory Glucose Profile: Representation of Verified Self-Monitored Blood Glucose Data“. In: *Diabetes care* 10 (Jan. 1987), S. 111–7. DOI: 10.2337/diacare.10.1.111 (zitiert auf S. 27).

[Pea01] K. Pearson F.R.S. „LIII. On lines and planes of closest fit to systems of points in space“. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), S. 559–572. DOI: 10.1080/14786440109462720. eprint: https://doi.org/10.1080/14786440109462720. URL: https://doi.org/10.1080/14786440109462720 (zitiert auf S. 20).

[PND20] V. Pham, N. Nguyen, T. Dang. „ContiMap: Continuous Heatmap for Large Time Series Data“. In: *2020 Visualization in Data Science (VDS)*. 2020, S. 42–51. DOI: 10.1109/VDS51726.2020.00009 (zitiert auf S. 19).

[SA13] R. Sathya, A. Abraham. „Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification“. In: *International Journal of Advanced Research in Artificial Intelligence* 2 (Feb. 2013). DOI: 10.14569/IJARAI.2013.020206 (zitiert auf S. 20).

[SAAF17] G. Shurkhovetskyy, N. Andrienko, G. Andrienko, G. Fuchs. „Data Abstraction for Visualizing Large Time Series“. In: *Computer Graphics Forum* 37 (Juli 2017). DOI: 10.1111/cgf.13237 (zitiert auf S. 15, 20).

[TAMS17] C. Tominski, W. Aigner, S. Miksch, H. Schumann. „Images of Time“. In: Jan. 2017, S. 23–42. ISBN: 9780415786324 (zitiert auf S. 17).

[VAT19] S. Velliangiri, S. Alagumuthukrishnan, S. I. Thankumar joseph. „A Review of Dimensionality Reduction Techniques for Efficient Computation“. In: *Procedia Computer Science* 165 (2019). 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019, S. 104–111. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.01.079. URL: https://www.sciencedirect.com/science/article/pii/S1877050920300879 (zitiert auf S. 20).

[VP20] B. V. Vishwas, A. Patel. *Hands-on Time Series Analysis with Python: From Basics to Bleeding Edge Techniques*. 2020, S. 1–3, 12–13. ISBN: 978-1-4842-5992-4. DOI: 10.1007/978-1-4842-5992-4 (zitiert auf S. 15, 16, 18).

[Wor21] World Health Organization. *Global report on diabetes*. Zuletzt besucht: 31.08.2021. Apr. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/diabetes (zitiert auf S. 13).

Alle URLs wurden zuletzt am 31. 08. 2021 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift