Gottfried Wilhelm
Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Praktische Informatik
Fachgebiet Software Engineering

Erhebung von Erklärbarkeitsanforderungen für Machine-Learning-Methoden von Blockheizkraftwerken

Bachelorarbeit

im Studiengang Informatik

von

Ronja Fuchs

Prüfer: Prof. Dr. rer. nat. Kurt Schneider
Zweitprüfer: Dr. rer. nat. Jil Klünder
Betreuer: M. Sc. Jakob Droste

Hannover, 14. August 2023

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 14. August 2023

Ronja Fuchs

Zusammenfassung

Die Nutzung von Machine-Learning-Methoden zur Prognose von Daten über Zeit gehört zunehmend in den Alltag vieler mittelständischer Unternehmen. Diese Arbeit fand in Zusammenarbeit mit einem solchen Unternehmen statt, welches Blockheizkraftwerke entwickelt, produziert und wartet. Einige der 95 Mitarbeitenden des Unternehmens arbeiten regelmäßig mit Prognosen über die erzeugte elektrische Energie einzelner Blockheizkraftwerke, welche durch Machine-Learning-Methoden erstellt wurden. Diese Mitarbeitende, denen die Arbeit mit Machine-Learning-Methoden häufig noch nicht vertraut ist, müssen den Prognosen nicht nur vertrauen, sondern sie auch verstehen und interpretieren.

Erklärungen helfen, das Vertrauen der Mitarbeitenden in die Prognosen zu steigern und die Arbeit mit diesen zu erleichtern. Im Zuge dieser Arbeit wurden acht Stakeholdergruppen aus dem Unternehmen zu Erklärungen von Machine-Learning Vorhersagen befragt. Für die Befragung wurde ein Prototyp erstellt, welcher einige Beispiele von Erklärungen veranschaulicht und zu welchem die Mitarbeitenden befragt wurden. Die Ergebnisse zeigen, dass eine Erklärung in Form eines Entscheidungsbaums, welcher visuell die Entscheidungen des Modells annähert, den Stakeholdergruppen am meisten hilft. Auch eine Erklärung mittels Angabe der Feature Importance der Parameter des Modells hilft den Befragten. Basierend auf den Ergebnissen der Befragung wurden Anforderungen an Erklärungen der Prognosen erhoben und in einem Konzept realisiert. Das Konzept umfasst globale Erklärungen der Feature Importance des Modells, einen Entscheidungsbaum zur Annäherung der Modell-Entscheidungen, und allgemeine Angaben zum Umfeld der Vorhersagen. Weiterhin wird die Umsetzbarkeit des Konzepts diskutiert.

Abstract

Elicitation of Explainability Requirements for Machine Learning Methods of Cogeneration Plants

The use of Machine-Learning methods to predict data over time is becoming an increasingly important part of the day-to-day business of many mid-sized companies. This thesis is conducted in cooperation with a mid-size company specializing in creating, manufacturing, and maintaining cogeneration plants. Many of the 95 employees of the company regularly work with Machine-Learning forecasts of the electrical energy created by individual cogeneration plants. These employees, who are often not accustomed to working with Machine-Learning methods, must trust, understand and interpret the predictions.

Explanations help increase the employees' trust in the forecasts and make it easier for them to work with the predictions. As part of this thesis, eight stakeholder groups from the company took a survey about explanations of machine learning forecasts. A high-fidelity prototype was created for the survey to illustrate some examples of explanations. The results show that an explanation in the form of a decision tree, which visually approximates the model's decisions, helps the stakeholders' understanding of the prediction the most. An explanation of the feature importance of the model's parameters also helps the respondents. Based on the survey results, requirements for explanations of the forecasts were defined and implemented in a concept. The concept includes global explanations of the model's feature importance, a decision tree to approximate the model's decisions, and general information about the forecast environment. Furthermore, the feasibility of the concept is discussed.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Diese Bachelorarbeit wird in Kollaboration mit der *Kraftwerk Kraft-Wärme-Kopplung GmbH* erstellt. Zur Vereinfachung der Lesbarkeit beziehe ich mich, wenn nicht explizit anders angegeben, auf die *Kraftwerk Kraft-Wärme-Kopplung GmbH* mit Kraftwerk.

1.1 Motivation

Die Firma Kraftwerk entwickelt, produziert und wartet seit 1996 Blockheizkraftwerke. Ein Blockheizkraftwerk (BHKW) ist eine Anlage zur Erzeugung von Energie in Form von Strom und Wärme, wobei die Wärme als Nebenprodukt bei der Stromerzeugung anfällt. Anders als bei herkömmlichen Methoden der Energieerzeugung wird bei BHKWs die Wärme in Form von Heizenergie weiterverwendet. Auf diese Weise können Nutzende von BHKWs Kosten und Energie sparen, da bei herkömmlichen Methoden der Energiegewinnung die entstehende Wärme nicht weiter eingesetzt wird. BHKWs kommen in verschiedenen Größen zum Einsatz, kleinere Anlagen sind eher für Einfamilienhäuser geschaffen, während größere Anlagen unter anderem Krankenhäuser, Wohnkomplexe und Schwimmbäder versorgen. Die aufgestellten BHKWs bedürfen über die Jahre ihrer Laufzeit hinweg mehrerer Wartungen und Beobachtungen, welche bei Kraftwerk über eine sogenannte Fernsteuerung realisiert wird. So lässt sich jedes BHKW zu jeder Zeit über das firmeninterne Verwaltungs-Tool, das Webgate, steuern. Dort werden über die Fernsteuerung hinaus auch allgemeine Informationen des BHKWs sowie historische Sensordaten dargestellt.

Jedes BHKW verfügt über Sensoren, die den Zustand verschiedener Teile der Anlage aufnehmen. Auf diesem Wege kann etwa der Verlauf des Motorwasserdrucks oder der Rücklauftemperatur des Heizkreises über Zeit gespeichert und eingesehen werden. Um diese Daten über ihre Speicherung hinaus zu nutzen, wird bei Kraftwerk Machine-Learning zur Prognose weiterer Verläufe angewandt. Die Anwendung von Machine-Learning-Methoden

zur Auswertung von Daten ermöglicht es Mitarbeitenden von Kraftwerk, Strukturen in den Daten zu erkennen, die mit dem bloßen Auge nicht identifizierbar sind. Beispielsweise können sich wiederholende Muster im Verlauf identifiziert werden. In diesem Zuge gibt es eine Analyse, die für jedes BHKW den produzierten Wert an elektrischer Leistung über einen bestimmten Zeitraum voraussagt. Dies ermöglicht es Mitarbeitenden, die Einstellungen oder Veränderungen am BHKW vorgenommen haben, die vorhergesagten Werte mit den tatsächlichen Werten zu vergleichen. Wenn eine Einstellung die Effizienz eines BHKWs verändert, kann dies durch das Abgleichen der vorausgesagten Werte mit den tatsächlichen direkt identifiziert werden. Zuvor basierten solche Entscheidungen allein auf Spekulationen oder mühsamer Abgleichung mit historischen Daten. Ungenaue Prognosen können zu fehlerhaften Einstellungen an BHKWs führen und somit zu Verlust an Effizienz.

Die Voraussage der elektrischen Leistung über Zeit fällt unter den Begriff Zeitreihenprognose (time series forecasting). Dabei wird mittels Methoden des Machine-Learning ein Modell von Zeitreihendaten erstellt, anhand dessen Vorhersagen über zukünftige Daten getroffen werden können. Derzeit werden bei Kraftwerk mithilfe eines neuronalen Netzes, dem sogenannten Long Short-Term Memory (LSTM) [9], Vorhersagen getroffen. Hierfür wurde *tensorflow*[1] verwendet, eine Open Source Library für Machine-Learning Methoden, die von *Google Brain*[2] entwickelt wurde. Da die Arbeit mit den Ergebnissen von Machine-Learning Methoden für die Mitarbeitenden bei Kraftwerk ungewohnt ist und sie Prognosen teils nicht vertrauen, soll die Arbeit mit Machine-Learning durch Erklärungen vereinfacht werden.

1.2 Problemstellung

Bei der Analyse von Daten zur Modellierung und der Vorhersage von Datenpunkten wird ein neuronales Netz genutzt. Da diejenigen Mitarbeitende bei Kraftwerk, die die Vorhersagen tatsächlich nutzen sollen, nicht unbedingt mit dieser Thematik vertraut sind, ist es wichtig, Vertrauen in die Methoden und Vorhersagen zu schaffen. Außerdem sollen die Mitarbeitende dabei unterstützt werden, die Vorhersagen nachzuvollziehen. Um dies zu realisieren, soll erarbeitet werden, welche Stakeholder an der Arbeit mit den Ergebnissen beteiligt sind und in welcher Form diese durch die Umsetzung von Methoden der Erklärbarkeit unterstützt werden können. Im Zuge dieser Arbeit wird eine Befragung der Mitarbeitenden durchgeführt, um zu spezifizieren, mit welchen Methoden das Vertrauen in die Prognosen gesteigert werden kann. Überdies sollen Erklärbarkeitsanforderungen der Mitarbeitenden erhoben und ein Konzept für Erklärungen entwickelt werden.

[1] `https://www.tensorflow.org/`, Stand 10.07.2023 16:40 Uhr
[2] `https://research.google/teams/brain/`, Stand 10.07.2023 16:40 Uhr

Die Einführung von Machine-Learning-Methoden und Datenanalysen verspricht eine Effizienzsteigerung des Anlagenbetriebs und der Anlagensteuerung. Diese Steigerung birgt jedoch einen nicht vernachlässigbaren Aufwand. Machine-Learning-Methoden in einem so spezifischen Feld erfordern nicht nur Fachwissen über die Thematik des Machine-Learning, sondern auch ein breites Verständnis von BHKWs. Die Beschaffung der Daten sowie deren Korrektur und Analyse erfordern Zeit und Ressourcen, welche investiert werden müssen. Nach der Einführung der Datenanalysen und Prognosen muss dafür gesorgt werden, dass Mitarbeitende bestmöglich mit diesen arbeiten können. Allein die Einführung der Vorhersage-Methoden würde sich weniger lohnen, wenn ihr Potenzial unausgeschöpft bleibt. Dies trifft primär auf Fälle zu, in denen Mitarbeitende die Prognosen nicht nachvollziehen können, oder sie den Ergebnissen der Vorhersage nicht vertrauen. Weiterhin sollen die vorausgesagten Daten bei der Arbeit helfen, und nicht aufgrund von fehlendem Vertrauen oder fehlender Nachvollziehbarkeit zur Unsicherheit führen. Daher ist es von erheblicher Bedeutung, die Ergebnisse mit den richtigen Methoden darzustellen und zu erklären.

1.3 Lösungsansatz

Um das Vertrauen und die Nachvollziehbarkeit der Ergebnisse zu steigern, ist es im ersten Schritt notwendig, die Sichtweise der Mitarbeitenden bezüglich der Darstellung der Prognosen zu erfassen. Nach Chazette et al.[3] helfen Erklärungen dabei, ein System durch Ethik, Fairness und Transparenz vertrauenswürdig zu machen. Erklärbarkeit wurde nicht nur als Mittel zur Erreichung all dieser drei Aspekte in Systemen identifiziert, sondern auch als Möglichkeit, das Vertrauen der Nutzer zu fördern. Die Autoren empfehlen außerdem eine nutzerorientierte Strategie, um Anforderungen an Erklärungen für ein System zu verstehen und zu entwickeln. Nach Chazette und Schneider [4] ist es von großer Relevanz, die Stakeholder der Erklärungen zu identifizieren. Dies helfe der Feststellung ihrer Motivation und Interessen und ermögliche zu erkennen, wie Stakeholder dabei unterstützt werden können, mit dem System oder den Daten zu arbeiten. Die Autoren erläutern weiterhin, dass Erklärungen in einem System Wissenslücken der Stakeholder beheben könnten.

Um herauszufinden, welche Anforderungen verschiedene Stakeholder an die Erklärungen haben, muss erarbeitet werden, welche Stakeholdergruppen mit den Prognosen arbeiten und sich auf die Ergebnisse verlassen müssen. Sind die Stakeholdergruppen herausgearbeitet, gilt es, herauszufinden, ob und wie sich ihre Anforderungen voneinander unterscheiden. Um die Erklärungen möglichst nah am End-Nutzer zu entwickeln und um die Unterschiede in den Anforderungen der Stakeholdergruppen zu erfassen, bietet sich eine Befragung der Mitarbeitenden an.

Durch die enge Arbeit mit den Mitarbeitenden kann die Meinung dieser direkt in die Entwicklung eines Konzeptes für Erklärungen der Prognosen einfließen. Diejenigen Erklärungsmethoden, die aus der Befragung das meiste Vertrauen und die beste Nachvollziehbarkeit ermöglichen, können direkt in das Konzept eingearbeitet werden. Außerdem kann so auf die Anforderungen der Stakeholder eingegangen werden.

1.4 Struktur der Arbeit

Diese Arbeit ist wie folgt strukturiert. In Kapitel 2 werden für diese Arbeit nötige Grundlagen erklärt. Weiterhin werden Grundkenntnisse über BHKWs in Kapitel 2.1 erläutert. In Kapitel 2.2 wird das bei Kraftwerk genutzte Machine-Learning-Modell kurz erläutert. In diesem Kapitel wird sowohl allgemeines Wissen über LSTM Modelle vermittelt als auch kurz auf die genaue Struktur des Modells zur Vorhersage von produzierter elektrischer Leistung an einem BHKW eingegangen. In Kapitel 3 wird der Stand der Systeme bei Kraftwerk vor der Arbeit sowie die Gruppierung der Stakeholder erläutert. In Kapitel 4 wird die Forschungsmethodik dieser Arbeit erläutert und die geführte Umfrage aufgeführt. Die Ergebnisse der Umfrage werden in Kapitel 5 dargestellt. Das daraus resultierende Konzept wird in Kapitel 6 dargelegt. Es folgt eine Diskussion in Kapitel 7, innerhalb welcher die Limitationen dieser Arbeit, die Umsetzbarkeit des entwickelten Konzepts und die Herausforderungen bei der Bearbeitung erörtert werden. Eine kurze Zusammenfassung dieser Arbeit sowie ein Ausblick in zukünftige Forschung sind in Kapitel 8 aufgeführt.

Kapitel 2

Grundlagen und verwandte Arbeiten

Zum Verständnis der genutzten Technologien werden in diesem Kapitel die Grundlagen im Rahmen dieser Arbeit erläutert. Dafür werden in Abschnitt 2.1 BHKWs im Grundsatz vorgestellt. In Abschnitt 2.2 wird die genutzte Machine-Learning-Methode erklärt.

2.1 Blockheizkraftwerke

Ein BHKW nutzt das Prinzip der Kraft-Wärme-Kopplung zur Erzeugung von elektrischer Energie und Wärme. Nach Schaumann et al. [20] beschreibt die Kraft-Wärme-Kopplung die gleichzeitige Gewinnung von mechanischer und thermischer Nutzenergie. Durch das Nutzen der Abwärme, die bei der Gewinnung elektrischer Energie anfällt, haben BHKWs einen höheren Gesamtnutzungsgrad gegenüber herkömmlichen Methoden der Energiegewinnung.

Die Firma Kraftwerk bietet mehrere Modelle von BHKWs an, welche sich in der Menge der produzierten Energie voneinander unterscheiden. Kleinere Modelle erzeugen eine Nettoleistung von bis zu 22 kW, wohingegen größere Modelle eine elektrische Nettoleistung von bis zu 50 kW erzeugen. Die kleineren Modelle versorgen etwa Mehrfamilienhäuser oder Schulen[1]. Größere Modelle versorgen dementsprechend größere Verbraucher wie Schwimmhallen oder Pflegeheime[2]. In dieser Arbeit wird sich mit den größeren Modellen von BHKWs bei Kraftwerk beschäftigt, den *Mephisto*

[1] `https://kwk.info/nachrichten/friedrich-froebel-schule-stade`, Stand 11.07.2023 13:50 Uhr

[2] `https://kwk.info/nachrichten/alten-und-pflegeheim-blumenhain`, Stand 11.07.2023 13:50 Uhr

G50[3].

Mephisto G50 sind meist wärmegeführt und werden durch den Wärmebedarf der Nutzenden gesteuert. Ein BHKW im Grundaufbau besteht aus einem Generator, der an einen Verbrennungsmotor gekoppelt ist. Die Funktionsweise eines BHKWs ist vereinfacht in Abbildung 2.1 dargestellt. Der Motor wird mit Kraftstoff betrieben und gibt seine mechanische Leistung an den Generator ab, welcher die mechanische Leistung in elektrische Leistung umwandelt und diese im angeschlossenen elektrischen Netz bereitstellt. Ein Wärmetauscher kühlt den Verbrennungsmotor, nimmt die anfallende Wärme auf und gibt diese in den Heizkreislauf. Dort wird die Wärme dann beispielsweise zum Heizen von Wohnungen genutzt und das zurückfließende Wasser, welches seine Wärme abgegeben hat, wird wieder zum Kühlen des BHKWs genutzt. Daraufhin beginnt das fließende Wasser seinen Kreislauf von vorn.

Bei wärmegeführten BHKWs wird nur so viel Wärme erzeugt, wie im Heizkreislauf genutzt oder in sogenannten Pufferspeichern gespeichert werden kann. Deshalb ist bei wärmegeführten BHKWs, im Gegensatz zu stromgeführten BHKWs, die produzierte elektrische Energie das Nebenprodukt. Da die erzeugte elektrische Energie jedoch eng mit der Erzeugung von Wärme zusammenhängt und einfacher berechnet werden kann als der Wärmeverbrauch, kann sie in Datenanalysen genutzt werden, um das Verhalten eines BHKWs zu untersuchen.

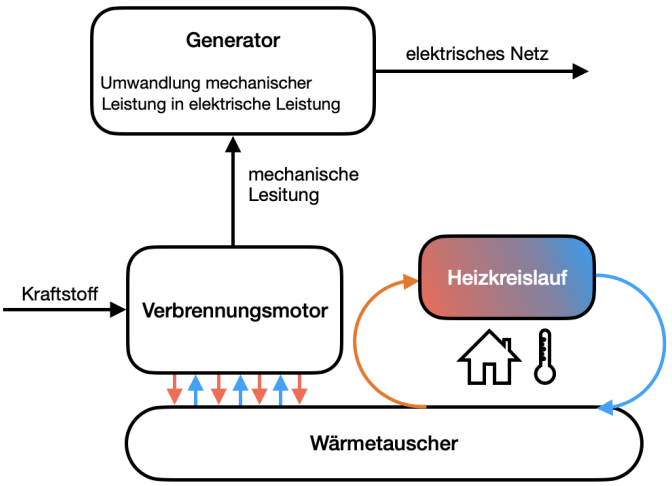

Abbildung 2.1: Vereinfachte Funktionsweise eines BHKWs

[3] `https://kwk.info/produkte/produkt/mephisto-g50-erdgas`, Stand 10.07.2023 18:40 Uhr

2.2 Machine-Learning-Methode

Die genutzte Machine-Learning-Methode zur Vorhersage der produzierten elektrischen Leistung eines BHKWs über Zeit ist ein Long Short-Term Memory-Modell, kurz LSTM. Diese wird im Folgenden oberflächlich erläutert.

2.2.1 Long Short-Term Memory-Modell

Nach Russell und Norvig [18] bauen viele Machine-Learning-Methoden auf der Verwendung von neuronalen Netzen auf. Die Autoren erläutern, dass sich neuronale Netze aus Knoten zusammensetzen, welche durch gerichtete Verknüpfungen verbunden sind. Verlaufen alle gerichteten Verbindungen in die gleiche Richtung, so spreche man von einem Feedforward Netz. Die Autoren erklären weiterhin, dass es bei einem Feedforward Netz bis auf die Gewichtung keinen internen Zustand gebe. Gäbe es im Netz jedoch Zyklen, das heißt gerichtete Pfade, die zurück zum Ursprung führen, spreche man von einem rekurrenten Netz. Eingaben in diese Knoten seien sogenannte Aktivierungen. Solche Knoten sind stark vereinfacht in Abbildung 2.2 dargestellt. So diene eine Verknüpfung des Knotens i zum Knoten j dazu, eine Aktivierung a_i von i nach j zu propagieren. Jeder Verknüpfung sei zudem Gewicht $w_{i,j}$ zugeordnet. Jede Einheit j berechne dann zunächst die gewichtete Summe ihrer Eingaben und wende dann eine Aktivierungsfunktion an, um die Ausgabe zu berechnen. Diese mathematische Struktur wird nach Russel und Norvig [18] auch Neuron genannt. Die Verknüpfung nur-linearer Aktivierungsfunktionen würde dazu führen, dass das Modell nur lineare Funktionen darstellen könne. Daher sei die Aktivierungsfunktion im Falle eines LSTM eine logistische Funktion. Durch die Anwendung einer logistischen Funktion sei gewährleistet, dass das gesamte Netz auch nichtlineare Funktionen darstellen kann.

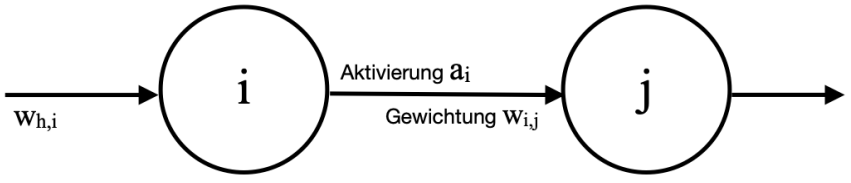

Abbildung 2.2: Knoten i, j eines neuronalen Netzes, Aktivierung a_i propagiert über eine Verknüpfung mit Gewichtung $w_{i,j}$.

Ein LSTM ist ein rekurrentes, oder auch rückgekoppeltes, neuronales Netz. So ist es dadurch ausgezeichnet, dass es seine eigenen Ausgaben

immer wieder in seine eigenen Eingaben einspeist. Russell und Norvig [18] erwähnen, dass diese Rückkopplung in sogenannten Aktivierungsebenen, auch Schichten genannt, resultiert. Dieses Verhältnis ist in Abbildung 2.3 vereinfacht dargestellt. Die Knoten i_1, j_1, k_1 bilden eine Schicht, durch welche eine Aktivierung propagiert. Die Ausgabe $output_1$ von Schicht L_1 führt zur Eingabe, $input_2$ von Schicht L_2. Diese Schichten könnten einen stabilen Zustand erreichen, aber auch zu Schwingungen neigen oder gar chaotisches Verhalten zeigen. Durch die Abhängigkeit der Ausgabe von der Eingabe, welche wiederum abhängig von den vorigen Ausgaben ist, welche selbst abhängig von vorigen Eingaben sind, könne ein rückgekoppeltes Netz ein Kurzzeitgedächtnis unterstützen. Der Einfluss, den eine Eingabe auf eine Ausgabe hat, wird genauso wieder eingegeben und kann sich durch mehrere Stufen Ein- und Ausgabe bewähren.

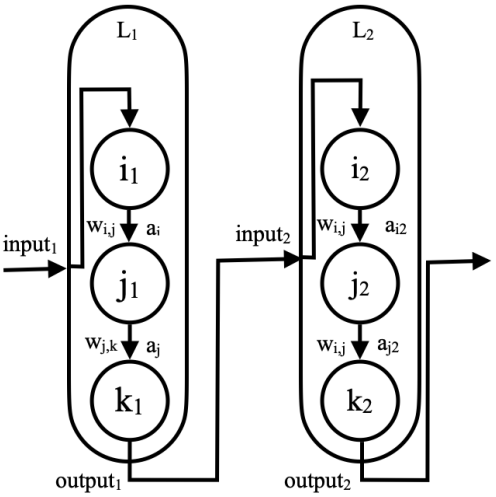

Abbildung 2.3: Schichten L_1 und L_2 eines neuronalen Netzes mit Rückkopplung. Innerhalb der Schichten befinden sich Knoten i_1, j_1, k_1, und i_2, j_2, k_2, verbunden durch gerichtete Verknüpfungen.

2.2.2 LSTM bei Kraftwerk

Für die Vorhersage der produzierten elektrischen Energie über Zeit werden Zeitreihendaten von BHKWs genutzt. Nach Rojat et al. [17] können Zeitreihendaten jede Variable darstellen, die sich über Zeit verändert. Die produzierte elektrische Leistung eines BHKWs wird im Fünf-Minuten-Takt gespeichert und als Eingabe für ein neuronales Netz genutzt. Ein großer Teil der Daten wird zum Training des Modells verwendet. Das Modell zu trainieren, beinhaltet die Eingabe von Daten, mit denen die Gewichte angepasst werden können. Innerhalb mehrerer Iterationen werden einzelne Prognosen gemacht und mit den Eingabedaten verglichen. Bei einer

Differenz zwischen der Prognose und den Eingabedaten werden die Gewichte angepasst. Nach dem Training können die aktuellen Daten der letzten Tage als Eingabe in das neuronale Netz gegeben werden, und für die Berechnung einer Prognose für die nächsten Tage genutzt werden.

Veränderungen in der Produktion von elektrischer Leistung eines BHKWs sind durch die große Menge an Daten sowie die nichtlineare Natur des Verlaufs nicht immer mit dem einfachen Auge zu erkennen. Deswegen sollen Prognosen helfen, den weiteren Verlauf eines BHKWs besser einzuschätzen. Sobald Einstellungen an einem bestimmten BHKW vorgenommen werden, assistieren die Prognosen dabei, Differenzen bei der produzierten Energie seit der Einstellung zu identifizieren und zu bewerten. Dabei kann der tatsächliche Verlauf der produzierten elektrischen Energie nach der Einstellung mit dem prognostizierten Verlauf verglichen und der Nutzen der Einstellung besser bewertet werden.

Weiterhin können Einflussfaktoren wie Wetterverhältnisse, Jahreszeit, Wochentag und Uhrzeit mit in die Erstellung des Modells einbezogen werden und die Prognose beeinflussen. Diese sogenannten "Features" haben einen Einfluss auf die produzierte elektrische Leistung. So wird zum Beispiel an heißen Tagen weniger geheizt und daher weniger Wärme produziert.

Für die Prognosen elektrischer Leistung an BHKWs nutzt Kraftwerk *tensorflow's Keras LSTM*[4]. Das Modell setzt sich aus 24 LSTM Schichten mit einem Dropout[5] von 0.2 zusammen. Dropout beschreibt die Regulation, mit der Ausgaben von Schichten nach Zufallsprinzip weggelassen werden.

Ein frühes Halten des Trainings ist durch die Implementierung einer "early stopping"-Funktion gegeben. Diese Funktion bestimmt, ab wann das Training angehalten wird, wenn sich eine bestimmte Metrik nicht mehr weiter verbessert. Die Aktivierungsfunktion der LSTM-Schichten ist eine Sigmoid Funktion. Nach Russel und Norvig [18] gewährleistet diese, dass das neuronale Netz auch nichtlineare Funktionen darstellen kann.

Weiterhin wird für das Training des Modells eine Verlustfunktion genutzt: die mittlere quadratische Abweichung (MSE). Überdies wird der Adam-Optimierer ("adaptive moment estimation"[6]) verwendet, um das Training zu verbessern. Der Optimierer ist ein Algorithmus, welcher durch die Anpassung der Gewichte des Netzwerks Fehler reduziert. Die Größe der Fehler wird mithilfe des sogenannten Verlustwerts ermittelt. Die Performance des Modells wird anhand der mittleren absoluten Abweichung (MAE) gemessen.

Die Daten zur Eingabe werden vor dem Training und vor den Vorhersagen

[4] `https://www.tensorflow.org/api_docs/python/tf/keras/layers/\acrshort{lstm}`, Stand 10.07.2023 16:42 Uhr

[5] `https://www.tensorflow.org/api_docs/python/tf/keras/layers/Dropout`, Stand 10.07.2023 16:42 Uhr

[6] `https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam`, Stand 10.07.2023 16:42 Uhr

jeweils normiert. Für das Erstellen des Modells bei Kraftwerk wurden insgesamt ca. 17400 Datenpunkte genutzt. Davon wurden 70 % zum Training, 20 % zur Validierung und 10 % zum Testen des Modells verwendet. Vor dem Training wurden die Daten normiert, um eine einfachere Verarbeitung zu ermöglichen. Die sonst übliche Standardisierung der Daten kam an dieser Stelle nicht infrage, da die Eingabe-Daten nicht normalverteilt sind. Für die Normierung wurde die Norm aller Trainings-Daten berechnet. Als Folge dessen sind die Modellausgaben immer normierte Werte, die erst im letzten Schritt mithilfe der zuvor berechneten Norm in die Größenordnung der erzeugten elektrischen Leistung umgerechnet werden.

Diese Komplexität dieser verwendeten Machine-Learning-Methode führt dazu, dass sie von Nutzenden nicht immer verstanden wird. Daher ist ein Ansatz erforderlich, der die Komplexität für Nutzende durch Erklärungen minimiert und es Nutzenden ermöglicht, besser mit der Methode zu arbeiten.

2.3 Erklärungen von Machine-Learning-Modellen

In diesem Abschnitt werden grundlegende Arten von Erklärungen für Machine-Learning Voraussagen (XAI) erläutert. Diese Erklärungen sollen dabei helfen, der eben beschriebenen Komplexität entgegenzuwirken und es Nutzenden ermöglichen, die Prognosen und ihr Zustandekommen besser zu verstehen.

Durch das auch im Alltag steigende Interesse an der Arbeit mit Machine-Learning-Methoden, unter anderem etwa durch Routenberechnungen bei Navigationsapplikationen oder Korrektur von Texten bei Schreibapplikationen, ist nach Chazette et al. [3] auch die Nachfrage nach Vertrauenssteigernden Methoden für diese Systeme gestiegen. Die Autoren beschreiben einen steigenden Bedarf nach Systemtransparenz und Fairness, welche durch zunehmendes Verlassen auf diese Systeme geprägt ist. Sie definieren ein erklärbares System als ein System, welches eine Erklärung zu einem bestimmten Systemaspekt an einen Adressaten gibt. Für die Erklärungen von Voraussagen der erzeugten elektrischen Leistung an BHKWs sind die Mitarbeitenden die Adressaten und der in diesem Kontext relevante Systemaspekt ist die Frage, wie das System zur Voraussage kommt. Die Erklärungen eines erklärbaren Systems sollen nach Chazette et al. [3] Wissenslücken schließen, die jedoch für jedes Individuum sehr spezifisch sind.

Um herauszufinden, welche Erklärungen das Nutzerverständnis der Prognosen steigern, muss erst definiert werden, was Erklärbarkeit und was eine Erklärung ist. Nach Chazette [2] ist Erklärbarkeit die Fähigkeit oder der Akt der Offenlegung von Informationen, die notwendig sind, damit ein Adressat einen bestimmten Aspekt eines Systems in einem bestimmten Kontext verstehen kann. Dies könne durch Bereitstellung von Erklärungen erreicht werden. Weiterhin definiert die Autorin, dass eine Erklärung eine Information

sei, die dazu beitrage, dass ein Adressat ein zu erklärendes Ereignis in einem bestimmten Kontext versteht. Erklärungen, die im Zuge dieser Arbeit entworfen werden, müssen daher an einen Adressaten gerichtet sein und dessen Verständnis eines zu erklärenden Ereignisses in einem bestimmten Kontext beeinflussen. Durch die Prognosen elektrischer Leistung über einen Zeitraum existiert ein zu erklärendes Ereignis, zu dem das Verständnis der Adressaten, den Mitarbeitenden, durch Informationen beeinflusst werden soll. Der Kontext, in dem sich die Adressaten befinden, ist für Mitarbeitende individuell. Folglich gilt es, diese Informationen zu konkretisieren und herauszuarbeiten, in welcher Form und in welchem Umfang diese das Verständnis der Adressaten steigern.

Chazette et al. [3] definieren, dass vor allem Machine-Learning-Modelle zwar akkurate Ergebnisse liefern können, jedoch von Nutzenden oft als sogenannte Black-Boxes gesehen werden, da weder ihr Grundprinzip leicht nachvollziehbar sei, noch die Ergebnisse leicht herleitbar seien. Mohseni et al. [12] definieren Black-Boxes als komplexe Machine-Learning-Modelle, welche nicht von Menschen interpretiert werden können. Chazette und Schneider [4] erläutern, dass Transparenz eines Systems informell als das Gegenteil von Undurchsichtigkeit oder "Blackboxness" definiert werden kann. Um Nutzenden zu helfen, die Black-Box, also das Machine-Learning-Modell, zu verstehen, müsse die Transparenz des Systems gesteigert werden. Chazette et al. [3] erwähnen, dass Erklärungen die Ethik, die Fairness und die Transparenz eines Systems erhöhen. Vor dem Hintergrund seien Erklärungen ein potenziell effektives Mittel, um das Vertrauen in die Machine-Learning-Methoden bei Kraftwerk zu steigern.

Auch Rojat et al. [17] erklären, dass die Genauigkeit eines Modells allein nicht bedeutet, dass das Modell vertrauenswürdig ist. Die Autoren erläutern weiterhin, dass sehr genaue Modelle meist komplexer als weniger performante Modelle sind, jedoch im Gegenzug weniger interpretierbar. Sie definieren Vertrauenswürdigkeit als "das Vertrauen, dass sich ein Modell bei bestimmten Problemen wie vorgesehen verhält" [17] und erläutern, dass Nutzende bei der Arbeit mit dem Modell zuversichtlicher sein könnten, wenn sie seine Funktionsweise verstehen.

Erklärungen von Machine-Learning-Modellen sind nach Mohseni et al. [12] in zwei große Kategorien zu unterteilen: globale und lokale Erklärungen. Die Autoren definieren, dass eine **globale Erklärung** das Modell im Allgemeinen oder die Gesamtheit dessen erklärt, während sich eine **lokale Erklärung** auf einzelne Ausgaben oder Ausgabebereiche bezieht. Dabei können auch Eingabe-Ausgabe Paare von lokalen Erklärungen erläutert werden. Amiri et al. [1] definieren weitere Unterteilungen von Erklärungen. Eine **post-hoc Methode** würde erst nach dem Training des Modells angewandt, während eine **intrinsische Methode** sich auf gesteigerte Interpretierbarkeit durch Reduzierung der Komplexität des Modells beziehe. Weiterhin geben nach den Autoren **Modell-spezifische Methoden** nur Informationen für

einen spezifischen Typ Modell, während **Modell-unabhängige Methoden** bei Modellen jeder Art angewandt werden können. Modell-unabhängige Methoden hätten keinen Zugriff auf die innere Gewichtung eines Modells und ihr Fokus liege auf der Untersuchung der Verhältnisse zwischen Ein- und Ausgabe.

Mohseni et al. [12] definieren Gestaltungsziele für verschiedene Zielgruppen von Erklärungen. Für Neulinge im Bereich des Machine-Learning erheben die Autoren vier Design-Ziele sowie vier Evaluationsmetriken für die Erklärungen. Die Design-Ziele umfassen algorithmische Transparenz, Vertrauen der Nutzenden, Bias-Minderung und Bewusstsein für Datenschutz. Da Mitarbeitende von Kraftwerk bei der Nutzung der Prognosen elektrischer Leistung über Zeit keinerlei personenbezogene Daten preisgeben, ist das Design-Ziel "Bewusstsein für Datenschutz" für diese Arbeit nicht relevant. Die Autoren identifizieren vier Evaluationsmetriken für Neulinge: das mentale Modell der Nutzenden, Zweckmäßigkeit und Zufriedenheit, Vertrauen der Nutzenden sowie Mensch-KI Performance. Das mentale Modell der Nutzenden beschreibe, in Anlehnung an die Theorien der kognitiven Psychologie, die Darstellung dessen, wie Nutzende ein System verstehen. Dabei könne der Grad an Genauigkeit zum tatsächlichen Modell für jede nutzende Person variieren. Nach Kirchhoff [10] kann eine Abweichung des mentalen Modells von dem tatsächlichen Systemmodell zur Frustration bei Endnutzenden führen. Mohseni et al. [12] erläutern, dass eine Prognose der Nutzenden über die Modellausgabe das mentale Modell der Nutzenden messen kann. Die letzte Evaluationsmetrik, die Mensch-KI Performance, wird als Metrik zum Messen einer Performance-Steigerung beschrieben. So könne etwa eine Erfolgsquote oder die Zeit bis zum Abschluss einer Aufgabe gemessen werden.

Weiterhin definieren Mohseni et al. [12], dass sich das Vertrauen von Nutzenden in ein System über die Zeit der Interaktion sowie über mehrere Interaktionen hinweg verändern kann. Sie identifizieren überdies Metriken zur Messung von Vertrauen bei Nutzenden, darunter Selbsteinschätzungen und Fragen mit einer Likert-Skala für die subjektive Messung. Auch für objektive Messungen, etwa die von Nutzenden wahrgenommene Systemkompetenz und Verständlichkeit, seien solche Fragen hilfreich. Die Autoren erklären zudem, dass transparente Erklärungen dazu beitragen könnten, negative Auswirkungen des Vertrauensverlustes in unerwarteten Situationen zu vermindern.

2.4 Relevante XAI-Methoden

Populäre Methoden für Erklärungen von Machine-Learning-Modellen sind LIME [16], SHAP [11], und TREPAN [5].

LIME, auch Local Interpretable Model-agnostic Explanations, ist auf

tabellarische Erklärungen oder Erklärungen in Form von Bildern von Klassifikationsmodellen ausgerichtet. Dieber und Kirrane [8] erklären, dass LIME eine lokale, interpretierbare, Model-agnostische Methode ist, komplexe Machine-Learning Modelle zu erklären. Nach Ribeiro et al. [16] nähert LIME die Entscheidungen eines komplexen Klassifikationsmodells durch das Erstellen eines weiteren, interpretierbaren Modells, an. Die Anwendung von LIME als Methode zur Erklärung von Zeitreihenprognosen ist nach Schlegel et al. [21] durch die Festlegung von Relevanz bei Zeitreihendaten möglich. Dabei beschreibt das durch LIME erstellte Modell nach Dieber und Kirrane [8] die Entscheidungen des zu interpretierenden Modells, indem es die relevanten Einflussfaktoren (Features) und die Wahrscheinlichkeiten des Eintritts (Confidence) zu jeder Klasse darstellt. In den tabellarischen Erklärungen werden Features, die einen Wert innehaben, welcher auf eine andere Klasse hinweist als die übrigen, hervorgehoben. So ließe sich für Nutzende visuell nachvollziehen, welche Entscheidung aufgrund von welchen Features getätigt wurde und welche Werte von Features gegen die Entscheidung sprechen.

Nach Schlegel et al. [21] können auch sogenannte SHAP-Methoden [11] (SHapley Additive exPlanations) durch Festlegung von Relevanz bei Zeitreihendaten zur Erklärung eines Zeitreihen-Modells eingesetzt werden. Die Autoren erläutern, dass SHAP-Methoden zur Erklärung von Entscheidungen eines Modells mittels Shapley Value [23] und Game Theory das relevanteste Feature aufzeigen. Molnar [13] erklärt, dass eine Shapley Value der durchschnittliche Beitrag eines Features über alle möglichen Vereinigungen von Features sei. Nach Russel und Norvig [18] beschreibt Game Theory die Modellierung logischer Entscheidungsfindung bei Spielern mit potenziellem Einfluss auf den Spielausgang. Molnar et al. [14] erklären, dass sich die Idee eines kollaborativen Spiels auf Machine-Learning-Methoden anwenden lässt. So würden die Features, das heißt die Spieler, zusammenarbeiten, um eine Vorhersage, also eine Auszahlung im Spiel, zu treffen. Ozyegen et al. [15] erläutern, dass SHAP eine Model-agnostische post-hoc Methode ist.

Ozyegen et al. [15] beschreiben, dass SHAP, ähnlich wie LIME, ein lokales, lineares Model nutzt, um ein komplexeres Model interpretierbarer darzustellen. Weiterhin könnte durch SHAP die Wichtigkeit der Features (Feature Importance) analysiert werden. So könnte die durchschnittliche sowie die relative Wichtigkeit der verschiedenen Einflussfaktoren auf eine Entscheidung dargestellt werden. Nutzende könnten dadurch einsehen, welche Features den größten oder kleinsten Einfluss auf die Entscheidungen haben. Ähnlich zur üblichen tabellarischen Darstellung der Feature Importance, könnte die Feature Importance eines Modells auch in einem Wärmebild angezeigt werden und es Nutzenden so ermöglichen, schnellstmöglich die wichtigsten Features zu identifizieren.

Der TREPAN-Algorithmus [5] erzeugt einen Entscheidungsbaum, an dem Nutzende die Entscheidungen eines komplexen Modells nachvollziehen

können. Nach De et al. [7] kann TREPAN mittels Annäherung an das Modell und durch das kontinuierliche Hinzufügen neuer Daten immer wieder neue Knoten erstellen. TREPAN kann nach Craven und Shavlik [6] auch Informationen aus Zeitreihen-Modellen extrahieren und darstellen. Ein solcher Entscheidungsbaum würde es Nutzenden ermöglichen, eine Entscheidung durch die im Baum präsentierten Features und ihre Werte nachzuvollziehen.

Kapitel 3

Momentaner Stand und Stakeholder

In diesem Kapitel wird der derzeitige Stand der Systeme zur Vorhersage elektrischer Leistung bei Kraftwerk sowie eine Aufteilung der Stakeholder der Erklärungen in Gruppen und ihre Motivation erläutert.

3.1 Momentaner Stand

Derzeit können Voraussagen der erzeugten elektrischen Leistung von ausgewählten Mitarbeitern im internen Verwaltungs-Tool "Webgate" eingesehen werden. Im Webgate ist unter einer Übersicht, welche für jedes BHKW eingesehen werden kann, ein eigener Reiter für die Prognose eingerichtet. Für jedes BHKW kann dort eine Anlagen-spezifische Vorhersage eingesehen werden. Die momentane Darstellung der Vorhersagen ist Abbildung 3.1 dargestellt.

Dort ist die erzeugte elektrische Leistung über Zeit sowie eine Vorhersage über den Rest des aufgeführten Tages abgebildet. Die Voraussagen sind in einem einfachen Diagramm abgebildet, welches die elektrische Leistung in kW über einen Zeitraum darstellt. Weiterhin kann in demselben Diagramm der bisherige Verlauf der elektrischen Leistung eingesehen werden, farblich von der Vorhersage differenziert. Weitere Inhalte zu der Vorhersage gibt es derzeit nicht. Im Zuge der Implementierung des Vorhersage-Diagramms wurden diejenigen Mitarbeiter, die die Vorhersage einsehen können, mittels kurzer Benachrichtigung im Webgate über das Update informiert. Außerdem wurde ein kurzer Bericht, welcher über künstliche Intelligenz im allgemeineren Sinne informiert, im Firmen-internen Blog veröffentlicht. Der Bericht erläutert Möglichkeiten der Nutzung von Machine-Learning bei Kraftwerk und soll das Verständnis der Mitarbeiter mit geringerer technischer Expertise für das Thema künstliche Intelligenz fördern.

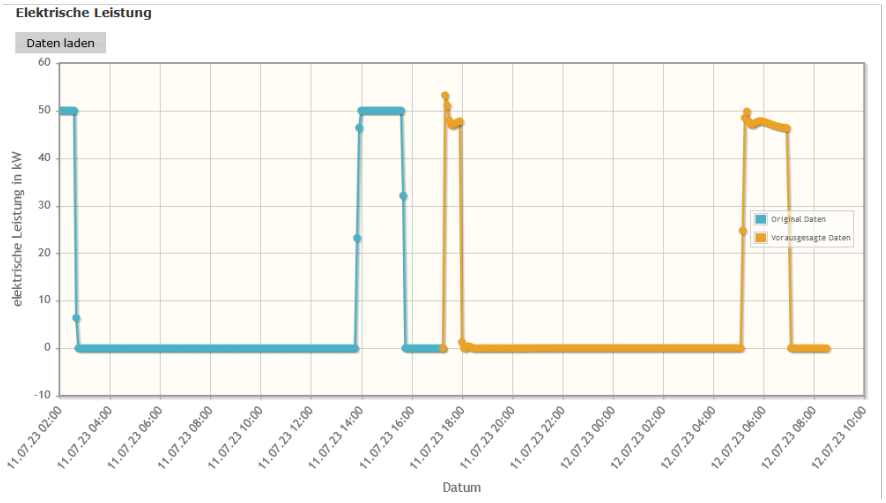

Abbildung 3.1: Anzeige einer Vorhersage im Webgate, Stand der Systeme vor der Arbeit.

3.2 Stakeholder der Erklärungen

Im Weiteren werden die einzelnen Stakeholdergruppen sowie ihre Motivation genauer erläutert. Nach Gesprächen mit verschiedenen Abteilungs-Leitenden stellte sich heraus, dass Mitarbeitende innerhalb derselben Abteilung im Allgemeinen ähnliche oder gleiche Motivationen an den Vorhersagen haben wie ihre Team-Kollegen. Daher wird in dieser Arbeit darauf verzichtet, die Mitarbeitenden als Stakeholder neu zu gruppieren. Im Gespräch mit den Leitern verschiedener Teams von Kraftwerk hat sich ergeben, dass insgesamt acht von zwölf Teams generelles Interesse an den Vorhersagen haben, wobei fünf regelmäßig mit den Vorhersagen arbeiten wollen. Tabelle 3.1 zeigt die acht Teams bzw. Stakeholdergruppen sowie ihre Motivation und eine nähere Beschreibung dieser.

Da die Teams allgemein die Endnutzenden der Vorhersagen sind, ist die typische Stakeholdergruppe “Endnutzende“ als solche hier nicht vertreten. Weiterhin ist auch die typische Gruppe “Entwickelnde“ nicht aufgeführt, da die Autorin dieser Arbeit die alleinige Entwicklerin der betroffenen Systemkomponenten ist. Die Stakeholdergruppen Steuerungstechnik, Maschinenbau, Service-Technik, Service-Büro sowie der Vertrieb sollen regelmäßig auf die Vorhersagen zurückgreifen. Insbesondere die Stakeholdergruppen Maschinenbau sowie Steuerungstechnik sind durch ihre Arbeit an der Effizienzsteigerung derzeitiger Systeme interessiert.

Die **Steuerungstechnik** beschäftigt sich vorwiegend mit der Forschung und Entwicklung neuer Methoden zur Optimierung und Effizienzsteigerung der BHKWs und ihrer Steuerung. Durch die derzeit an Methoden des Machine-Learning orientierte Entwicklung der Forschung und Literatur

Tabelle 3.1: Stakeholdergruppen sowie ihre Motivation an den Vorhersagen

Stakeholdergruppe	**Motivation**	**Motivations-Beschreibung**
Steuerungstechnik	Entwicklung	Entwicklung neuer Techniken und Optimierung derzeitiger Abläufe, Entwicklung und Einschätzung des Nutzens von Lösungen mittels Machine-Learning
Maschinenbau	Entwicklung	Entwicklung neuer Techniken und Optimierung derzeitiger Abläufe
Service-Technik	Wartungsprognosen, Betriebsoptimierung	Optimierung derzeitiger Abläufe sowie bessere Einschätzung von Wartungszeiten der BHKWs
Service-Büro	Wartungsprognosen, Betriebsoptimierung	Optimierung derzeitiger Abläufe sowie bessere Einschätzung von Wartungszeiten der BHKWs
Vertrieb	Betriebsoptimierung	Optimierung derzeitiger Abläufe sowie vertriebliche Vermarktung derzeitiger Projekte
Projektbetreuung	Betriebsoptimierung	Optimierung derzeitiger Abläufe sowie Verbesserung von Lösungsansätzen der Projekte
Buchhaltung	Abrechnungen	Erstellung spezifischer Abrechnungen aufgrund der Vorhersagen bei veränderter Wartungs-Struktur
Geschäftsleitung	Verantwortung	Tragende der Verantwortung über Betriebsabläufe sowie effiziente Unternehmensentscheidungen

sowie das große Interesse an effizienterer Produktion von Energie, liegt es nahe, Machine-Learning Methoden in vielen Bereichen der Steuerung zu implementieren. Die Steuerungstechnik als solche umfasst bei Kraftwerk unter anderem die (Weiter-)Entwicklung der Anlagensteuerung und die Entwicklung interner Software-Tools zur Verwaltung und Fernsteuerung der BHKWs. Die Stakeholdergruppe Steuerungstechnik hat eine stark technische Orientierung, daher geht es diesem Team vor allem um ein umfangreiches technisches Verständnis neuer Methoden.

Weiterhin hat auch die Stakeholdergruppe **Maschinenbau** potenziell ein großes Interesse an einer technischen Auseinandersetzung mit Machine-Learning Methoden. Durch die Vorhersage der erzeugten elektrischen Leistung ergeben sich viele Möglichkeiten der Anlagenoptimierung im Rahmen der Forschung und Entwicklung. Ähnlich zur Steuerungstechnik befasst sich der Maschinenbau bei Kraftwerk vorrangig mit der (Weiter-)Entwicklung der Anlagen und hat eine stark technische Orientierung.

Die Stakeholdergruppe **Geschäftsleitung** soll, begründet mit ihrer Natur der Verantwortungsübernahme, zusätzlich ein besonderes Verständnis gegenüber den Vorhersagen aufbringen. Ihnen obliegt es nicht nur, die Effizienz der einzelnen BHKWs zu steigern, sondern auch die Verantwortung über die Arbeit von Mitarbeitenden zu übernehmen und das Unternehmen in der freien Wirtschaft zu repräsentieren. Da alle Personen aus dem Team Geschäftsleitung auch Teammitglieder anderer Teams sind, ist ihre Motivation an den Vorhersagen nicht nur durch das Tragen der Verantwortung, sondern auch durch die Arbeit in ihrem zweiten Team geprägt.

Im starken Kontrast dazu stehen die Stakeholdergruppen **Service-Technik** sowie **Service-Büro**. Das Service-Büro beschäftigt sich primär mit der Wartung schon installierter Anlagen und dem Service für Kunden. Es ist hauptsächlich an der Optimierung und Wartung laufender Anlagen interessiert und verspricht sich durch Machine-Learning-Methoden eine effizientere Planung von Wartungsaktivitäten. Auch die Service-Technik ist hauptsächlich mit der Wartung installierter Anlagen beschäftigt. Die Mitarbeitende dieser Stakeholdergruppe stellen sich durch die Prognosen eine effizientere Lösungsfindung bei Problemen und ein tieferes Verständnis von auftretenden Anomalien in Aussicht.

Ebenso ist die Stakeholdergruppe **Vertrieb** an den Prognosen interessiert. Der Vertrieb verspricht sich nicht nur eine Optimierung des Betriebes, sondern darüber hinaus auch eine vertriebliche Vermarktung der Prognosen. Die Mitarbeitenden im Vertrieb kommen größtenteils von einem nicht-technischen Hintergrund, deswegen ist diese Stakeholdergruppe als weniger technikaffines Team bei der Entwicklung von Erklärungen besonders relevant. Der Vertrieb ist insbesondere an einem oberflächlichen Verständnis der Prognosen interessiert.

Ähnlich zum Vertrieb ist auch die Stakeholdergruppe **Projektbetreuung** an einem vordergründigen Verständnis der Vorhersagen interessiert.

Durch ihre große thematische Entfernung gegenüber der technischen Seite des Machine-Learning geht es auch dieser Stakeholdergruppe um eine einfache Arbeit mit den Prognosen. Zuletzt tritt auch die **Buchhaltung** durch die Notwendigkeit der Erstellung spezieller Abrechnungen mit den Prognosen in Kontakt. Auch für dieses Team ist der genaue technische Hintergrund der Vorhersagen nicht von Interesse.

Kapitel 4

Forschungsmethodik

In diesem Kapitel wird die angewandte Forschungsmethodik sowie die Forschungsfragen dieser Arbeit erläutert. Auch die Entwicklung des Prototyps, welcher im Zuge dieser Arbeit erstellt wurde, wird beschrieben.

4.1 Forschungsfragen

Um ein Konzept für Erklärungen der Voraussagen bei Kraftwerk zu entwerfen, stellt sich vorerst die Frage, inwiefern sich die Anforderungen der Mitarbeitenden voneinander unterscheiden. Durch die Aufteilung der Stakeholder in Stakeholdergruppen ergeben sich mehrere Gruppen mit verschiedenen Motivationen an den Erklärungen. Die erste Forschungsfrage bezieht sich auf Unterschiede im Bedarf nach Erklärungen verschiedener Stakeholdergruppen. Dies stellt sicher, dass auch Stakeholder mit verschiedener Motivation einen Mehrwert aus den Erklärungen ziehen können. Die zweite Forschungsfrage bezieht sich auf die spezifischen Anforderungen der Stakeholdergruppen und bietet explizit Raum, das derzeitige System bei Kraftwerk zu verbessern.

4.1.1 RQ1: Unterschiede im Bedarf nach Erklärungen

Durch das Herausstellen der Unterschiede im Bedarf nach Erklärungen der Stakeholdergruppen kann sichergestellt werden, dass auf die verschiedenen Motivationen dieser zur Nutzung der Prognosen eingegangen wird. Folglich hat etwa der Vertrieb, der hauptsächlich an der Betriebsoptimierung interessiert ist, eine andere Motivation, die Prognosen zu nutzen, als der Maschinenbau, welcher die Prognosen primär für die Entwicklung nutzt. Um diese Unterschiede in der Motivation nicht zu vernachlässigen, müssen die Differenzen des Bedarfs nach Erklärungen herausgearbeitet werden.

RQ1: Wie unterscheidet sich der Bedarf nach Erklärungen innerhalb der Stakeholdergruppen?

4.1.2 RQ2: Verbesserungen des bestehenden Systems durch Erklärungen

Die zweite Forschungsfrage geht auf die spezifischen Verbesserungen ein, die am bestehenden System bei Kraftwerk vorgenommen werden können, um die Arbeit mit den Voraussagen zu verbessern. Es geht primär darum, die unterschiedlichen Anforderungen der Stakeholdergruppen zu kombinieren und dabei der Gesamtheit der Stakeholdergruppen Erklärungen zu liefern, die ihre Anforderungen abdecken.

RQ2: Wie kann das bestehende System durch Erklärungen an den Voraussagen verbessert werden?

Um die Forschungsfragen zu bearbeiten und deren Beantwortung nah an den Endnutzenden zu erarbeiten, hat sich für diese Arbeit eine Umfrage angeboten. So können möglichst viele Teams mit verschiedener Motivation und abweichendem technischem Hintergrund einen Einfluss auf die Konzeptentwicklung nehmen. Eine andere mögliche Methode der Anforderungserarbeitung wären Interviews mit den Endnutzenden gewesen, jedoch sollte eine möglichst hohe Anzahl der Endnutzenden befragt werden und das Halten von Interviews hätte die Zahl an Teilnehmenden aus Zeitgründen stark begrenzt.

4.2 Methodologie

Um sinnvolle Metriken für die Beantwortung der Forschungsfragen zu finden, wurde die Goal-Question-Metric (GQM) Methode genutzt. Dabei wurde als Goal die Evaluation des Bedarfs an Erklärungen der Voraussagen bei Kraftwerk gesetzt.

G1: Evaluation des Bedarfs an Erklärungen der Voraussagen bei Kraftwerk

Aus diesem Ziel ergeben sich drei Fragen, die in Abbildung 4.1 gezeigt werden. Die Frage $Q1$ bezieht sich auf den derzeitigen Stand der Systeme bei Kraftwerk. Diese Frage ist wichtig, um einen grundsätzlichen Bedarf an Erklärungen bei den Stakeholdergruppen festzustellen und $RQ1$ 4.1.1 zu beantworten.

Um eine Antwort auf $Q2$ zu finden, werden in der Umfrage zu Erklärungen von Voraussagen bei Kraftwerk, mehrere Beispiele gegeben. Diese Frage ist relevant, um herauszufinden, welche grundsätzlichen Typen von Erklärungen das Verständnis und Vertrauen der Teilnehmenden in die Voraussagen

steigern können. Durch Frage $Q3$ können explorativ weitere Alternativen von Erklärungen erforscht werden. Weiterhin dient sie, zusammen mit $Q2$, dazu, die zweite Forschungsfrage 4.1.2 zu beantworten.

Q1: Wie gut wird das System, wie es ist, verstanden?
Q2: Inwiefern helfen die gegebenen Erklärungsbeispiele, das System zu verstehen?
Q2: Was für Erklärungen wünschen sich die Teilnehmenden?

Aus den GQM-Fragen ergeben sich zwei Metriken, die zur Messung und Beantwortung der Fragen dienen. Es bietet sich an, die Performance der Teilnehmenden abzufragen und dadurch herauszufinden, ob bestimmte Erklärungen diese beeinflussen. Daher sollen Teilnehmende der Umfrage daraufhin abgefragt werden, ob sie selbst eine Prognose treffen können. Damit können die Antworten der Teilnehmenden direkt mit dem tatsächlichen Wert verglichen und Differenzen evaluiert werden.

M1: Prüfung der Korrektheit der Prognosen der Teilnehmenden
M2: Selbsteinschätzungen

So kann die Antwort auf die Fragen $Q1$ sowie $Q2$ durch $M1$, eine Performance-Abfrage der Teilnehmenden, gemessen werden. Die Frage $Q3$ kann durch $M2$, Selbsteinschätzungen der Teilnehmenden, beantwortet werden.

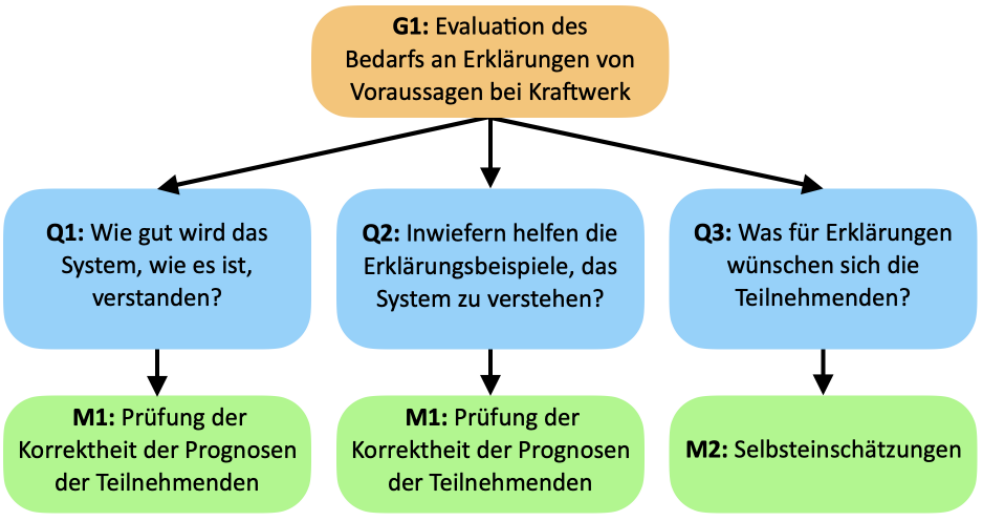

Abbildung 4.1: Goal-Question-Metric

4.2.1 Umfrage

In diesem Abschnitt wird die Umfrage, die im Zuge dieser Arbeit bei Kraftwerk ausgeführt wurde, erläutert. Die Umfrage wurde mit *LimeSurvey*[1] durchgeführt und durch Beispiele auf einem Prototyp unterstützt. Der Prototyp setzt sich aus mehreren Beispielen zusammen, welche in Kapitel 4.2.2 erläutert werden. Zu jedem der genannten Beispiele wurde ein Set aus Fragen erstellt, welches jeweils aus drei Kategorien bestand.

Zuerst wurden Teilnehmende gebeten, in einer Performance-Abfrage selbst eine Prognose abzugeben, um ihr mentales Modell zu messen, wie in Kapitel 2 erklärt. Dazu wurde in den Beispielen nicht nur die erzeugte elektrische Leistung über Zeit, sondern auch der tägliche Durchschnitt dieser angegeben. Zum letzten Tag jeder Prognose sollten Teilnehmende angeben, ob der Tagesdurchschnitt steigen, sinken, oder gleich bleiben würde. Weiterhin wurden Teilnehmende gefragt, wieso sie sich für ihre Antwort entschieden haben.

In der zweiten Fragekategorie je Beispiel wurden den Teilnehmenden mehrere Fragen mittels einer Likert-Skala gestellt. In dieser hatten Teilnehmende die folgenden Antwortmöglichkeiten der Zustimmung.

1. Stimme überhaupt nicht zu (1)
2. Stimme nicht zu (2)
3. Stimme weder zu noch nicht zu (3)
4. Stimme zu (4)
5. Stimme völlig zu (5)

Dabei wurden die Teilnehmenden unter anderem gefragt, ob sie sich bei ihrer Prognose sicher seien und ob ihnen die gezeigten Erklärungselemente bei der Beantwortung der Performance-Abfrage geholfen haben. Weiterhin wurden Teilnehmende gefragt, ob sie einschätzen könnten, ob die Prognosen tatsächlich eintreten und ob sie den Umfang der gegebenen Erklärungen für angemessen halten. Zuletzt hatten Teilnehmende in der dritten Fragekategorie die Möglichkeit anzugeben, was für Elemente ihnen in der Darstellung gefehlt haben, oder welche Elemente sie nicht verstanden haben.

Am Ende der Umfrage wurden demographische Fragen gestellt, unter anderem für wie technikaffin sich Teilnehmende halten und wie lange sie schon bei Kraftwerk oder einem vergleichbaren Unternehmen angestellt sind. Ein vergleichbares Unternehmen wurde in der Umfrage als ein Unternehmen definiert, welches entweder auch mit Blockheizkraftwerken arbeitet, oder in der Energie-Branche sehr eng mit Kraftwerk verwandt ist. Die Frage nach der

[1] `https://www.limesurvey.org/de/`, Stand 11.07.2023 15:45 Uhr

Technikaffinität der Teilnehmenden setzte sich aus drei Teilfragen zusammen und konnte mittels der Likert-Skala beantwortet werden.

1. Bei Problemen am Computer (z. B. kein Internetzugriff) frage ich jemand anderen nach einer Lösung.
2. Ich behebe Probleme am Computer zum größten Teil selbst.
3. Ich halte mich für technikaffin.

Weiterhin sollten Teilnehmende angeben, in welchem Team sie bei Kraftwerk arbeiten.

4.2.2 Prototyp

Teilnehmende wurden im Zuge der Umfrage zu einem Prototyp mehrere Fragen, welche in Kapitel 4.2.1 erklärt wurden, gestellt. Es wurde eine externe Website erstellt, um mehrere Beispiele von Erklärungen an den Prognosen darzustellen. Die Website wurde mit Javascript, HTML und SCSS realisiert und ist eine rein statische Website. Zur Darstellung der Diagramme wurde die Library *ChartJS*[2] benutzt. Teilnehmende müssen, ausgenommen der Navigation durch die Website, keine Aktionen ausführen. Die Website ist in sieben Schritte aufgeteilt, wobei der **erste Schritt** eine Willkommensnachricht enthält und kurz den Umfrageverlauf beschreibt. Die weiteren sechs Schritte der Website beinhalten jeweils eine Anzeige eines Verlaufs elektrischer Leistung sowie einer Prognose sowie je ein Beispiel für Erklärungen.

Schritt zwei zeigt ein Diagramm der elektrischen Leistung eines Blockheizkraftwerks. Die vereinfachte Darstellung kann in Abbildung 4.2 eingesehen werden. Das Diagramm zeigt den bisherigen Verlauf elektrischer Leistung der letzten zwei Tage sowie eine Prognose. Weiterhin wird die durchschnittliche elektrische Leistung pro Tag angezeigt, für den Verlauf sowie für die Prognose. Der Wert wird in kW gemessen und dargestellt. Ein Diagramm dieser Art wird im Folgenden als "Standard-Diagramm" (SD) bezeichnet, da es, ausgenommen des Durchschnitts-Wertes, dem derzeitigen Stand der Anzeige im internen Verwaltungs-Tool entspricht und noch nicht um Erklärungen bereichert wurde. Das Standard-Diagramm ist in allen Beispielen enthalten und wird durch die jeweiligen Erklärungen ergänzt.

In **Schritt drei** wird das Standard-Diagramm mit zwei Graphen und einem kurzen Text erweitert. Es wird die Feature Importance dargestellt, erläutert in Kapitel 2.4, sowie der Verlauf der Wetterverhältnisse und das prognostizierte Wetter der nächsten zwei Tage. Dabei enthält das Diagramm der Wetterverhältnisse die Werte für die Außentemperatur in Grad Celsius, die durchschnittliche Luftfeuchtigkeit sowie der Niederschlag

[2] `https://www.chartjs.org/`, Stand 11.07.2023 15:45 Uhr

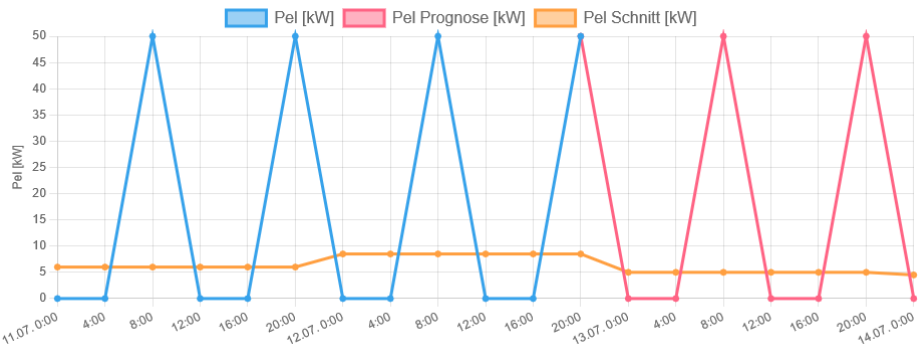

Abbildung 4.2: Standard Diagramm aus dem Prototyp, Darstellung der vergangenen Datenpunkte über Zeit sowie einer Vorhersage

in Prozent und die durchschnittliche Windgeschwindigkeit in km/h. Da diese Art von Diagramm auch in anderen Schritten wiederkehrt, wird auf dieses als "Wetter-Diagramm" Bezug genommen. Das Diagramm "Parameter und ihre "Wichtigkeit" für die Prognose" zeigt Einflussfaktoren des Modells und ihre relative Wichtigkeit. Der kurze Text-Abschnitt erläutert, dass die "Wichtigkeit" eines Parameters den Grad an Einfluss beschreibt, den dieser Parameter auf die Gesamtheit der Prognose genommen hat. Dieses Beispiel stellt die globale Erklärung der Features (GlobalFeat) dar, das heißt, die Erklärung bezieht sich auf das Modell insgesamt und nicht auf einzelne Datenpunkte.

In **Schritt vier** werden Teilnehmende dazu aufgefordert, mit dem Mauszeiger über die Datenpunkte der Prognose zu fahren. Beim Herüberfahren über einen Datenpunkt der Prognose öffnet sich ein Tooltip, in dem weitere Informationen über diesen Wert der Prognose vermittelt werden. Für jeden Zeitschritt wird erklärt, welche Parameter den meisten Einfluss auf diesen Wert genommen haben. Dabei wird die Feature Importance des jeweiligen Parameters genannt. Weiterhin wird im Tooltip die Confidence in die Entscheidung, erläutert in Kapitel 2.4, genannt sowie kurz erläutert. Überdies ist unter dem Standard-Diagramm ein Wetter-Diagramm abgebildet. Dieses Beispiel stellt die lokale Erklärung der Features (LocalFeat) dar.

Schritt fünf beinhaltet ein Standard-Diagramm sowie ein Wetter-Diagramm. Zusätzlich wird ein Entscheidungsbaum in Anlehnung an den in Kapitel 2.4 erläuterten TREPAN-Algorithmus dargestellt und durch einen kurzen Text erklärt. Der Entscheidungsbaum kann in Abbildung 4.3 eingesehen werden. Er nähert die Entscheidungen des Modells an, wobei bestimmte Schwellenwerte für jedes Feature gegeben werden. Je nachdem, wie der tatsächliche Wert des Features im Vergleich zu dessen Schwellenwerten steht, verfällt die Prognose anders. Um den Entscheidungsbaum übersichtlich und interpretierbar zu halten, wird in den Entscheidungskno-

ten der Wert der durchschnittlichen elektrischen Leistung an einem Tag dargestellt. Teilnehmende können anhand des Wetter-Diagramms erkennen, welche Schwellenwerte zu welchem Zeitpunkt erreicht werden, und somit die Entscheidungen des Modells nachvollziehen. Dieses Beispiel stellt die globale Erklärung mittels Entscheidungsbaums (DT) dar.

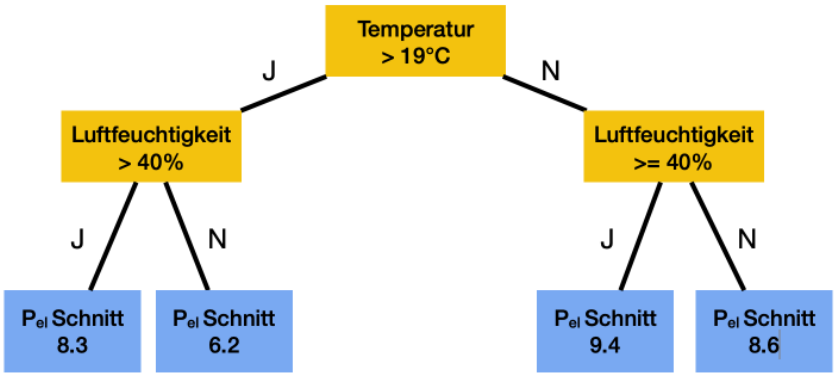

Abbildung 4.3: Darstellung aus dem Prototyp, Entscheidungsbaum, welcher die Entscheidungen des Modells annähert

Schritt sechs zeigt, vergleichbar mit Schritt vier, erneut ein Standard-Diagramm sowie ein Wetter-Diagramm. Weiterhin weist ein kurzer Text Teilnehmende darauf hin, dass sie mit dem Mauszeiger über Datenpunkte des Diagramms fahren können, um mehr Informationen über die Prognose zu erhalten. Beim Herüberfahren über einzelne Datenpunkte der Prognose mit dem Mauszeiger öffnet sich erneut ein Tooltip. Dieses beinhaltet einen Fließtext, welcher die Features für den jeweiligen Zeitschritt sowie deren Wert erklärt. Zusätzlich wird die Confidence in die Vorhersage erläutert und angezeigt. Dieses Beispiel stellt die lokale textuelle Erklärung der Voraussage (LocalText) dar.

Im **siebten Schritt** sehen Teilnehmende ein Standard-Diagramm sowie einen Fließtext. In diesem Text wird oberflächlich erläutert, wie ein LSTM-Modell strukturiert ist. Dabei werden kurz auf Begriffe wie “Training“, “Gedächtnis“ und “Schichten“ eingegangen. Zudem werden zwei Grafiken dargestellt. Die erste Grafik zeigt ein stark abstrahiertes Netzwerk von Neuronen und Schichten, die untereinander verbunden sind. Die zweite Grafik stellt, auch stark vereinfacht, eine LSTM-Zelle dar, indem Eingangs-, Vergessen- und Ausgangs-Tor gezeigt sowie eine einfache Rückkopplung dargestellt werden. Diese kann in Abbildung 4.4 eingesehen werden. Zuletzt wird der Text in einem Abschnitt “In Kurz“ stark zusammengefasst. Dieses Beispiel stellt die globale textuelle Erklärung des Modells (GlobalModel) dar.

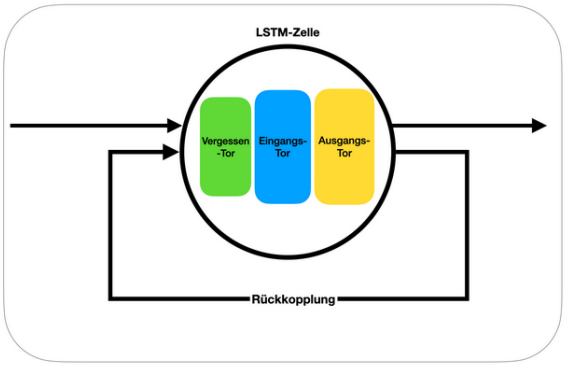

Abbildung 4.4: Darstellung aus dem Prototyp, eine LSTM-Zelle, stark vereinfachte Darstellung des eingesetzten Modells

4.2.3 Pilotierung

Bevor die Umfrage für Teilnehmende von Kraftwerk freigeschaltet wurde, wurden insgesamt drei Pilotierungen durchgeführt. Die **erste Pilotierung** wurde mit einem Mitarbeiter aus dem Team Steuerungstechnik bei Kraftwerk durchgeführt. Dabei ging es hauptsächlich darum, die Beispiele des Prototyps für Mitarbeitende von Kraftwerk anzupassen. Auch der Umfang der Umfrage wurde bearbeitet, da die vorerst neun Beispiele in einer Umfragedauer von über einer Stunde resultierten. Eine globale numerische Erklärung der Fehlerwerte des Modells ist zwar für Entwickler eine interessante Angabe zum Modell, jedoch nicht zielführend für Endnutzende. Die Stakeholder sind nicht mit den genauen Begrifflichkeiten der numerischen Eigenschaften vertraut, und können die Werte auch nicht mit Erfahrungswerten abgleichen. Deshalb wurde dieses Beispiel entfernt. Eine Darstellung eines Entscheidungsbaumes für jeden Zeitschritt gleicht der Darstellung des globalen Entscheidungsbaumes und führt keine neuen Erklärungsmethoden ein. Daher wurde sie ebenso entfernt.

Die **zweite Pilotierung** wurde mit einer studierenden Person der Informatik durchgeführt. Die Person hat über vier Jahre Arbeitserfahrung in der Software-Entwicklung. In diesem Durchlauf wurden redundante Fragen entfernt und dadurch die Umfragedauer weiter gekürzt.

Die **dritte Pilotierung** wurde mit einem wissenschaftlichen Mitarbeiter des Instituts für Praktische Informatik im Fachgebiet Software Engineering an der Gottfried Wilhelm Leibniz Universität abgehalten. Auch hier wurde bestätigt, dass die bereits entfernten Beispiele am wenigsten zur Erhebung von Erklärungsanforderungen beitragen. Insbesondere die Bezeichnung der dargestellten Graphen wurde verbessert. Außerdem wurden Fragen vereinfacht und Anweisungen der Umfrage deutlicher gemacht.

Insgesamt wurden mehrere Fragen sowie zwei Beispiele vom Prototyp

entfernt. Weiterhin wurden Begriffe deutlicher erklärt und für Mitarbeitende von Kraftwerk angepasst. Um den Pool der Teilnehmenden nicht weiter zu verkleinern, wurden keine Personen mit geringerer technischer Affinität bei Kraftwerk zur Pilotierung gewählt.

4.2.4 Kodierung

Die Fragen, die in eigenen Worten beantwortet werden konnten, wurden in zwei Schritten zur Auswertung kodiert: In vivo Kodierung und durch Pattern Codes. Nach Saldaña [19] werden bei der In Vivo Kodierung Codes aus den eigenen Antworten der Teilnehmenden genutzt. In Vivo Codes würden vor allem dabei helfen, die Nuancen der Ansichten der Teilnehmenden in der Kodierung nicht zu verlieren. Im zweiten Schritt wurde Pattern Coding angewandt. Nach Saldaña [19] werden beim Pattern Coding untereinander ähnliche Codes herausgearbeitet und gruppiert. Auch würden Pattern Codes nicht nur sortieren, sondern den Codes weiterhin eine Bedeutung zuschreiben. Der Autor erläutert, dass Pattern Codes erklären oder schlussfolgern und ein Thema, eine Konstruktion oder Erklärung identifizieren. Durch die Zusammenfassung der Codes in Muster würden sie dabei helfen, aus einer Menge an Antworten kleine Einheiten zu entwickeln. Eine detaillierte Aufstellung aller erhobenen Codes kann in Tabelle A.2 entnommen werden.

Kapitel 5

Ergebnisse

In diesem Kapitel werden die Ergebnisse der Umfrage bei Kraftwerk erläutert. Insgesamt haben 23 Mitarbeitende der ausgewählten Teams die Umfrage vollständig ausgefüllt. Weitere 37 Mitarbeitende haben die Umfrage begonnen, aber nicht vollständig beendet. Für diese Arbeit wurden ausschließlich die vollständigen Antworten ausgewertet.

5.1 Demographie

Aus fast jeder Stakeholdergruppe nahmen mindestens zwei Personen an der Umfrage teil, nur aus dem Team Buchhaltung hat niemand die Umfrage abgeschlossen. Eine Person gab an, in einem anderen Team als den genannten zu sein. Unter den Teilnehmenden gaben 81,61 % bei der Frage nach ihrer Geschlechtsidentität “Männlich“ an. Jeweils eine Person identifiziert sich als “Weiblich“ oder “Sonstige“. Zwei Personen enthielten sich ihrer Antwort. Die Ergebnisse der Frage nach der Geschlechtsidentität der Teilnehmenden sind im Anhang in Tabelle A.1 dargestellt. Die Mehrheit der Teilnehmenden (56,39 %) gab ein Alter von unter 40 Jahren an. Außerdem gaben 47,83 % der Teilnehmenden an, schon über sieben Jahre bei Kraftwerk oder einem vergleichbaren Unternehmen angestellt zu sein. 8,7 % gaben an, fünf bis sieben Jahre bei Kraftwerk oder einem vergleichbaren Unternehmen zu arbeiten, 17,39 % gaben an, drei bis fünf Jahre angestellt zu sein. 26,09 % der Teilnehmenden gab an, weniger als drei Jahre bei Kraftwerk oder einem vergleichbaren Unternehmen angestellt zu sein. Damit arbeitet die Mehrheit der Teilnehmenden (56,53 %) seit über fünf Jahren bei Kraftwerk oder in einem vergleichbaren Unternehmen.

34,78 % gaben an, seit über sieben Jahren mit den derzeitigen Systemen bei Kraftwerk zu arbeiten, 26,8 % gaben an, fünf bis sieben Jahre mit den Systemen zu arbeiten. 39,13 % gaben an, seit weniger als drei Jahren mit den Systemen zu arbeiten.

Die Mehrheit der Teilnehmenden (82,61 %) gaben an, dass sie sich

für technikaffin halten. Lediglich eine teilnehmende Person gab an, sich für nicht-technikaffin zu halten. Auch gaben nur zwei Teilnehmende an, dass sie Probleme an ihrem Computer zum größten Teil nicht selbst beheben. Dahingegen gab die Mehrheit (52,17 %) der Teilnehmenden an, bei Problemen am Computer jemand anderen nach einer Lösung zu fragen.

5.2 Performance-Abfrage

Die allgemeine Korrektheit der Antworten der Teilnehmenden in der Performance-Abfrage ist in Abbildung 5.1 zu sehen. Dort sind alle Beispiele sowie die erreichte Korrektheit der Teilnehmenden in Prozent dargestellt.

Die Teilnehmenden beantworteten die Performance-Abfrage zu Beispiel DT in Schritt fünf des Prototyps, wie in Kapitel 4.2.2 erklärt, zu 82,6 % korrekt. Damit war die Korrektheit bei der Erklärung mittels Entscheidungsbaums die höchste, gefolgt von einer Korrektheit von 78,3 % in Schritt vier, der Erklärung LocalFeat. Mit einer Korrektheit von 73,9 % lag auch die Erklärung GlobalFeat in Schritt drei über der Korrektheit des Standard-Diagramms, welche nur eine Korrektheit von 21,7 % aufwies. Von den neu eingeführten Beispielen schnitt die Erklärung GlobalModel allgemein mit 26,1 % am schlechtesten ab. Die Erklärung LocalText wies eine Korrektheit von 46,7 % auf.

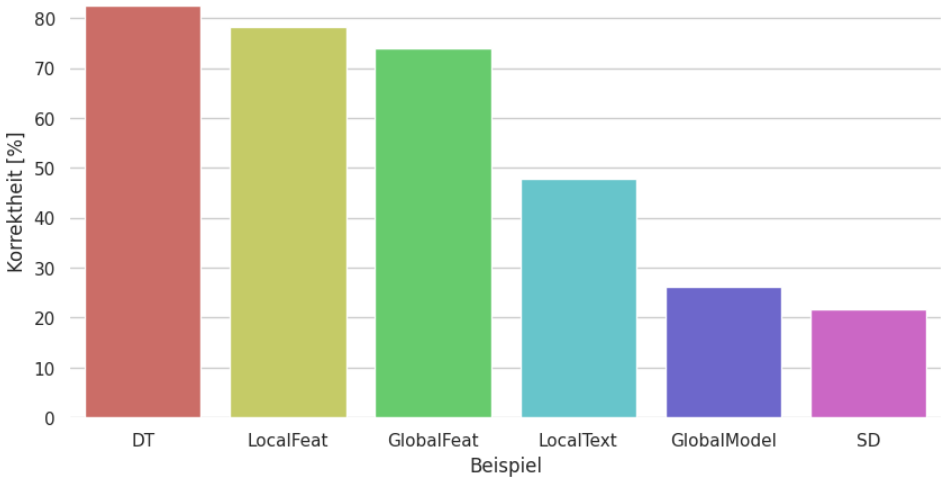

Abbildung 5.1: Allgemeine Korrektheit der Antworten je Beispiel in Prozent

Die Performance der jeweiligen Stakeholdergruppen je Beispiel sowie die durchschnittliche Korrektheit der Gruppen ist in Tabelle 5.1 dargestellt. Die Stakeholdergruppe Steuerungstechnik beantwortete die Performance-Abfrage sowohl zum Beispiel SD als auch zur Erklärung GlobalFeat zu 100 % korrekt. Die Mitarbeitenden dieser Gruppe gaben an, dass ihnen die

Erklärung GlobalFeat sowie die Erklärung DT am meisten geholfen haben. Weiterhin hielt diese Stakeholdergruppe die Angabe des Wetters immer für hilfreich. Dahingegen gaben diese Mitarbeitenden bei der Frage, ob die Erklärung GlobalModel hilfreich sei, mindestens "Stimme nicht zu." an.

Die Stakeholdergruppe Maschinenbau beantwortete die Performance-Abfrage bei den Erklärungen GlobalFeat, LocalFeat und DT zu 100 % korrekt. Die Mitarbeitenden dieses Teams stimmten alle völlig zu, dass ihnen die Erklärung DT zusammen mit der Darstellung des Wetters geholfen hätte. Dahingegen stimmten die Befragten dieser Stakeholdergruppe mindestens nicht zu, dass ihnen die Erklärung LocalText und die Erklärung GlobalModel geholfen hätten.

Die Stakeholdergruppe Service-Technik beantwortete die Abfrage bei den Erklärungen GlobalFeat, LocalFeat, DT sowie LocalText zu 100 % korrekt. Die Teilnehmenden aus dieser Gruppe stimmten nur bei der Erklärung DT zu, dass ihnen die Darstellung geholfen habe. Insbesondere stimmten sie mindestens nicht zu, dass ihnen die Erklärung LocalText sowie GlobalModel geholfen hätte.

Die Stakeholdergruppen Service-Büro, Geschäftsleitung sowie Vertrieb beantworteten nur das Beispiel DT zu 100 % korrekt. Die Teilnehmenden dieser Gruppen waren neutral oder gaben an, dass ihnen die Beispiele der Erklärung LocalText sowie GlobalModel nicht geholfen hätte. Lediglich eine teilnehmende Person aus dem Vertrieb stimmte zu, dass die Erklärung GlobalModel bei der Performance-Abfrage geholfen hätte.

Die Stakeholdergruppe Projektbetreuung lag nur bei der Erklärung LocalFeat bei einer vollständigen Korrektheit. Die Teilnehmenden dieser Gruppe stimmten mindestens zu, dass ihnen die Erklärungen DT sowie GlobalFeat geholfen hätte. Zur globalen Erklärung des Modells stimmten die Teilnehmenden höchstens neutral.

Die teilnehmende Person, welche "Sonstige" als Team angegeben hatte, und somit nicht in die anderen Stakeholdergruppen gezählt werden kann, hat keine der Performance-Abfragen richtig beantwortet.

Tabelle 5.1: Korrektheit bei der Performance-Abfrage je Stakeholdergruppe je Beispiel sowie durchschnittlich in Prozent

Stakeholder-gruppe	**DT**	**Local Feat**	**Global Feat**	**Local Text**	**Global Model**	**SD**	ϕ
Steuerungs-technik	67%	67%	100%	100%	33%	0%	61%
Maschinen-bau	100%	100%	100%	33%	67%	67%	78%
Vertrieb	100%	67%	67%	0%	33%	33%	50%
Projekt-betreuung	60%	100%	60%	60%	0%	0%	47%
Geschäfts-leitung	100%	50%	50%	50%	0%	50%	50%
Service-Büro	100%	75%	75%	25%	25%	25%	54%
Service-Technik	100%	100%	100%	100%	50%	0%	75%

Im Allgemeinen identifizierten sich 34,9 % der Teilnehmenden als nicht technikaffin. Eine teilnehmende Person ist nicht technikaffin, wenn sie mindestens eine der Fragen aus Tabelle 5.2 dementsprechend beantwortet hat. Nicht-technikaffine Mitarbeitende haben die Performance-Abfrage mit der Erklärung DT am häufigsten korrekt beantwortet (87,5 %).

Tabelle 5.2: Antworten, die Technikaffinität verneinen

Frage	**Antwort**
Bei Problemen am Computer (z.B. kein Internetzugriff) frage ich jemand anderen nach einer Lösung.	Mindestens “Stimme zu“
Ich behebe Probleme am Computer zum größten Teil selbst.	Mindestens “Stimme nicht zu“
Ich halte mich für technikaffin	Mindestens “Stimme nicht zu“

5.3 Einschätzung des Prognoseeintritts

Allgemein gaben Teilnehmende an, dass sie bei der Erklärung DT am besten einschätzen könnten, ob die Prognose tatsächlich eintritt. Aus der Likert-Skala ergab sich hier ein Schnitt von 2.91. Die Erklärung GlobalModel gab den Teilnehmenden als einzige das Gefühl, schlechter einschätzen zu können, ob die Prognose tatsächlich eintritt. Zum Standard-Diagramm ergab sich hier ein Schnitt von 2.39.

Der Durchschnitt der Antworten zur Frage, ob die Teilnehmenden einschätzen können, ob die Prognosen tatsächlich eintreten, ist in Abbildung 5.2 dargestellt.

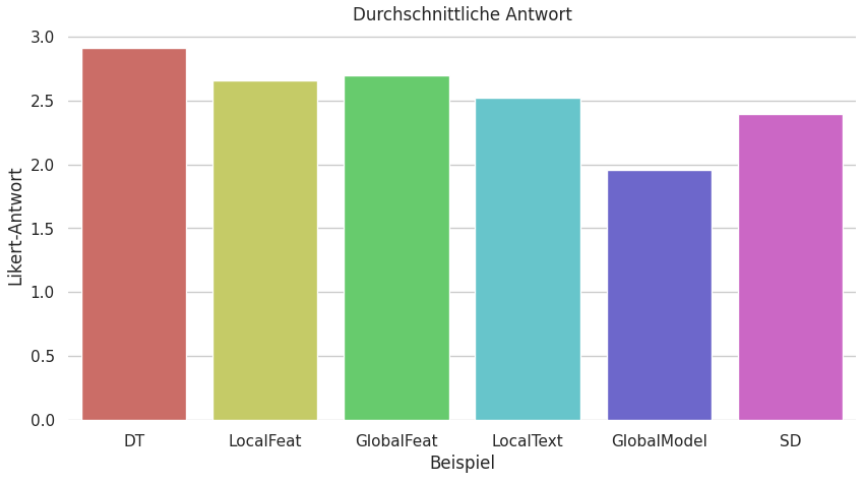

Abbildung 5.2: Durchschnittlicher Likert-Wert der Antworten der Teilnehmenden, zur Frage ob sie einschätzen können, ob die Prognosen tatsächlich eintreten.

5.4 Ergebnisse der Abfrage nach Nützlichkeit

Innerhalb der Umfrage wurden Teilnehmende gefragt, für wie nützlich sie bestimmte Elemente der Beispiele im Prototyp hielten. Diese Fragen konnten mit der Likert-Skala, wie sie in Kapitel 4.2.1 erklärt wird, beantwortet werden. Teilnehmende beantworteten die Frage, ob sie die Wetter-Diagramme als hilfreich empfanden, zu 72,82 % mit mindestens “Stimme zu“. 56,52 % der Teilnehmenden stimmten überhaupt nicht zu, dass ihnen die Erklärung GlobalModel geholfen hätte. 78,26 % der Teilnehmenden stimmten mindestens nicht zu, dass sie die Erklärung LocalText als hilfreich empfinden würden. Die Beispiele und Prozent der Teilnehmenden, die mindestens zustimmten, dieses als hilfreich empfunden zu haben, sind in Tabelle 5.3 dargestellt.

Tabelle 5.3: Beispiele und Prozent der Teilnehmenden, die mindestens zustimmten, das jeweilige Beispiel als hilfreich empfunden zu haben

Beispiel	Stimme zu	Stimme völlig zu	Summe
GlobalFeat	48 %	26 %	74 %
LocalFeat	43 %	9 %	52 %
DT	35 %	48 %	83 %
LocalText	9 %	0 %	9 %
GlobalModel	4 %	0 %	4 %

5.5 Ergebnisse der Kodierung

Nach der Kodierung der Freitext-Antworten, wie in Kapitel 4.2.4 beschrieben, ergaben sich 13 Codes. Die Häufigkeiten der Pattern Codes sind im Anhang in Tabelle A.2 dargestellt. In Tabelle A.3 ist eine Legende der Pattern Codes und ihrer Abkürzungen zu finden. Die meisten Teilnehmenden erwähnten, dass ihnen die Angabe der Wetterverhältnisse geholfen habe. Auch der Entscheidungsbaum wurde vermehrt erwähnt. Im Vergleich dazu wurden weitere Metriken zur Messung der "Güte" des Modells sowie weitere Angaben bezogen auf die Zeit weniger häufig angesprochen. Die Häufigkeit der Pattern Codes zur Frage, aus welchem Grund sich die Teilnehmenden für ihre Antwort bei der Performance-Abfrage entschieden haben, sind im Anhang in Tabelle A.4 dargestellt. Dort sind alle Häufigkeiten der Codes je Beispiel im Prototyp ausgeführt. Die Pattern Codes zur Frage, welche Elemente die Teilnehmenden vermisst oder welche sie nicht verstanden haben, sind im Anhang in Tabelle A.5 dargestellt. Zu beachten ist, dass mehrere Teilnehmende nach der Beantwortung der Fragen zum Beispiel DT angegeben haben, dass ihnen ein Entscheidungsbaum in den darauffolgenden Darstellungen gefehlt habe. Die Stakeholdergruppen Maschinenbau sowie Service-Technik wünschten sich am häufigsten die Angabe der Länge der Hoch-Intervalle im Diagramm. Weiterhin wurde ausschließlich von Mitarbeitenden aus der Stakeholdergruppe Maschinenbau erwähnt, dass ihnen die Angabe eines längeren Zeitraums fehlte. Die Angabe der Wetterverhältnisse und der Entscheidungsbaum wurden am häufigsten von den Stakeholdergruppen Maschinenbau sowie Service-Technik angeführt. Die Stakeholdergruppen Service-Büro und Vertrieb erwähnten am häufigsten, dass sie eins oder mehrere Elemente der Darstellung nicht verstanden hätten. Weiterhin erwähnte die Stakeholdergruppe Service-Büro am häufigsten, bei ihrer Antwort zur Performance-Abfrage geraten zu haben.

Kapitel 6

Konzept für Erklärbarkeit

In diesem Kapitel wird das erstellte Konzept für Erklärungen an den Prognosen erläutert. Weiterhin werden die aufgestellten Anforderungen erklärt.

6.1 Erhebung von Anforderungen

Zur Erhebung der konkreten Anforderungen der Stakeholdergruppen wurden die Umfrageergebnisse ausgewertet. Die Erhebung der Anforderungen erfolgte nach Schneider [22] in drei Schritten: die Identifikation der Anforderungen, die Strukturierung der identifizierten Anforderungen und die Konkretisierung der Anforderungen. Zur Identifikation der Anforderungen wurden sowohl die Korrektheit bei der Performance-Abfrage als auch die Abfrage mittels Likert-Skala genutzt. Weiterhin wurden die Pattern Codes aus den Fragen nach dem Grund der Entscheidungen bei der Performance-Abfrage sowie aus den explorativen Fragen nach weiteren Verbesserungen einbezogen. Dabei ergaben sich insgesamt neun identifizierte Anforderungen. Im Rahmen der Strukturierung wurden diese Anforderungen in vier Klassen eingeteilt. Anschließend wurden die identifizierten Anforderungen im Zuge der Konkretisierung zu neun Anforderungen zusammengefasst und konkretisiert. Diese konkretisierten Anforderungen sind in Tabelle 6.1 dargestellt.

Die Priorisierung der Anforderungen erfolgt aufgrund von drei Fragen:

1. Wie korrekt war das Ergebnis der Performance-Abfrage?
2. Wie hilfreich wurde eine Darstellung empfunden?
3. Wie häufig kam etwas in der Kodierung vor?

Die Performance-Abfrage in Verbindung mit den Beispielen des Prototyps, welche in Kapitel 4.2.2 erklärt werden, hat die höchste Priorität bei der Sortierung der Anforderungen. Die Anforderungen $R01$ bis $R04$

Tabelle 6.1: Priorisierte Anforderungen an die Erklärungen der Prognosen

Code	Anforderung
*R*01	**Wetterverhältnisse über Zeit**: Die Wetterverhältnisse über Zeit sollen einsehbar sein und mit der Historie sowie Prognose vergleichbar.
*R*02	**Entscheidungsbaum ersichtlich**: Ein Entscheidungsbaum, welcher die Entscheidungen des Modells annähert, soll angezeigt werden.
*R*03	**Zusammenhang von Einflussfaktoren**: Der Zusammenhang von Einflussfaktoren und der Prognose soll explizit dargestellt werden, es sollen alle Parameter des Modells genannt werden.
*R*04	**Übersichtlichkeit des Diagramms**: Die Übersichtlichkeit des Diagramms soll gesichert werden, innerhalb des Diagramms soll kein langer Text angezeigt werden. Weiterhin sollen längere Texte nur auf Abruf dargestellt werden.
*R*05	**Länge der Hochintervalle**: Die Länge der Hoch-Intervalle in der Historie sowie Prognose sollen explizit angegeben werden.
*R*06	**Längerer Historien-Zeitraum**: Es soll ein Zeitraum von einer Woche in der Historie dargestellt werden.
*R*07	**Allgemeine Informationen zum BHKW**: Allgemeine Informationen des BHKWs sollen ersichtlich sein: Objekttyp, Adresse, Modell, Wartungsintervall.
*R*08	**Angaben zu Zeiteigenschaften**: Es sollen Angaben zu Zeiteigenschaften explizit ersichtlich sein: Wochentage, eine verkürzte Datumsanzeige sowie eine Markierung von Feiertagen soll gegeben sein
*R*09	**Metrik zur Messung der "Güte" des Modells**: Eine Metrik zur Messung der "Goodness" soll angeben, wie das Modell in der Vergangenheit im Vergleich zu tatsächlichen Werten abgeschnitten hat.

haben die Performance der Teilnehmenden im Vergleich mit dem Standard-Diagramm verbessert und erhalten somit eine höhere Priorisierung in ihrer Implementierung. Innerhalb dieser Anforderungen zählen die Ergebnisse der Frage, für wie hilfreich Teilnehmende die Beispiele jeweils empfunden haben. Die weiteren Anforderungen $R05$ bis $R09$ werden nach der Häufigkeit der Pattern Codes, welche mit den jeweiligen Anforderungen in Verbindung stehen, priorisiert.

6.2 Paper-Prototype

Für die Konzeptentwicklung wurde ein Paper-Prototype der angestrebten Software erstellt. Ein Graph, welcher die Elemente des Paper-Prototype in Verbindung miteinander konzeptuell darstellt, ist in Abbildung 6.1 dargestellt. Der Paper-Prototype kann in Abbildung 6.2 eingesehen werden. Dieser stellt das fertige Konzept unter Berücksichtigung der in Kapitel 6.1 erläuterten Anforderungen dar. Der Paper-Prototype könnte einen ersten Ansatzpunkt für die tatsächliche Umsetzung der Software bilden.

Insgesamt besteht der Prototyp aus fünf Abschnitten. Die Anzeige der Prognose und Historie, bilden den zentralen Teil der Prognose-Ansicht. Dort wird zu jedem Hoch-Intervall die zeitliche Länge dargestellt ($R05$). Darunter befindet sich zum direkten Vergleich eine Darstellung der Wetterverhältnisse über Zeit ($R01$). Eine gemeinsame Zeitachse verbindet diese beiden Diagramme, um das Vergleichen der Wetterverhältnisse mit der Historie und Prognose zu vereinfachen ($R04$). Weiterhin wird durch die zusammengefasste Achse die Übersichtlichkeit gesteigert, da keine redundanten Informationen dargestellt werden ($R04$). Die Achse beinhaltet eine Datumsangabe für jeden Tag sowie eine Angabe zum jeweiligen Wochentag ($R08$). Der “heutige“ Tag sowie Sonn- und Feiertage werden farblich markiert ($R08$). Seitlich befindet sich ein ausklappbarer Reiter für allgemeine Informationen zum BHKW ($R04$, $R07$). Ist er ausgeklappt, wird dort angegeben, um welchen Objekttyp es sich handelt ($R07$). Außerdem werden alle Einflussfaktoren der Prognose sowie ihr Feature-Score dort dargestellt ($R03$). Ein Info-Button führt zu einem Dialog, in welchem die einzelnen Faktoren sowie der Feature-Score erläutert werden ($R04$, $R09$). Seitlich befindet sich weiterhin eine Anzeige eines Entscheidungsbaumes ($R02$). Dieser nähert die Entscheidungen des Modells, wie im Prototyp in Kapitel 4.2.2 erläutert, an. Ein Info-Button führt zu einem Dialog, in welchem erklärt wird, wie der Entscheidungsbaum zu lesen ist ($R04$). Durch diesen Button wird sichergestellt, dass die gesamte Darstellung zwar übersichtlich bleibt, aber Nutzende, welche den Baum nicht direkt verstehen, genügend Informationen bekommen können. Der fünfte Abschnitt ist ein Button “Informationen zur Prognose“. Dieser führt zu einem Dialog, in welchem die Funktionsweise des Modells erklärt wird, und weitere Informationen verlinkt werden können ($R04$). Weiterhin wird

dort eine Metrik zur Messung der Performance des Modells angezeigt und erklärt ($R09$). Die Adresse, der nächste Wartungsintervall sowie das Modell des Blockheizkraftwerks werden in vom System übergeordneten Reitern bei Auswahl eines spezifischen Blockheizkraftwerks ohnehin angezeigt, müssen also in diesem Konzept nicht implementiert werden.

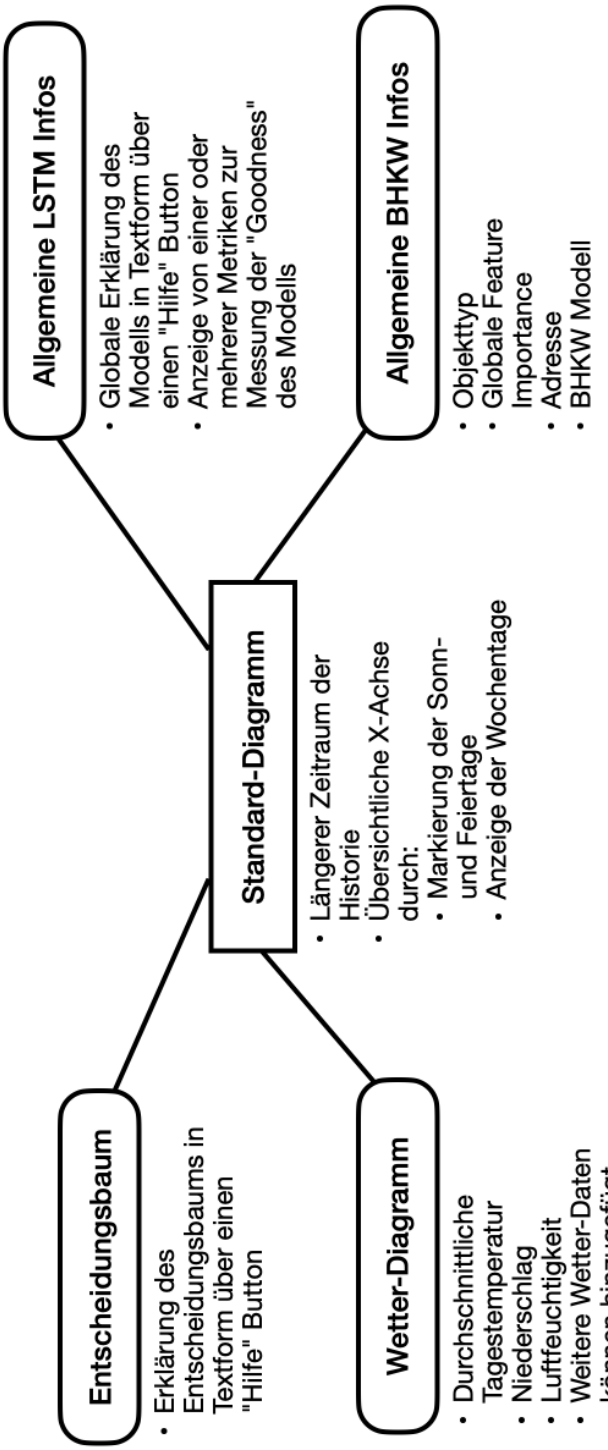

Abbildung 6.1: Konzeptueller Graph für Elemente des Paper-Prototypes

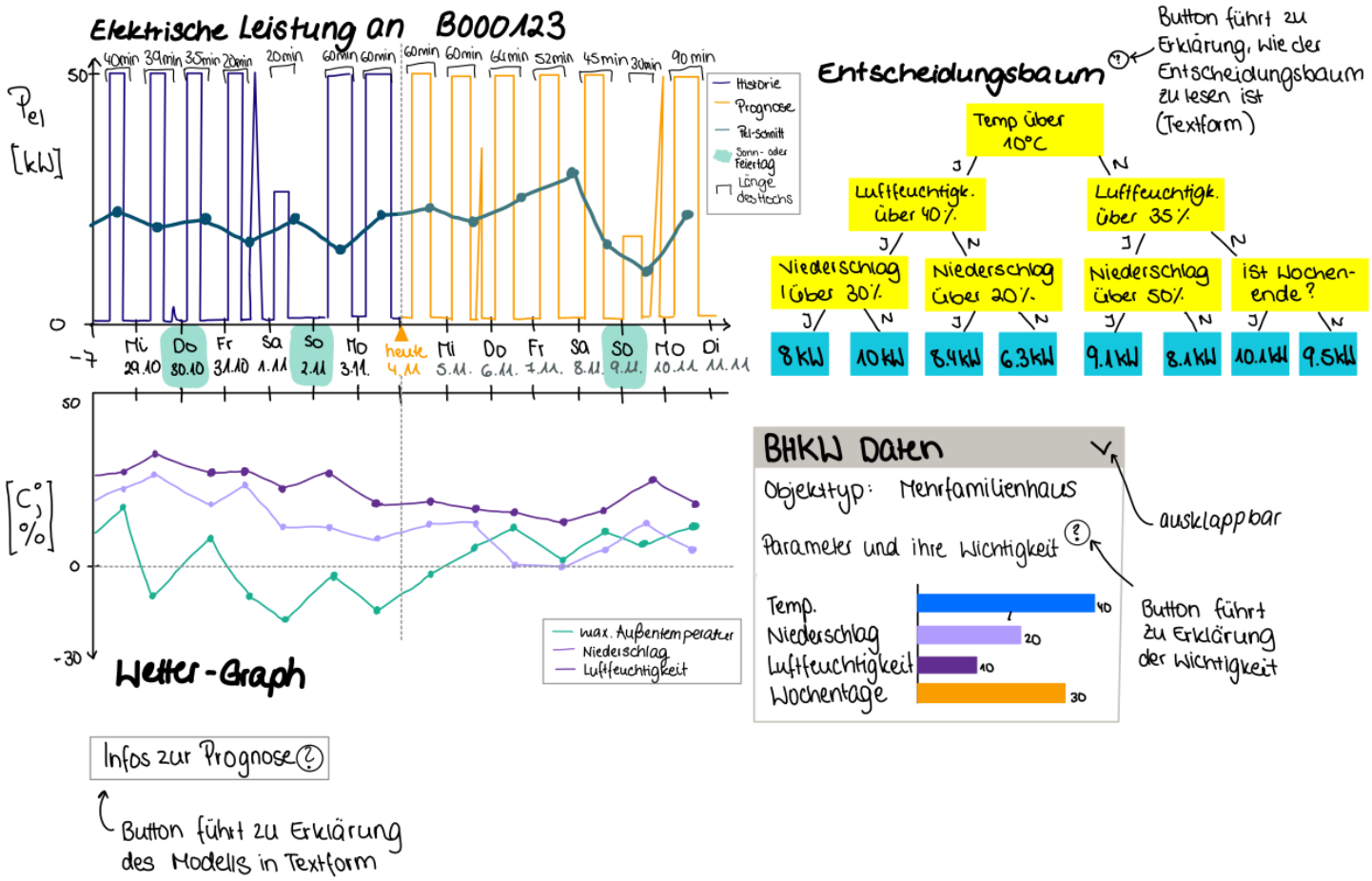

Abbildung 6.2: Paper-Prototype des erstellten Konzeptes für Erklärungen an Vorhersagen für erzeugte elektrische Leistung an Blockheizkraftwerken

Kapitel 7

Diskussion

7.1 Beantwortung der Forschungsfragen

In diesem Abschnitt werden die Forschungsfragen beantwortet, welche in Kapitel 4.1 erläutert werden.

7.1.1 Beantwortung von RQ1

Die Forschungsfrage *RQ*1 beschäftigt sich mit dem Unterschied im Bedarf nach Erklärungen an den Prognosen der Stakeholdergruppen.

> **RQ1:** Wie unterscheidet sich der Bedarf nach Erklärungen innerhalb der Stakeholdergruppen?

Auffällig ist, dass die Stakeholdergruppen Maschinenbau sowie Service-Technik die höchste Korrektheit der Performance-Abfrage erzielten. Durch ihren vermehrten Bezug auf die Pattern Codes “Int“, “Wea“ und “DT“ (siehe Tabellen A.2 und A.3) ist davon auszugehen, dass den Gruppen mit der höchsten Korrektheit die Informationen über die Länge der Hoch-Intervalle, die Angaben zu den Wetterverhältnissen sowie der Entscheidungsbaum am meisten geholfen haben.

Durch ihre geringere Korrektheit bei den Performance-Abfragen lässt sich schließen, dass die Stakeholdergruppen Vertrieb, Projektbetreuung, Geschäftsleitung und Service-Büro den größten Bedarf an Erklärungen der Prognosen haben. Sie beantworteten weniger als 55 % aller Fragen richtig. Durch die erhöhte Erwähnung und Wunsch nach weiteren, allgemeinen Angaben zum BHKW in den Gruppen Projektbetreuung und Vertrieb ist der Bedarf nach allgemeinen Informationen in diesen Gruppen höher als in den anderen Stakeholdergruppen. Alle Teams sind sich einig, dass ihnen das Beispiel GlobalModel am wenigsten geholfen hat. Einen Bedarf nach allgemeinen Informationen zum LSTM-Modell gibt es daher in keiner Stakeholdergruppe.

Bis auf die Stakeholdergruppen Steuerungstechnik und Projektbetreuung erzielten alle Gruppen eine vollständige Korrektheit am Beispiel DT, weshalb der Bedarf nach einer Erklärung in Form eines Entscheidungsbaums besonders hoch ist. Weiterhin erwähnte die Gruppe Service-Büro am häufigsten, bei ihrer Antwort geraten zu haben, gefolgt von der Gruppe Geschäftsführung. Auch diese Gruppen beantworteten die Performance-Abfrage am Beispiel DT zu 100 % korrekt.

7.1.2 Beantwortung von RQ2

Die Antwort auf die Forschungsfrage *RQ*2 kann direkt mit dem Paper-Prototype beantwortet werden, welcher in Kapitel 6.2 beschrieben ist.

> **RQ2:** Wie kann das bestehende System durch Erklärungen an den Voraussagen verbessert werden?

Die konkreten Verbesserungen durch Erklärungen am System lassen sich durch die erhöhte Korrektheit der Antworten in der Performance-Abfrage zum Beispiel DT und der hohen Erwähnung des Pattern Codes "Wetterverhältnisse über Zeit" entnehmen. Besonders die Angabe eines Entscheidungsbaums zur Annäherung der Entscheidungen des Modells sowie die Angabe der historischen sowie prognostizierten Wetterverhältnisse steigerten die Performance der Teilnehmenden. Wie in Kapitel 2.3 erwähnt, misst eine solche Performance-Abfrage das mentale Modell der Nutzenden. Die hohe Korrektheit am Beispiel DT lässt schließen, dass der Entscheidungsbaum das mentale Modell der meisten Nutzenden näher zum tatsächlichen Modell gebracht hat.

Auch die Angabe der Feature Importance erzielte eine erhöhte Korrektheit der Antworten, daher kann diese Angabe das mentale Modell der Teilnehmenden dabei unterstützen, zu verstehen, welche Faktoren Einfluss auf die Prognosen nehmen. Um die Übersichtlichkeit der Erklärungen und der Prognose zu erhalten, werden keine langen Erklärungen in Textform direkt dargestellt, sondern erst nach explizitem Anfragen durch Nutzende in Dialogen gezeigt. Der Paper-Prototype, gezeigt in Abbildung 6.2 ist ein Beispiel für eine konkrete Verbesserung des bestehenden Systems durch Erklärungen an den Voraussagen.

7.2 Umsetzbarkeit

In diesem Abschnitt wird die Umsetzbarkeit des in Kapitel 6 beschriebenen Konzepts diskutiert. Der in Kapitel 6.2 beschriebene Entscheidungsbaum, welcher die Entscheidungen des Modells annähern soll und somit interpretierbarer für Nutzende macht, ist an ein potenzielles Ergebnis des TREPAN-Algorithmus nach Craven [5] angelehnt. Wie in Kapitel 2.4

beschrieben, kann der TREPAN-Algorithmus auch die Entscheidungen von Zeitreihen-Modellen annähern. Craven und Shavlik [6] erklären, dass TREPAN keine spezielle Netzarchitektur oder Trainingsmethode benötigt, da der Algorithmus jedes Modell als Black-Box Modell betrachtet. Die Autoren erläutern weiterhin, dass der TREPAN-Algorithmus aus einem trainierten neuronalen Netz ein prägnantes, symbolisches Konzept des Netzes extrahieren könne. Jedoch könne TREPAN keine tatsächlichen Werte in den Blattknoten des Entscheidungsbaumes darstellen. Daher müsse der Output des Modells entsprechend angepasst werden. Die Autoren erläutern, dass der Algorithmus etwa Aussagen darüber treffen könne, wann ein bestimmter Wert steigt oder sinkt. Im Falle des Modells zur Vorhersage von erzeugter elektrischer Leistung eines BHKWs könnte etwa eine Prognose zum Wert der durchschnittlich erzeugten elektrischen Leistung am Tag erstellt, und durch den TREPAN-Algorithmus erklärt werden. Der fertige Baum würde dann erklären, wieso die durchschnittlich erzeugte elektrische Leistung steigt oder sinkt.

Nach Dieber und Kirrane [8] macht auch der LIME-Algorithmus Klassifizierungen erklärbar. Um den Algorithmus für das Entnehmen von Feature Importance bei Vorhersagen erzeugter elektrischer Leistung an BHKWs einzusetzen, müsste das Modell, ähnlich wie beim Einsatz von TREPAN, eine Klassifizierung erbringen. Dafür würde sich in etwa eine Einstufung der durchschnittlich erzeugten elektrischen Leistung pro Tag eignen, um die Klassen “Low“, “Medium“ und “High“ für eine geringe, mittlere und hohe Produktion elektrischer Leistung pro Tag vorherzusehen. Der Einsatz des LIME-Algorithmus würde es damit möglich machen, die Feature Importance für einzelne Tage zu extrahieren.

Ozyegen et al. [15] erklären, dass SHAP Methoden, wie in Kapitel 2.4 erläutert, die Prognosen von Zeitreihen-Modellen interpretierbar machen können. Die Autoren beschreiben, dass die Performance von SHAP Methoden in Verbindung mit einem LSTM-Modell abhängig von der Performance des Modells selbst ist. Je niedriger die Genauigkeit des LSTM-Modells, desto mehr Fehler würden die Erklärungen der SHAP Methoden aufweisen. SHAP Methoden könnten zur Messung von Feature Importance bei Vorhersagen erzeugter elektrischer Leistung an BHKWs eingesetzt werden.

Zusammenfassend kann gesagt werden, dass die Umsetzung der oben genannten Methoden für die in Kapitel 6.2 vorgeschlagenen Erklärungen unter Berücksichtigung der besprochenen Einschränkungen technisch möglich ist.

7.3 Limitierungen der Validität

In diesem Abschnitt werden die Limitierungen der Validität dieser Arbeit erläutert. Diese Diskussion richtet sich nach Wohlin et al. [24].

Die *construct validity* wird durch die quantitative Natur der in der

Befragung erhobenen Daten limitiert. Um ein tieferes Verständnis für den Bedarf der Stakeholder zu erhalten, ist es sinnvoll, weitergehende qualitative Daten zu erheben. Aus diesem Grund wurden Kommentarfelder in die Befragung aufgenommen und die Antworten kodiert.

Die *internal validity* wird durch Einschränkungen im Kodierungsprozess limitiert. Da diese Arbeit eine Abschlussarbeit ist, wurden Freitextantworten nur durch eine Person kodiert. Um die Ergebnisse weiter zu festigen, kann der Prozess zur Kodierung der Antworten erweitert werden. Zur Erweiterung des Kodierungsprozesses könnten etwa weitere Personen zur Kodierung herangezogen werden und ein Interrater Agreement erstellt werden.

Die *conclusion validity* wird durch die Demographie der befragten Personen limitiert. Bei den Befragten handelt es sich ausschließlich um Mitarbeitende eines Unternehmens in Deutschland, weshalb es womöglich wenig Variation in den kulturellen Werten der Teilnehmenden gab. Weiterhin identifizierten sich die Teilnehmenden vorwiegend als männlich. Mit mehr Teilnehmenden oder einer anderen Demographie hätte dieselbe Befragung gegebenenfalls zu anderen Ergebnissen geführt. Ebenso beschränkt der Stichprobenumfang von 23 vollständigen Antworten die *conclusion validity*.

Die *external validity* wird durch die ausschließliche Ausführung der Befragung innerhalb des Unternehmens limitiert. Es ist unklar, inwieweit sich die Ergebnisse auf andere Kontexte generalisieren lassen. Jedoch ist es vorstellbar, dass die erhobenen Ergebnisse auch für andere Zeitreihendaten in ähnlichen Anwendungsfeldern anwendbar sind.

Kapitel 8

Zusammenfassung und Ausblick

In diesem Kapitel werden die Ziele dieser Arbeit mit den Ergebnissen in einen Kontext gesetzt. Zuletzt werden mögliche Wege für die zukünftige Forschung vorgeschlagen.

8.1 Zusammenfassung

Das Ziel dieser Arbeit war, das Vertrauen von Mitarbeitenden der Firma Kraftwerk in Voraussagen der erzeugten elektrischen Leistung von BHKWs mittels Machine-Learning-Methoden zu steigern. In diesem Zuge werden die Stakeholder der Voraussagen ermittelt und in Stakeholdergruppen aufgeteilt. Es ergeben sich acht Stakeholdergruppen, welche verschiedene Motivationen an den Voraussagen haben. Durch die Ähnlichkeit der Motivation von Mitarbeitenden in den selben Teams entspricht die Aufteilung der Stakeholdergruppen der Aufteilung von Mitarbeitenden in Teams. Die Motivationen dieser Stakeholdergruppen reichen von der Forschung und Entwicklung über die Betriebsoptimierung bis hin zur Übernahme unternehmerischer Verantwortung.

Die verwendete Machine-Learning-Methode für die Prognosen ist ein LSTM-Modell, welches die nichtlineare Natur des Verlaufs erzeugter elektrischer Leistung darstellen kann. Die Prognosen werden von den Stakeholdern dazu genutzt, Veränderungen an BHKWs nach einer getätigten Einstellung mit dem prognostizierten Verlauf vor der Einstellung zu vergleichen. Durch die gesteigerte Komplexität der Machine-Learning-Methode sollen Erklärungen das Vertrauen der Stakeholder in die Voraussagen steigern und sie dabei unterstützen, die Voraussagen zu interpretieren. Dafür wird im Zuge dieser Arbeit ein Konzept für Erklärungen entwickelt, welches die Voraussagen ergänzt und es den Stakeholdern erleichtert, mit den Prognosen zu arbeiten.

Um die Meinung der Stakeholdergruppen bezüglich potenzieller Erklä-

rungen zu erarbeiten, wird eine Umfrage mit den tatsächlichen Stakeholdern durchgeführt. Diese Umfrage wird durch einen selbst entwickelten High-Fidelity-Prototype unterstützt, welcher verschiedene Beispiele von Erklärungen an den Prognosen darstellt. Die Beispiele umfassen unter anderem die Darstellung allgemeiner Informationen zum Modell, die Feature Importance der Einflussfaktoren wie Wetter und Jahreszeit sowie die Darstellung eines Entscheidungsbaums, welcher die Entscheidungen des Modells annähert und durch Schwellenwerte darstellt. Die Stakeholder beantworteten eine Performance-Abfrage für jedes Beispiel des Prototyps, um die Eignung der Beispiele zu ermitteln. Dabei zeichnete sich eine klare Bevorzugung der Erklärung mittels Entscheidungsbaum sowie der Darstellung der Wetterverhältnisse über Zeit ab. Die Auswertung der Umfrage erfolgte in zwei Schritten: der statistischen Auswertung und der Kodierung der Freitext-Antworten. Auch die Ergebnisse der Kodierung zeigten einen starken Trend zugunsten der Erklärungen mittels Entscheidungsbaum und Darstellung der Wetterverhältnisse über Zeit.

Auf Grundlage dieser Ergebnisse wurden insgesamt neun Anforderungen an die Erklärungen zusammengefasst. Diese wurden zunächst identifiziert, dann klassifiziert und schließlich konkretisiert. Die identifizierten Anforderungen wurden in einem Paper-Prototype implementiert. Anschließend wurde die Umsetzbarkeit des Konzepts diskutiert. Durch die Anlehnung der Konzeptteile an bestehende XAI-Methoden erscheint die Umsetzung des Konzepts innerhalb des Systems von Kraftwerk mit wenigen Modifikationen realistisch.

8.2 Ausblick

Während im Zuge dieser Arbeit ein Konzept für Erklärungen an den Prognosen erzeugter elektrischer Leistung von BHKWs entwickelt wurde, muss dieses Konzept noch praktisch umgesetzt werden. Die Implementierung des Konzepts in die interne Verwaltungs-Software der BHKWs umfasst nicht nur die Bearbeitung der momentanen Darstellung, sondern auch die Beschaffung und Darstellung der zusätzlichen Daten entsprechend des Konzepts. Nicht nur historische Wetterdaten sowie deren Vorhersage müssen abgebildet, sondern auch weitere weniger komplexe Machine-Learning-Modelle trainiert werden, um die Entscheidungen des LSTM-Modells anzunähern und Entscheidungsbäume und Feature Importance darzustellen. Nach der Entwicklung dieser Software muss diese in Bezug auf Funktionalität und Nutzbarkeit getestet werden. Dies könnte durch eine Nutzerstudie inklusive einer Performance-Abfrage erfolgen. Im nächsten Schritt können die Ergebnisse dann mit den Ergebnissen der Beispiele im Prototyp verglichen werden. Daraufhin können die Anforderungen dementsprechend angepasst werden. Da diese Arbeit ausschließlich im Rahmen der Firma Kraftwerk

erfolgte, wäre im Anschluss die Erforschung der Übertragbarkeit dieser Arbeit auf andere Umfelder denkbar. Zuletzt ist die Anwendbarkeit dieser Arbeit auf andere Zeitreihendaten vorstellbar, müsse aber weiter untersucht werden.

Abbildungsverzeichnis

Tabellenverzeichnis

Glossar

BHKW Blockheizkraftwerk. 1–10, 15–18, 38, 39, 43, 45, 47, 48, 51

DT Decision Tree. 27, 32–36, 43, 44, 58

Feature Importance Wichtigkeit der Features. 13, 25, 26, 44, 45, 48

GlobalFeat Globale Erklärung der Feature Importance. 26, 32, 33, 36

GlobalModel Globale Erklärung des Modells in Textform. 27, 32, 33, 35, 36, 43

Kraftwerk Kraftwerk Kraft-Wärme-Kopplung GmbH. 1, 2, 4, 5, 9–12, 15, 16, 18, 21, 22, 24, 25, 28, 29, 31, 47, 48

LIME Local Interpretable Model-agnostic Explanations. 12, 13, 45

LocalFeat Lokale Erklärung der Feature Importance. 26, 32, 33, 36

LocalText Lokale Erklärung in Textform. 27, 32, 33, 35, 36

LSTM Long Short-Term Memory. 2, 4, 7, 9, 27, 28, 43, 45, 47, 48, 51

SD Standard-Diagramm. 25, 32

XAI Explainable Artificial Intelligence. 10, 48

Anhang A

Anhang

In diesem Anhang werden weitere Ergebnisse der Befragung tabellarisch dargestellt.

Tabelle A.1: Zusammenfassung der Antworten auf die Frage nach der Geschlechtsidentität der Teilnehmenden, zu Kapitel 5.1

Antwort	**Prozentsatz**
Männlich	81,61 %
Weiblich	4,35 %
Divers	0 %
Sonstige	4,35 %

Tabelle A.2: Pattern Codes und ihre Häufigkeit, zu Kapitel 5.5

Pattern Code	Frequenz
Wea	63
NoFeat	27
Gue	20
Int	19
DT	18
No	15
His	10

Pattern Code	Frequenz
Fore	9
Wrong	7
Vis	5
Feat	4
Met	1
Time	1

Tabelle A.3: Legende von Abkürzungen der Pattern Codes

Wea	Wetterverhältnisse über Zeit
Int	Länge der Hoch-Intervalle
Gue	Ich bin mir unsicher oder habe geraten.
No	Ich habe in oder mehrere Elemente nicht verstanden.
His	Wunsch nach Darstellung eines längeren Zeitraums der Historie
Fore	Bezug auf die Prognose der elektrischen Leistung genommen
Met	Metrik zur Messung, wie "richtig" das Modell in der Vergangenheit lag, Vergleich der Historie mit vergangenen Prognosen
NoFeat	Der Zusammenhang zwischen Einflussfaktoren und Prognose ist nicht ersichtlich, oder es fehlen Einflussfaktoren.
DT	Der Entscheidungsbaum hat gefehlt oder hat mir geholfen.
Wrong	Fälschlicherweise Aussagen anderer Schritte einbezogen.
Time	Angaben über Eigenschaften auf die Zeit bezogen (Wochentage, Jahreszeiten, Tag-Nacht)
Vis	Andere Art und Weise der Visualisierung ist gewünscht.

Tabelle A.4: Pattern Codes und ihre Häufigkeit zum Grund der Entscheidung zur Performance-Abfrage, zu Kapitel 5.5

Code	**DT**	**Local Feat**	**Global Feat**	**Local Text**	**Global Model**	**SD**
Int		3	3		4	6
Fore		1		2	1	3
Feat					1	
Wea	9	14	16	12		1
Gue	1	2	3	2	5	3
No		1		1	2	
His	2			1		
Met						
NoFeat	1	1	1	2	4	
DT	14					
Wrong		2	1	3	1	
Time						
Vis						

Tabelle A.5: Pattern Codes und ihre Häufigkeit zu nicht verstandenen oder von Teilnehmenden gewünschten Elementen, zu Kapitel 5.5

Code	**DT**	**Local Feat**	**Global Feat**	**Local Text**	**Global Model**	**SD**
Int			1	1		1
Fore			1			
Feat			1			3
Wea			2		6	3
Gue				1		3
No		4	2	2	2	1
His	1	1	2	1		2
Met		1				
NoFeat	2	3	3	1	7	2
DT	1			2	1	
Wrong						
Time			1			
Vis	1	1		2	1	

Literaturverzeichnis

[1] S. S. Amiri, S. Mottahedi, E. R. Lee, and S. Hoque. Peeking inside the black-box: Explainable machine learning applied to household transportation energy consumption. *Computers, Environment and Urban Systems*, 88:101647, 2021.

[2] L. Chazette. *Requirements engineering for explainable systems*. Dissertation, Gottfried Wilhelm Leibniz Universität Hannover, 2023.

[3] L. Chazette, J. Klünder, M. Balci, and K. Schneider. How can we develop explainable systems? insights from a literature review and an interview study. In *Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering*, pages 1–12, 2022.

[4] L. Chazette and K. Schneider. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, 25(4):493–514, 2020.

[5] M. W. Craven. *Extracting comprehensible models from trained neural networks*. Dissertation, The University of Wisconsin-Madison, 1996.

[6] M. W. Craven and J. W. Shavlik. Understanding time-series networks: A case study in rule extraction. *International Journal of Neural Systems*, 8(04):373–384, 1997.

[7] T. De, P. Giri, A. Mevawala, R. Nemani, and A. Deo. Explainable ai: a hybrid approach to generate human-interpretable explanation for deep learning prediction. *Procedia Computer Science*, 168:40–48, 2020.

[8] J. Dieber and S. Kirrane. Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*, 2020.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] C. I. Kirchhoff. Designing an experiment on triggering explanations based on mental model conflicts. Bachelor's thesis, Gottfried Wilhelm Leibniz Universität Hannover, 2023.

[11] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[12] S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

[13] C. Molnar. *Interpretable machine learning*. Lean Publishing, 2020.

[14] C. Molnar, G. Casalicchio, and B. Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 417–431. Springer, 2020.

[15] O. Ozyegen, I. Ilic, and M. Cevik. Evaluation of interpretability methods for multivariate time series forecasting. *Applied Intelligence*, pages 1–17, 2022.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[17] T. Rojat, R. Puget, D. Filliat, J. D. Ser, R. Gelin, and N. D. Rodríguez. Explainable artificial intelligence (XAI) on timeseries data: A survey. *CoRR*, abs/2104.00950, 2021.

[18] S. Russell and P. Norvig. *Künstliche Intelligenz, 3. Auflage*. Pearson Studium, 2012.

[19] J. Saldaña. *The Coding Manual for Qualitative Researchers*. SAGE Publications Inc., Thousand Oaks, CA, USA, 2nd edition, 2013.

[20] G. Schaumann, K. W. Schmitz, et al. *Kraft-Wärme-Kopplung*, volume 8. Springer, 2010.

[21] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim. Towards a rigorous evaluation of xai methods on time series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4197–4201, 2019.

[22] K. Schneider. Vorlesungsfolien: Grundlagen der Softwaretechnik, 2019. Gottfried Wilhelm Leibniz Universität Hannover, Fachgebiet Software Engineering.

[23] L. S. Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[24] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.