

Entwicklung und Einsatz automatisierter maschineller Lernmodelle für die Vorhersage von Polymer-Bandlücken

Thesis

at

University of Bayreuth

Chair of
Computational Material Science

von Ibrahim Karademir

Supervised by Prof. Dr. Christopher Künneth
Cosupervised by Prof. Dr. Daniel Buschek

Contents

Zusammenfassung	1
Einleitung	2
1 Hintergrund	3
1.1 Verwandte Arbeiten	3
1.2 Grundlagen der Polymere und Polymer Informatik	5
1.2.1 Klassifikation von Polymeren	5
1.2.2 Struktur von Polymeren	6
1.2.3 Elektrische Bandlücke in Polymeren	8
1.2.4 Einführung in die Chemoinformatik	8
1.2.5 SMILES	9
1.2.6 Polymere SMILES	9
1.2.7 Polymer fingerprints	10
1.3 Maschinelles Lernen und Auto-ML	11
1.3.1 Regression	12
1.3.2 Neuronale Netze und maschinelles Lernen	14
1.3.3 Model Ensembles im maschinellen Lernen	16
1.4 HCI und Nutzerstudien	17
1.4.1 Arten von Interviews	18
1.4.2 Beobachtungen	19
1.4.3 Fragebögen	20
2 AutoML-Frameworks	22
2.1 Auswahl der AutoML-Frameworks	22
2.2 Vorstellung AutoML-Bibliotheken	23
2.2.1 AutoGluon	23
2.2.2 AutoKeras	24
2.2.3 Auto-sklearn	25
2.2.4 PyCaret	25

2.2.5	Ludwig	26
2.3	Datenvorbereitung	28
2.4	Modelltraining und Evaluation	28
3	Entwicklung des Webtools	30
3.1	Ziel des Webtools	30
3.2	Aufbau und Entwicklung	30
3.2.1	My Project Tab	31
3.2.2	Polymere Tab	32
3.2.3	Automated Machine Learning Tab	33
3.2.4	Train Your Model Tab	33
3.2.5	About Me Tab	35
4	Durchführung der Nutzerstudie	37
4.1	User Tasks	37
4.2	Semi-Strukturierte Online-Befragung	41
5	Ergebnisse	43
5.1	Leistung der AutoML-Modelle	43
5.2	Ergebnisse der Nutzerstudie	45
5.3	Evaluation der Semi-Strukturierten Online-Befragung	46
6	Diskussionen	50
7	Schlussfolgerungen und Ausblick	52
	References	55

Zusammenfassung

Diese Arbeit untersucht die Leistungsfähigkeit von Automated Machine Learning (AutoML) bei der Vorhersage von Bandlücken in Polymeren, einem Schlüsselement für die Entwicklung neuer Materialien mit spezifischen elektronischen Eigenschaften. Angesichts der komplexen Natur der Materialwissenschaft und der steigenden Anforderungen an präzise Vorhersagemodelle zielt diese Studie darauf ab, die Effektivität von AutoML-Lösungen zu bewerten, um zu bestimmen, welche AutoML-Ansätze am besten für die Vorhersage von Bandlücken in Polymeren geeignet sind. Dazu wurden verschiedene AutoML Frameworks systematisch getestet, um ihre Genauigkeit und Anwendbarkeit zu vergleichen. Die Ergebnisse zeigen, dass bestimmte AutoML Modelle eine hohe Vorhersagegenauigkeit erzielen, was ihre Eignung für diese spezifische Anwendung unterstreicht. Basierend auf den gewonnenen Erkenntnissen wurde ein Webtool entwickelt, das es Nutzern ermöglicht, eigene Modelle basierend auf ihren Daten mit der leistungsfähigsten AutoML-Bibliothek zu trainieren. Dieses Tool bietet nicht nur Zugang zu den Forschungsergebnissen, sondern auch eine Plattform für weitere experimentelle Studien.

Einleitung

Die rasanten Fortschritte im Bereich des maschinellen Lernens (ML) haben transformative Auswirkungen auf zahlreiche Wissenschafts- und Industriebereiche, insbesondere in der Materialwissenschaft, wo sie das Potenzial haben, die Entdeckung und Entwicklung neuer Materialien zu revolutionieren. Trotz der beeindruckenden Fortschritte bleibt die Anwendung von ML-Techniken oft auf Experten beschränkt, da sie tiefgreifendes Fachwissen in Datenwissenschaft und Programmierung erfordern. Diese Barriere verhindert, dass ein breiteres Spektrum von Forschern und Entwicklern ML-Modelle effektiv nutzen können, insbesondere in spezialisierten Anwendungen wie der Vorhersage von Bandlücken in Polymeren. Bandlücken sind ein kritischer Parameter für die Funktionalität von Materialien in elektronischen Anwendungen, und ihre präzise Vorhersage ist für die Materialentwicklung von entscheidender Bedeutung. Automated Machine Learning (AutoML) bietet eine vielversprechende Lösung für diese Herausforderung, indem es den Prozess der Auswahl, Konfiguration und Optimierung von ML-Modellen automatisiert. Die zentrale Motivation dieser Arbeit ist es daher, die Anwendbarkeit und Effizienz von AutoML für die spezifische Aufgabe der Bandlücken-Vorhersage in Polymeren zu untersuchen. Durch die Evaluation verschiedener AutoML-Frameworks wird das Ziel verfolgt, deren Eignung für die spezifische Aufgabe zu bestimmen und zu ermitteln, welches Framework die genauesten Vorhersagen liefert. Ferner wird die Notwendigkeit hervorgehoben, die Zugänglichkeit und Anwendbarkeit von maschinellem Lernen (ML) für ein breites Spektrum von Anwendern zu verbessern. Vor diesem Hintergrund wurde basierend auf den Erkenntnissen der vorliegenden Studie ein webbasiertes Tool entwickelt, um die Implementierung von ML-Verfahren zu vereinfachen. Dieses Tool ermöglicht es Nutzern, ohne tiefgreifende Kenntnisse in ML oder Programmierung, eigene Daten zu verwenden, um Modelle für die Vorhersage von Bandlücken zu trainieren. Das Tool stellt eine direkte Anwendung der Forschungsergebnisse dar und demonstriert, wie AutoML die Zugänglichkeit und Anwendbarkeit von ML-Technologien erweitern kann.

Chapter 1

Hintergrund

1.1 Verwandte Arbeiten

In der Studie von Omar et al. wurde H2O AutoML eingesetzt, um die Ausbreitung von Rissen in Polymeren automatisch vorherzusagen. Die Forschungsarbeit, veröffentlicht im Journal Sensors, zielt darauf ab, mithilfe von maschinellem Lernen (ML) und automatisierter Modellierung (AutoML) die Rissausbreitung in polymeren Materialien zu prognostizieren.

Die Autoren haben H2O AutoML verwendet, weil es verschiedene ML-Algorithmen automatisch trainiert und optimiert, um das beste Modell für die vorliegende Aufgabe zu finden. H2O AutoML führt eine Vielzahl von Aufgaben aus, darunter Datenvorbereitung, Feature-Engineering, Modellauswahl und Hyperparameter-Optimierung. In ihrer Forschung haben Omar et al. gezeigt, dass H2O AutoML effektiv in der Vorhersage von Rissausbreitungen in Polymeren sein kann, was für die Materialwissenschaft und Ingenieurpraxis von großer Bedeutung ist.

Die Verwendung von H2O AutoML ermöglichte es den Forschern, präzise Vorhersagemodelle zu entwickeln, ohne dass umfassende manuelle Modellierung erforderlich war. Dies stellt einen signifikanten Vorteil in der Effizienz und Genauigkeit der Vorhersagen dar und zeigt das Potenzial von AutoML-Systemen in der angewandten Materialwissenschaft. [44]

In einem ähnlichen Kontext, jedoch in einem anderen Anwendungsbereich, haben Hariri-Ardebili et al. [23] eine umfassende Untersuchung der Anwendung von AutoML-Lösungen für die Vorhersage der Druckfestigkeit von Beton durchgeführt. Diese Studie hebt die Bedeutung der Integra-

tion fortschrittlicher maschineller Lernmethoden in den Bereich der Betontechnologie hervor, um die Genauigkeit bei der Vorhersage der Betoneigenschaften zu verbessern. Besonders bemerkenswert ist, dass die Autoren verschiedene AutoML-Tools bewertet und ihre Leistung anhand mehrerer Datensätze zur Betonfestigkeit verglichen haben. Die Ergebnisse dieser Untersuchung unterstreichen die hohe Genauigkeit der AutoML-Modelle in der Vorhersage der Druckfestigkeit von Beton, was die Potenziale dieser Technologie in der Bauindustrie verdeutlicht.

Ähnlich wie in der von Hariri-Ardebili et al. durchgeführten Forschung, nutze auch ich in meiner Arbeit das AutoML-Tool PyCaret, um die Bandbreite von Polymeren vorherzusagen. Die Wahl von PyCaret als AutoML-Lösung für diese Aufgabe basiert auf seiner Benutzerfreundlichkeit und Effizienz in der Automatisierung des maschinellen Lernprozesses. Durch den Einsatz von PyCaret wird der Bedarf an manueller Hyperparameter-Optimierung minimiert, während gleichzeitig präzise und zuverlässige Vorhersagemodelle generiert werden können. Die Arbeit von Hariri-Ardebili et al. dient somit als wertvolle Referenz und Bestätigung für die Anwendbarkeit und Wirksamkeit von AutoML-Lösungen im Kontext der Betontechnologie.

Beide Studien verdeutlichen die breite Anwendbarkeit und Effizienz von AutoML-Technologien in der Materialwissenschaft. Dies bildet eine solide Grundlage für meine eigene Forschung, in der ich das AutoML-Tools verwende, um die Eigenschaften von Polymeren vorherzusagen. Somit ergänzen diese Untersuchungen die methodische Basis meiner Arbeit und bekräftigen die Relevanz von AutoML-Technologien in verschiedenen Bereichen der Materialwissenschaft.

1.2 Grundlagen der Polymere und Polymer Informatik

Polymere, etymologisch hergeleitet aus den altgriechischen Begriffen "polýs" für "viel" und "méros" für "Teil", repräsentieren fundamentale chemische Entitäten, die eine essentielle Funktion im Kontext der zeitgenössischen Technologie und Wissenschaft innehaben. Diese Materialien sind charakterisiert durch ihre Zusammensetzung aus langkettigen Makromolekülen, welche wiederum aus repetitiven Strukturen, den sogenannten Monomeren, bestehen [22]. Obschon die individuellen Makromoleküle eines Polymers hinsichtlich der Anzahl ihrer konstituierenden Monomere und folglich ihrer Molekülmasse variieren können, ist es gerade diese Heterogenität, die zu ihrer umfangreichen Anwendbarkeit beiträgt. Polymere, die sich von synthetischen und halbsynthetischen Varianten, welche als fundamentale Bestandteile für Kunststoffprodukte fungieren, bis zu natürlichen Biopolymeren erstrecken, die essentiell für vitale biologische Prozesse sind, manifestieren ihre Präsenz in praktisch allen Facetten des alltäglichen Lebens. Ihre Anwendung reicht von omnipräsenten Objekten wie Kunststoffverpackungen, Textilprodukten und Baustoffen bis hin zu spezialisierten Einsatzgebieten in der Biotechnologie und der medizinischen Forschung und Praxis.

1.2.1 Klassifikation von Polymeren

Polymere werden nach der Anzahl der Monomere, aus denen sie aufgebaut sind, klassifiziert:

- **Homopolymere:** Bestehen aus einer Art von Monomer. [6] [26] Ein Beispiel hierfür ist Polystyrol [siehe 1.1], das aufgrund seiner vielseitigen Anwendungsmöglichkeiten in Verpackungen und Behältern anerkannt ist, sowie Naturkautschuk, der für seine Elastizität und den Einsatz in Reifen und Dichtungen bekannt ist. [12]
- **Copolymere:** Entstehen durch die Polymerisation von zwei oder mehr verschiedenen Monomerarten. [6] Dies führt zu Materialien mit einzigartigen Eigenschaften, welche sich bei der Herstellung gezielt steuern lassen [4], wie zum Beispiel ABS, das für seine Schlagfestigkeit in der Automobil- und Elektronikindustrie geschätzt wird, und Styrol-Acrylnitril, das wegen seiner Transparenz und Chemikalienbeständigkeit in Konsumgütern verwendet wird. Ein weiteres Beispiel für ein Copolymer ist Styrol-Butadien-Kautschuk (SBR) [siehe 1.2], bekannt für seine Verwendung in der Reifenherstellung

und Industrie aufgrund seiner guten Abriebfestigkeit und Alterungsbeständigkeit.

- **Polymerblends:** Sind Mischungen aus unterschiedlichen Polymeren, die durch physikalisches Vermischen entstehen, ohne dass eine chemische Bindung zwischen ihnen erforderlich ist. [45] Wie in [65] diskutiert, können Eigenschaften verschiedener Polymere intelligent zu komplexen, mehrkomponentigen Legierungen kombiniert werden.

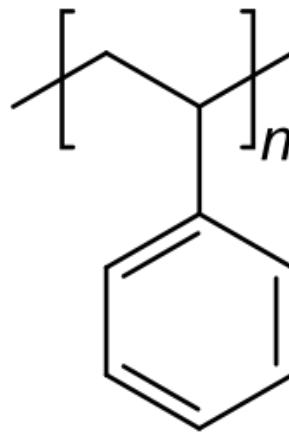


Figure 1.1: Polystyrol-Struktur.[47]

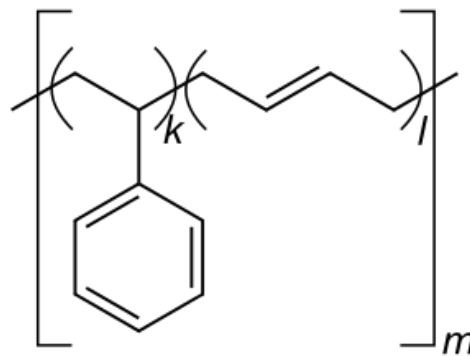


Figure 1.2: Styrol-Butadien-Kautschuk.[63].

1.2.2 Struktur von Polymeren

Die physikalischen Eigenschaften eines Polymers werden durch die Grundstruktur seiner Makromoleküle bestimmt.[60] Diese Strukturen können in

verschiedene Kategorien unterteilt werden, die jeweils spezifische Eigenschaften und Anwendungen haben.

Lineare Polymere

Lineare Polymere sind Makromoleküle, die aus wiederholten Monomereinheiten in einer unverzweigten Kette bestehen. Ihre viskoelastischen Eigenschaften variieren zwischen flüssig und elastisch, abhängig von der Deformationsrate und der Temperatur. Die Eigenschaften dieser Polymere, speziell die Übergänge zum hochelastischen Zustand, sind durch kritische Spannungsniveaus und Deformationsraten definiert, welche signifikant von der molekularen Masse abhängen. [66] Lineare Polymere mit periodischer Struktur, geringer Verzweigung und Stereoregularität (z.B. nicht ataktisch) weisen zusätzlich im festen Zustand eine halbkristalline Struktur auf.[64].

Verzweigte Makromoleküle

Verzweigte Makromoleküle, auch bekannt als dendritische Polymere, zeichnen sich durch eine strukturierte Hierarchie von Verzweigungen aus, die von einem zentralen Punkt ausgehen. Diese Polymere haben eine hohe Dichte an Funktionalitätsstellen und können komplexe Strukturen bilden, was sie in der Materialwissenschaft besonders wertvoll macht. [67]

Teilkristalline Strukturen

Teilkristalline Strukturen in Polymeren sind charakterisiert durch die Koexistenz von kristallinen und amorphen Bereichen, die ihre physikalischen Eigenschaften prägen. Diese Strukturen tragen zu einer erhöhten mechanischen Festigkeit und Steifigkeit bei und beeinflussen das Deformationsverhalten des Materials. Die Morphologie, insbesondere die kristallinen Lamellen und deren räumliche Anordnung, spielt eine entscheidende Rolle bei der Bestimmung der Materialcharakteristika wie Festigkeit, Zähigkeit und Schmelzverhalten. [20]

Vernetzungen

Vernetzte Polymere entstehen durch die kovalente Verbindung von Makromolekülketten, die ein dreidimensionales Netzwerk formen. Diese Transformation vom thermoplastischen zum duroplastischen Zustand führt zu verbesserter thermischer Stabilität und chemischer Beständigkeit. Vernetzte Polymere sind für ihre schwer lösliche und schwer schmelzbare Beschaffen-

heit bekannt, was sie besonders geeignet für Anwendungen unter extremen Bedingungen macht. [32]

1.2.3 Elektrische Bandlücke in Polymeren

Einleitung

Die elektrische Bandlücke, ein zentrales Konzept in der Festkörperphysik, spielt eine entscheidende Rolle bei der Bestimmung der elektronischen und optischen Eigenschaften eines Materials. Sie wird definiert als der energetische Abstand zwischen dem Valenzband und dem Leitungsband eines Festkörpers. Die Größe dieser Bandlücke bestimmt, ob ein Material als Leiter, Halbleiter oder Isolator fungiert. Die elektrische Bandlücke in Polymeren ist ein Maß für die Energie, die benötigt wird, um ein Elektron aus dem Valenzband in das Leitungsband zu befördern, und bestimmt die elektronischen Eigenschaften des Materials. [48]

Grundlagen der Bandlücke

Im Bändermodell sind elektronische Zustände innerhalb von Energiebändern organisiert, wobei das Valenzband im Grundzustand mit Elektronen gefüllt ist und das Leitungsband leer bleibt. Die Bandlücke oder verbotene Zone, beschreibt den Energiebereich zwischen diesen Bändern, in dem keine stabilen elektronischen Zustände vorhanden sind. Diese Lücke bestimmt, ob das Material ein Isolator, Halbleiter oder Leiter ist. [37]

Materialien mit großen Bandlücken agieren als Isolatoren oder schlechte elektrische Leiter, da eine Anregung der Elektronen vom Valenz- ins Leitungsband erschwert ist. Kleinere Bandlücken oder deren Abwesenheit erleichtern diese Anregung und erhöhen somit die Leitfähigkeit. Leiter haben eine Bandlücke von 0 eV, wohingegen Halbleiter einen Wert von 0,1 bis 4 eV besitzen. [24] Nichtleiter haben einen Wert größer als 4 eV.

1.2.4 Einführung in die Chemoinformatik

Chemoinformatik ist ein Bereich der Informatik, der sich mit der Anwendung von computergestützten Methoden und Techniken auf chemische Daten und Probleme befasst. Es ist ein Anwendungsgebiet des maschinellen Lernens, das sich insbesondere auf die Analyse und Vorhersage von molekularen Eigenschaften konzentriert.[38]

Um Untersuchungen an Molekülen durchführen zu können, müssen diese zunächst in einer für Algorithmen verständlichen Form repräsentiert wer-

den. In frühen Studien wurden Moleküle anhand vorab festgelegter struktureller Eigenschaften dargestellt und untersucht. Dabei wurden wichtige strukturelle Merkmale ausgewählt und in einem geeigneten Format codiert, um sie für algorithmische Analysen zugänglich zu machen. [16]

1.2.5 SMILES

In der Domäne der chemischen Informatik markiert das Simplified Molecular Input Line Entry System (SMILES), initiiert von David Weininger im Jahre 1988, einen signifikanten Meilenstein. Dieses Verfahren kodiert chemische Strukturen in eine linearisierte, textuelle Notation, die fundamentale Aspekte wie die molekulare Komposition, Konnektivität sowie fakultativ die Stereochemie abbildet. Die Etablierung von SMILES ist auf seine Effizienz bei der Abbildung komplexer molekularer Konfigurationen in einer sowohl kompakten als auch interpretierbaren Form zurückzuführen, wodurch es einen substantiellen Beitrag zur Erleichterung des Informationsaustausches, der Analyse und des Managements chemischer Daten leistet.

Weininger präsentierte 1988 SMILES als eine vielseitige chemische Sprache, die nicht nur eine Vielzahl molekularer Strukturen abbilden kann, sondern auch eine effiziente Verarbeitung und Suche von Informationen in chemischen Datenbanken unterstützt. SMILES ermöglicht es, Moleküle durch eine Zeichensequenz darzustellen, in der jedes Zeichen spezifische Atome, Bindungen oder stereochemische Konfigurationen symbolisiert. Wasser (H_2O) wird in SMILES beispielsweise mit "O" dargestellt, wobei die Wasserstoffatome implizit bleiben. Ethanol ($\text{CH}_3\text{CH}_2\text{OH}$) wird durch "CCO" repräsentiert, was dessen lineare Molekülstruktur veranschaulicht. [68]

Die Implementierung von SMILES stellte einen bedeutenden Fortschritt in der chemischen Informatik dar, welcher die Digitalisierung und Speicherung chemischer Daten nachhaltig beeinflusste und die Basis für die Entwicklung weiterführender chemischer Kodierungssysteme und Datenbanktechnologien bildete. Aufgrund seiner Flexibilität und Einfachheit ist SMILES zu einem zentralen Instrument in der chemischen Forschung geworden.

1.2.6 Polymere SMILES

SMILES-Notation für Polymere: Die chemischen Strukturen von Polymeren werden mit PSMILES-Strings dargestellt. Diese folgen der herkömmlichen SMILES-Syntax, verwenden jedoch zwei Sternzeichen (*), um die beiden Endpunkte der repetitiven Einheit des Polymers zu kennzeichnen.[34]

Um zu illustrieren, wie SMILES-Strings für Moleküle in PSMILES-Strings für Polymere überführt werden, betrachten wir Ethanol. Ein einfaches Molekül wie Ethanol, das in SMILES als CCO dargestellt wird, kann in einen PSMILES-String umgewandelt werden, indem die Endpunkte der Monomereinheit mit Sternen markiert werden. Für Ethanol als wiederholende Einheit in einem Polymer würde der entsprechende PSMILES-String [*]CCO[*] lauten. Dies zeigt an, dass die Moleküle an diesen Stellen verbunden werden können, um das Polymer zu bilden. [34]

1.2.7 Polymer fingerprints

In der modernen Materialwissenschaft und insbesondere im Bereich der Polymertechnologie spielt die Datenanalyse eine entscheidende Rolle bei der Entdeckung und Charakterisierung neuer Materialien. Um diese Daten effektiv für maschinelles Lernen und computergestützte Analysemethoden nutzbar zu machen, ist die Konvertierung in ein standardisiertes, numerisches Format erforderlich. Dies führt zum Konzept der Polymer-Fingerprints, einem innovativen Ansatz, der es ermöglicht, komplexe chemische Strukturen in eine Form zu bringen, die von Algorithmen effizient verarbeitet werden kann.

Polymer fingerprints, oder einfach Fingerabdrücke, dienen als numerische Repräsentationen der chemo-strukturellen Informationen eines Polymers. Sie sind so konzipiert, dass sie eng mit den materiellen und physikalischen Eigenschaften der Polymere verbunden sind, was eine präzise und aussagekräftige Analyse ermöglicht. Diese Fingerabdrücke sind unerlässlich, um die umfangreichen und oft unstrukturierten Datenmengen in eine maschinenlesbare Form zu transformieren, die für maschinelles Lernen und andere datengetriebene Forschungsansätze geeignet sind. [51, 41, 54, 28, 58, 27]

Ein zentraler Aspekt der Polymer fingerprints ist ihre Fähigkeit, die molekulare Struktur von Polymeren effektiv zu kodieren. Dies geschieht durch die Nutzung des SMILES-Systems (Simplified Molecular-Input Line-Entry System), einer Methode, die ursprünglich für die Darstellung kleiner Moleküle entwickelt wurde und auf Polymere erweitert worden ist. Die SMILES-Notation ermöglicht es, die Wiederholeinheiten von Polymeren als eine Sequenz von Zeichen zu beschreiben, was eine präzise und kompakte Darstellung ihrer chemischen Struktur ermöglicht. Polymer Genome unterstützen die Darstellung zweier Hauptklassen von Polymeren, linearer und Leiterpolymere, durch die Angabe der Verbindungspunkte in den Wiederholeinheiten, was eine differenzierte Analyse und Modellierung ermöglicht.

Die Erstellung von Fingerabdrücken in Polymer Genome basiert auf der

Eingabe von Polymer-SMILES-Strings, aus denen dann numerische Fingerabdruckvektoren generiert werden. Diese Vektoren bestehen aus verschiedenen Komponenten, die unterschiedliche Längenskalen der Polymerstruktur abbilden - von atomaren Details bis hin zu charakteristischen Merkmalen der Polymerketten. Diese mehrstufige Darstellung ist entscheidend, um eine Vielzahl von physikalischen und chemischen Prozessen zu erfassen, die die Eigenschaften der Polymere bestimmen. [16]

Es gibt verschiedene Arten von Polymer-Fingerprints, die sich in ihrer Dimensionalität und Dichte unterscheiden. Einige Fingerprints können viele Komponenten haben und sind größtenteils leer (spärlich besetzt), was bedeutet, dass viele ihrer Werte Null sind. Diese spärlichen Fingerprints erfassen oft das Vorhandensein oder Fehlen spezifischer chemischer Gruppen in einem Polymer. Andere Fingerprints sind dicht besetzt und haben weniger Komponenten, was bedeutet, dass fast alle ihre Werte Informationen tragen. [34]

1.3 Maschinelles Lernen und Auto-ML

Maschinelles Lernen (ML) ist ein zentraler Zweig der künstlichen Intelligenz und befähigt Computer, aus Daten zu lernen und Entscheidungen autonom zu treffen. Ein Bereich im ML sind neuronale Netze, die komplexe Muster in Daten erkennen können. Diese Fähigkeit wird durch einen Prozess erreicht, der als *Training* bekannt ist, währenddessen ein Modell aus einem spezifischen Datensatz lernt. Ziel ist es, Vorhersagen oder Entscheidungen über neue, bisher unbekannte Daten zu treffen. [57] Darüber hinaus verbessern Modell-Ensembles, die mehrere Lernmodelle kombinieren, die Vorhersagegenauigkeit und Robustheit. Im Kontext der Polymerwissenschaft ermöglicht AutoML (Automated Machine Learning) die automatisierte Erstellung und Optimierung von Modellen, um Eigenschaftsvorhersagen treffen zu können.

Automatisiertes Maschinelles Lernen repräsentiert den Prozess der Automatisierung des gesamten Workflows der Anwendung von maschinellem Lernen zur Problemlösung. Dieser Prozess, wie in Abbildung 1.3 dargestellt, umfasst verschiedene Schritte, die darauf abzielen, maschinelles Lernen für Nicht-Experten zugänglich zu machen, die Effizienz des maschinellen Lernens zu verbessern und die maschinelle Lernforschung zu beschleunigen. [31]

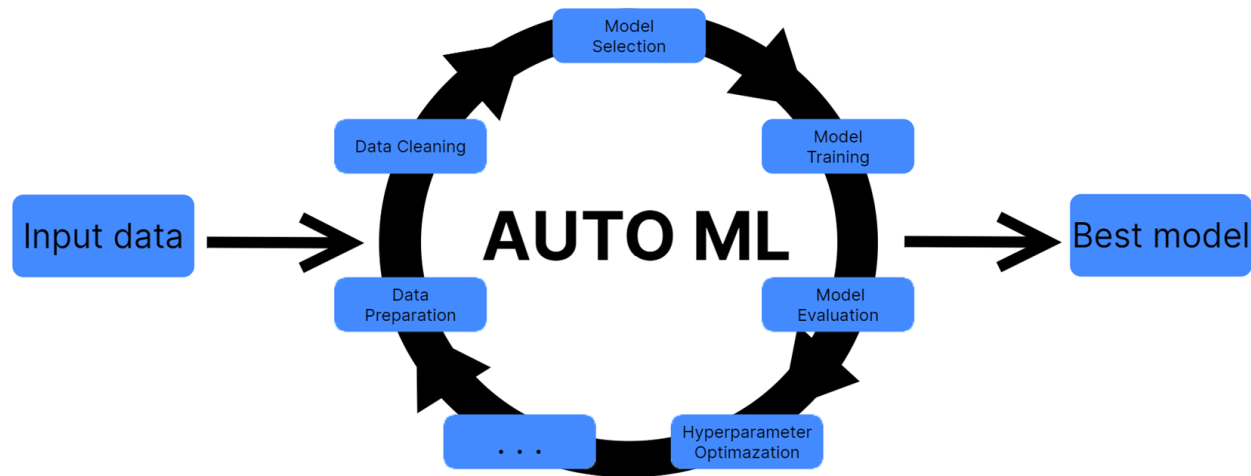


Figure 1.3: Der Prozess des Automatisierten Maschinellen Lernens (AutoML)

1.3.1 Regression

In der vorliegenden Arbeit wird Regression verwendet, um die elektrische Bandlücke in Polymeren vorherzusagen. Regression ist eine statistische Methode zur Untersuchung der Beziehungen zwischen abhängigen und unabhängigen Variablen. Sie wird eingesetzt, um zu verstehen, wie die abhängige Variable (hier die Bandlücke) auf eine oder mehrere unabhängige Variablen reagiert. Diese Methode ist besonders relevant, wenn es darum geht, kausale Beziehungen zu erkennen und zukünftige Werte zu prognostizieren. [69]

Um das Prinzip der Regression zu veranschaulichen, wird häufig die lineare Regression herangezogen. Lineare Regression, als eine der grundlegendsten Formen der Regression, zielt darauf ab, eine lineare Beziehung zwischen den abhängigen und unabhängigen Variablen herzustellen. Sie wird durch die Gleichung

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \epsilon$$

dargestellt, wobei y die abhängige Variable, x_1, \dots, x_n die unabhängigen Variablen, β_0, \dots, β_n die Regressionskoeffizienten und ϵ den Fehlerterm repräsentieren. Durch das Training des Modells wird versucht, die Koeffizienten so anzupassen, dass der Fehler zwischen den vorhergesagten und den tatsächlichen Werten minimiert wird. [40]

Bewertungsmetriken für Regressionsmodelle

In dieser Arbeit werden spezifische Bewertungsmetriken herangezogen, um die trainierten Regressionsmodelle zu testen und zu validieren. Diese Metriken dienen dazu, die Leistung der Modelle hinsichtlich ihrer Genauigkeit und Zuverlässigkeit in der Vorhersage zu quantifizieren. Zu den zentralen Metriken, die verwendet werden, zählen das Bestimmtheitsmaß (R^2), der mittlere quadratische Fehler (MSE) und der mittlere absolute Fehler (MAE). Diese Größen sind essenziell, um zu beurteilen, wie gut die Modelle die Daten abbilden und Vorhersagen treffen können.

R^2 (Bestimmtheitsmaß)

Das R^2 -Maß quantifiziert, welcher Anteil der Varianz der abhängigen Variable durch das Modell erklärt wird. Es wird wie folgt berechnet:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1.1)$$

MSE (Mean Squared Error)

Der MSE misst den durchschnittlichen quadratischen Fehler zwischen den tatsächlichen und vorhergesagten Werten:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1.2)$$

MAE (Mean Absolute Error)

Der MAE gibt den durchschnittlichen absoluten Unterschied zwischen den tatsächlichen und vorhergesagten Werten an:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1.3)$$

Diese Fehlermaße sind entscheidend für die Bewertung der Modellleistung, wobei R^2 die Anpassungsgüte, MSE die Varianz und MAE die durchschnittliche Abweichung angibt [14].

y_i repräsentiert die beobachteten Werte, \hat{y}_i die vorhergesagten Werte des Modells und \bar{y} den Mittelwert der beobachteten Werte. n steht für die Anzahl der Beobachtungen in der Stichprobe.

1.3.2 Neuronale Netze und maschinelles Lernen

Einführung und Klassifizierungsaufgaben

Künstliche neuronale Netze (KNN) bestehen aus zahlreichen einfachen Einheiten, den sogenannten Neuronen, die in größere Strukturen miteinander vernetzt sind. Die Klassifizierungsleistung des Netzwerks wird durch die Gewichte der Verbindungen zwischen den Neuronen bestimmt. Eine zentrale Aufgabe im maschinellen Lernen ist es, Algorithmen bereitzustellen, die fähig sind, Gewichtungen zu finden, die zu einer guten Klassifizierungsleistung führen. Dieser Prozess, gemeinhin als Training des neuronalen Netzwerks bezeichnet, ist entscheidend für die Entwicklung effektiver KNN-Modelle. [33]

Multilayer Perceptrons (MLP)

Ein Multilayer Perceptron (MLP) ist ein künstliches neuronales Netz, das aus mehreren Schichten von Neuronen besteht, wobei jede Schicht mit der nächsten vollständig verbunden ist. Diese Struktur ermöglicht es dem MLP, komplexe nichtlineare Beziehungen zwischen Eingabe- und Ausgabedaten zu modellieren. Die Grundfunktion eines Neurons in einem MLP ist es, eine gewichtete Summe der eingehenden Signale zu bilden und diese Summe einer Transfer- bzw. Aktivierungsfunktion zu unterziehen. Die am häufigsten verwendete Aktivierungsfunktion in frühen MLPs ist die Sigmoid-Funktion, definiert als $f(a) = \frac{1}{1+e^{-a}}$, wobei a die gewichtete Summe der Eingaben ist. Diese Funktion hat die Eigenschaft, dass sie einen Eingabewert in den offenen Intervall (0,1) abbildet, wobei der Wertebereich ideal für die Modellierung von Wahrscheinlichkeiten ist. [33]

Vorwärtspropagierung und universelle Klassifikationseigenschaften

Die Vorwärtspropagierung ist der Prozess, bei dem Eingabedaten durch das Netzwerk geleitet werden, um eine Vorhersage zu generieren. Dies geschieht, indem die Eingabewerte mit den entsprechenden Gewichten multipliziert, summiert und durch die Aktivierungsfunktion transformiert werden, bis die Ausgabeschicht erreicht wird. Das Ergebnis dieses Prozesses ist eine Reihe von Ausgabewerten, die als Evidenz für die Zugehörigkeit zu bestimmten

Klassen interpretiert werden können. Die MLP-Architektur kann prinzipiell jede realistische Funktion mit beliebiger Genauigkeit approximieren, was sie zu einem universellen Klassifikator macht. Dies bedeutet, dass MLPs grundsätzlich für fast jede Klassifizierungsaufgabe eingesetzt werden können, obwohl die spezifische Anzahl an versteckten Neuronen und die exakten Gewichtswerte für eine optimale Leistung sorgfältig ermittelt und eingestellt werden müssen. [33]

Fehlerbewertung in Neuronalen Netzen

Um die Technik für das Training von MLPs einzuführen, ist es notwendig, genauer zu betrachten, wie die Genauigkeit ihrer Klassifikationsentscheidungen quantifiziert werden können. Ein grundlegendes Maß ist die Fehlerrate, die jedoch bestimmte Limitationen aufweist. Die Fehlerrate errechnet sich aus der Anzahl der Fehler dividiert durch die Anzahl der klassifizierten Beispiele. Dieses Maß gibt allerdings nur ein grobes Bild des Klassifikationsverhaltens wieder und berücksichtigt nicht die Fähigkeit der Sigmoid-Funktion, die Größe jedes Fehlers zu messen. [33]

Backpropagation des Fehlers Die Backpropagation des Fehlers ist ein Verfahren zur systematischen Anpassung der Gewichte in einem Multilayer Perceptron (MLP), um die Klassifikationsleistung zu optimieren. Der Prozess beginnt mit der Initialisierung der Gewichte mit kleinen Zufallszahlen, typischerweise im Intervall $[-0.1, 0.1]$. Die Trainingsbeispiele werden nacheinander präsentiert und durch das Netzwerk vorwärts propagiert, um die Diskrepanz zwischen dem Ausgabevektor des Netzwerks und dem Zielvektor des Beispiels zu bestimmen. [33]

Formeln zur Anpassung der Gewichte Die Anpassung der Gewichte basiert auf dem Gradienten der Fehlerfunktion. Für die Neuronen der Ausgangsbeschicht wird die Verantwortung für den Gesamtfehler $\delta_i^{(1)}$ wie folgt berechnet:

$$\delta_i^{(1)} = y_i(1 - y_i)(t_i - y_i)$$

Hierbei bedeuten:

- y_i : Die Ausgabe des Neurons i in der Ausgangsbeschicht.
- t_i : Der tatsächliche Wert der Ausgabe für das Eingabebeispiel i .

Diese Berechnung nutzt die erste Ableitung der Sigmoid-Funktion, die die Vorhersage des Neurons bezüglich der Klassenzugehörigkeit eines Beispiels

ausdrückt. Für die Neuronen der verborgenen Schicht wird die Verantwortung $\delta_j^{(2)}$ durch Rückwärtspropagierung der Verantwortungen der Ausgabeneuronen ermittelt:

$$\delta_j^{(2)} = h_j(1 - h_j) \sum_i \delta_i^{(1)} w_{ji}^{(1)}$$

Hierbei bedeuten:

- h_j : Die Aktivierung des Neurons j in der verborgenen Schicht.
- $w_{ji}^{(1)}$: Das Gewicht zwischen Neuron i in der Ausgabeschicht und Neuron j in der verborgenen Schicht.

[33]

Gewichtsaktualisierungen Die Gewichte werden aktualisiert, indem die ermittelten Verantwortungen mit der Lernrate η und dem Ausgangswert des vorherigen Neurons (für Ausgabeneuronen) oder dem Eingabewert (für verborgene Neuronen) multipliziert werden. Die Anpassungen sind typischerweise klein, aber relativ signifikant für die Optimierung des Netzwerkverhaltens.

$$\begin{aligned} w_{ji}^{(1)} &:= w_{ji}^{(1)} + \eta \delta_i^{(1)} h_j \\ w_{kj}^{(2)} &:= w_{kj}^{(2)} + \eta \delta_j^{(2)} x_k \end{aligned}$$

[33]

1.3.3 Model Ensembles im maschinellen Lernen

Im Bereich des maschinellen Lernens hat sich der Einsatz von Model Ensembles als effektive Strategie erwiesen, um die Genauigkeit von Vorhersagemodellen zu verbessern und das Risiko von Überanpassung (Overfitting) zu verringern. Ein Ensemble besteht aus mehreren Modellen, deren Vorhersagen kombiniert werden, um ein robustes Gesamtmodell zu schaffen. Die gängigsten Techniken zur Erstellung von Ensembles umfassen Bagging, Boosting und Stacking. [33]

Bootstrap Aggregating

Bootstrap Aggregating (Bagging) generiert mehrere unabhängige Modelle durch zufälliges Ziehen von Stichproben aus den Trainingsdaten. Diese Modelle werden mit einem einheitlichen Lernverfahren erstellt. Ihre Vorhersagen werden entweder durch Mehrheitsentscheidung (bei Klassifikation)

oder durch Mittelwertbildung (bei Regression) zu einer Gesamtvorhersage vereint. Ein prominentes Beispiel für Bagging ist der *Random-Forest*-Algorithmus, der auf Entscheidungsbäumen basiert. [33]

Boosting

Im Gegensatz zu Bagging baut Boosting die Modelle sequentiell auf. Jedes Modell lernt dabei von den Fehlern des vorherigen, wodurch insbesondere schwierige Fälle besser erkannt werden sollen. Dieser Ansatz kann die Leistung verbessern, birgt jedoch ein erhöhtes Risiko für Overfitting. Ein bekanntes Beispiel für Boosting ist das *Adaptive Boosting* (AdaBoost), das die Leistung durch Fokussierung auf die Fehler der vorherigen Modelle steigert. [33]

Stacking

Stacking unterscheidet sich von Bagging und Boosting dadurch, dass es verschiedene Lernverfahren für die Erstellung der Modelle im Ensemble verwendet. Die Vorhersagen dieser Modelle können, ähnlich wie bei Bagging, durch Mehrheitsentscheid oder Mittelwert kombiniert werden. Ein alternativer Ansatz ist jedoch, ein weiteres Modell zu trainieren, das diese Vorhersagen als Eingaben nutzt, um eine finale Vorhersage zu erzeugen. Diese Methode ermöglicht eine effektive Integration verschiedener Modellperspektiven.

Ensemble-Methoden haben sich in der Praxis als äußerst wertvoll erwiesen, um die Präzision und Robustheit maschineller Lernmodelle zu steigern. Sie bieten vielfältige Ansätze, um die Stärken einzelner Modelle zu kombinieren und ihre Schwächen zu mindern. [33]

1.4 HCI und Nutzerstudien

Nutzerstudien sind ein wesentlicher Bestandteil des Human-Computer Interaction (HCI) Forschungsbereichs. Sie dienen dazu, die Interaktion zwischen Nutzern und Computersystemen zu verstehen und zu verbessern. In meiner Studie wurde die Think-Aloud Methode und die Likert-Skala-Evaluierung verwendet, um Einblicke in die Benutzererfahrung und die Bedienbarkeit der untersuchten Systeme zu gewinnen.

1.4.1 Arten von Interviews

Interviews stellen eine grundlegende Methode dar, um mit potenziellen Nutzern in Kontakt zu treten und Informationen zu sammeln, die dazu dienen, das Design interaktiver Technologien zu gestalten und zu bewerten. Ihr Zweck besteht darin, ein tieferes Verständnis für die Bedürfnisse, Verhaltensweisen und Anforderungen der Nutzer zu erlangen, um darauf aufbauend Designs zu entwickeln und zu verbessern. Der Begriff des Interviews kann als "gezielte Konversation" betrachtet werden, wie bereits von Kahn und Cannel (1957) beschrieben. Es existieren verschiedene Formen von Interviews, darunter offene, strukturierte und halbstrukturierte Interviews sowie Gruppeninterviews oder Fokusgruppen.

Offene Interviews

Offene Interviews zeichnen sich dadurch aus, dass weder der Interviewer noch der Interviewte klare Erwartungen hinsichtlich des Formats oder der Inhalte der Antworten haben. Vielmehr werden die Gespräche von einem offenen Austausch geprägt, der es ermöglicht, die Gedanken und Perspektiven der Beteiligten freizulegen. Dies fördert eine explorative Herangehensweise und ermöglicht es, tiefgreifende Einblicke zu gewinnen. [15]

Strukturierte Interviews

Strukturierte Interviews hingegen folgen einem vordefinierten Schema, bei dem standardisierte Fragen in jedem Interview gestellt werden. Diese Fragen sind in der Regel kurz und geschlossen formuliert, was bedeutet, dass den Befragten eine begrenzte Auswahl an Antwortmöglichkeiten gegeben wird. Diese Art von Interview eignet sich besonders dann, wenn klare Ziele und Fragestellungen vorliegen, da sie es ermöglicht, Daten systematisch zu sammeln und zu analysieren. [15]

Halbstrukturierte Interviews

Halbstrukturierte Interviews stellen eine Mischform dar, bei der zwar vorab bestimmte Fragen festgelegt werden, jedoch Raum für Flexibilität und offene Diskussionen bleibt. Dies ermöglicht es, bestimmte Themenbereiche gezielt zu erforschen, während gleichzeitig genügend Freiraum für spontane Reaktionen und neue Erkenntnisse bleibt.

Die Auswahl der Interviewmethode sollte stets an die spezifischen Ziele der Studie sowie an die Art der Informationen, die gesammelt werden sollen,

angepasst werden. Dabei ist es wichtig, die Vor- und Nachteile der verschiedenen Ansätze zu berücksichtigen und die Methode auszuwählen, die am besten zur Untersuchungsfrage und zum Forschungskontext passt. [15]

1.4.2 Beobachtungen

Nachdem wir die verschiedenen Interviewmethoden betrachtet haben, ist es wichtig, auch die potenziellen Probleme zu erkennen, die mit diesen Methoden verbunden sein können.

Ein Hauptproblem liegt in der Diskrepanz zwischen den Erwartungen und dem tatsächlichen Verhalten der Menschen. Oft handeln Menschen nicht so, wie sie sagen oder denken, dass sie handeln werden. Darüber hinaus kann das menschliche Gedächtnis unzuverlässig sein, da Details aus vergangenen Erfahrungen vergessen werden können, die für das Design relevant wären.

Um diese Herausforderungen zu bewältigen, ist es entscheidend, Beobachtungen durchzuführen, um die Kontexte, Aufgaben und Bedürfnisse der Nutzer besser zu verstehen. Beobachtungen können dazu dienen, Menschen bei der Nutzung eines Designs zu beobachten und somit Informationen für die Evaluation zu liefern. [42]

Es gibt verschiedene Arten von Beobachtungen, darunter direkte und indirekte Beobachtungen. Direkte Beobachtungen ermöglichen es, das Verhalten der Menschen mit eigenen Augen zu beobachten, während indirekte Beobachtungen auf der Analyse von Aufzeichnungen der Aktivitäten der Menschen basieren, wie beispielsweise Tagebuchstudien. [3][52]

Des Weiteren unterscheiden sich Beobachtungen hinsichtlich des Ortes, an dem sie durchgeführt werden. Laborbeobachtungen finden in kontrollierten Umgebungen statt, während Feldbeobachtungen in den natürlichen Umgebungen der Nutzer durchgeführt werden. Beide Ansätze haben ihre Vor- und Nachteile, und die Wahl zwischen ihnen hängt von den spezifischen Anforderungen der Forschung ab.[49] [18]

Think-Aloud

Die Think-Aloud-Methode ist eine effektive Technik, die es ermöglicht, Einblicke in die kognitiven Prozesse der Nutzer zu gewinnen. Wie Jørgensen (1990) hervorhebt, fördert diese Methode die kognitive Ergonomie, indem sie zeitnahe, authentische und anwendbare Rückmeldungen an die Designer im Designkontext bietet [30]. Eccles und Aarsal (2017) betonen weiterhin die Nützlichkeit der Methode zur Untersuchung des Denkens und beschreiben

detailliert, wie sie in verschiedenen Forschungskontexten angewendet werden kann [17].

Feldforschung

Die Feldforschung ist eine weitere wichtige Methode der Beobachtung, bei der Beobachtungen in der natürlichen Umgebung der Benutzer durchgeführt werden. Indem man das Büro verlässt und Menschen in ihrer alltäglichen Umgebung beobachtet, können unerwartete Erkenntnisse gewonnen werden. Trotz ihrer Vorteile können Feldstudien jedoch teuer sein und aufgrund von Datenschutz- und Sicherheitsbedenken sowie anderen Einschränkungen Herausforderungen mit sich bringen. [49]

Laborbeobachtung

Laborbeobachtungen sind eine Schlüsselkomponente in den Sozialwissenschaften, die es Forschern ermöglicht, Verhaltensweisen in einer kontrollierten Umgebung zu studieren. Durch die Kontrolle von Variablen können Forscher kausale Beziehungen zwischen Faktoren identifizieren. Falk und Heckman (2009) argumentieren, dass Laborexperimente eine bedeutende Quelle des Wissens in den Sozialwissenschaften sind, da sie präzise und kontrollierbare Bedingungen für die Untersuchung menschlichen Verhaltens bieten. Trotz der Kritik, dass Laborexperimente möglicherweise nicht die Komplexität realer Situationen widerspiegeln, betonen sie die Wichtigkeit von Laborexperimenten für die Erzeugung replizierbarer und überprüfbarer Ergebnisse, was für die Validierung von Theorien in den Sozialwissenschaften unerlässlich ist [18].

1.4.3 Fragebögen

Fragebögen sind eine verbreitete Methode, um Nutzer selbstständig zu befragen und Informationen für die Gestaltung interaktiver Technologien zu sammeln oder zu bewerten. Sie ermöglichen es, ein Verständnis für Menschen, Aufgaben und Bedürfnisse zu entwickeln und tragen zur Gestaltung und Bewertung von Designs bei. [10]

Vorteile von Fragebögen

Fragebögen bieten einige Vorteile:

- **Weite Verbreitung:** Fragebögen ermöglichen die Sammlung einer

breiten Palette von Informationen von einer großen Anzahl von Personen.

- **Wichtiges Instrument:** Bei korrekter Konstruktion und Verwaltung werden Fragebögen zu einem vitalen Instrument, das präzise Aussagen über bestimmte Gruppen oder Personen ermöglichen kann.
- **Vielfältige Informationsgewinnung:** Sie sind effizient in der Datenerhebung und können genutzt werden, um umfassende Daten aus der Zielgruppe zu erhalten.
- **Kritische Konstruktionsaspekte:** Die angemessene Erstellung von Fragebögen, einschließlich angemessener Fragenstellung und korrekter Skalierung, ist entscheidend für den Erfolg einer Umfrage.

[53]

Design von Fragebögen

Das Design von Fragebögen ist entscheidend für die Gewinnung aussagekräftiger Daten. Laut Bee und Murdoch-Eaton (2016) sollten beim Design folgende Aspekte berücksichtigt werden:

- Der **Zweck des Fragebogens** muss klar definiert sein, um sicherzustellen, dass er die benötigten Informationen effizient sammelt.
- Das **Format und die Formulierung** der Fragen sind entscheidend; der Fragebogen sollte attraktiv, leicht navigierbar sein und seine Länge sollte so kurz wie möglich gehalten werden.
- Die Entscheidung zwischen **offenen oder geschlossenen Fragen** oder einer Kombination von beiden beeinflusst die Qualität der gewonnenen Daten. Geschlossene Fragen bieten leicht handhabbare Daten, während offene Fragen tiefere Einblicke geben können.
- **Fragekonstruktion:** Fragen sollten unmissverständlich formuliert und auf einen einzigen Punkt fokussiert sein, in einer Sprache, die für die Zielgruppe verständlich ist.

Diese Prinzipien tragen dazu bei, dass ein Fragebogen sinnvolle Antworten liefert und seine Ergebnisse eine zuverlässige Grundlage für Forschungserkenntnisse bilden. [11]

Chapter 2

AutoML-Frameworks

2.1 Auswahl der AutoML-Frameworks

Die richtige Auswahl von AutoML Frameworks spielt eine entscheidende Rolle in der Forschung zur Vorhersage von Bandlücken in Polymeren. AutoML vereinfacht die Entwicklung und Anwendung von maschinellen Lernmodellen erheblich, indem es die Komplexität und den Zeitaufwand, der mit traditionellen Machine Learning-Prozessen verbunden ist, reduziert. Durch Automatisierung von Aufgaben wie der Auswahl von Modellen, der Optimierung von Hyperparametern und der Validierung von Modellen ermöglicht AutoML auch Forschenden ohne tiefgreifende Expertise im Bereich des maschinellen Lernens den Zugang zu fortschrittlichen ML-Techniken. Dies ist von besonderer Bedeutung in spezialisierten Feldern wie der Materialwissenschaft, wo die genaue Vorhersage spezifischer Eigenschaften – wie beispielsweise der Bandlücken in Polymeren – essentiell für die Entwicklung neuer Materialien ist. Die Auswahl von AutoML-Frameworks, die ein breites Spektrum maschinellen Lernens abdecken – von Ensemble-Methoden und neuronalen Netzwerken bis hin zu traditionellen ML-Algorithmen –, ist daher von entscheidender Bedeutung, um nicht nur die Genauigkeit und Effizienz der Vorhersagemodelle zu maximieren, sondern auch um systematisch zu untersuchen, welcher der maschinellen Lernansätze im Kontext von AutoML die beste Performance bietet.

2.2 Vorstellung AutoML-Bibliotheken

Um ein breites Spektrum an maschinellen Lernmethoden abzudecken und die Genauigkeit sowie Effizienz der Vorhersagemodelle für Bandlücken in Polymeren zu optimieren, wurden fünf führende AutoML-Bibliotheken ausgewählt: AutoGluon, AutoKeras, Auto-sklearn, PyCaret und Ludwig. Die Auswahl dieser Bibliotheken wird im Diagramm in Abbildung 2.1 dargestellt. Diese Bibliotheken wurden speziell aufgrund ihrer umfassenden Unterstützung für verschiedene maschinelle Lernansätze, einschließlich Ensemble-Methoden, neuronalen Netzen und traditionellen ML-Algorithmen, ausgewählt. Jede dieser Bibliotheken bringt spezifische Stärken in den Forschungsprozess ein, die im Folgenden näher erläutert werden.

2.2.1 AutoGluon

AutoGluon repräsentiert eine fortschrittliche Bibliothek im Bereich des Automated Machine Learning (AutoML), die speziell für die Arbeit mit tabellarischen Daten konzipiert wurde. Eines der Hauptmerkmale von AutoGluon ist seine Fähigkeit, hochgenaue Modelle durch einen einfachen Aufruf der `fit()` Funktion zu erstellen. Diese Modelle sind in der Lage, die Werte einer Spalte eines Datensatzes basierend auf den Werten der übrigen Spalten vorherzusagen. Diese Eigenschaft macht AutoGluon zu einem vielseitigen Werkzeug, das sowohl für Klassifikations- als auch für Regressionsprobleme eingesetzt werden kann. [9]

Standardmäßig versucht AutoGluon, verschiedene Arten von Machine learning Modellen zu trainieren, einschließlich neuronaler Netzwerke und Baumensembles. Jeder Modelltyp verfügt über verschiedene Hyperparameter, die traditionell vom Benutzer spezifiziert werden müssen. AutoGluon automatisiert diesen Prozess durch die automatische und iterative Testung von Hyperparameterwerten, um die beste Leistung auf den Validierungsdaten zu erzielen. Dies umfasst das wiederholte Trainieren von Modellen unter verschiedenen Hyperparametereinstellungen und die Bewertung ihrer Leistung.

Dieser Prozess kann rechenintensiv sein, weshalb `fit()` diesen Vorgang mithilfe von Ray über mehrere Threads parallelisiert. Durch diese Automatisierung der Hyperparameteroptimierung kann AutoGluon effizient Modelle erstellen, die hochgradig angepasst sind und eine optimale Leistung erbringen, ohne dass vom Nutzer eine manuelle Feinabstimmung erforderlich ist.[9]

Neben der Automatisierung des Hyperparameter-Tunings bietet AutoGluon auch fortschrittliche Funktionen im Bereich des Feature-Engineerings für

tabellarische Daten. Feature-Engineering ist der Prozess, bei dem rohe tabellarische Daten so umgewandelt werden, dass sie von einem maschinellen Lernmodell effektiv genutzt werden können. Dies umfasst die Konvertierung der Daten in ein für das Modell lesbares Format und die Verbesserung bestimmter Spalten (Features), um den ML-Modellen mehr Informationen zu liefern und somit genauere Ergebnisse zu erzielen. [8] AutoGluon automatisiert einen Teil dieses Prozesses und unterstützt eine Vielzahl von Feature-Typen, einschließlich boolescher, numerischer, kategorischer, Datums- und Textdaten. Es erkennt und verarbeitet diese Datentypen automatisch, wobei spezielle Behandlungen für jede Kategorie angewendet werden, um die Effektivität der darauf trainierten Modelle zu maximieren. Darüber hinaus bietet AutoGluon Optionen zur Konfiguration und Erweiterung dieser automatisierten Feature-Engineering-Prozesse, um die Anforderungen spezifischer Anwendungsfälle zu erfüllen.

2.2.2 AutoKeras

AutoKeras, ein fortschrittliches AutoML-System basierend auf Keras, wurde vom DATA Lab an der Texas A&M University entwickelt. Es zielt darauf ab, maschinelles Lernen einem breiteren Publikum zugänglich zu machen, indem es den Prozess der Modellauswahl und Hyperparameter-Feinabstimmung automatisiert. Mit einem einfachen und zugänglichen Interface ermöglicht AutoKeras auch Nutzern mit begrenzter Erfahrung in maschinellem Lernen und Programmierung, Standardprobleme des maschinellen Lernens mit nur wenigen Codezeilen zu lösen. AutoKeras ist für die Praxis konzipiert und basiert auf Keras und TensorFlow, was eine einfache Exportierung und Implementierung der mit AutoKeras erstellten Modelle ermöglicht. [29]

Ein besonderes Merkmal von AutoKeras ist der **StructuredDataRegressor**, der für die Arbeit mit strukturierten Daten konzipiert wurde. Dieses Werkzeug ist bemerkenswert flexibel in Bezug auf das Datenformat und unterstützt CSV-Dateien, `numpy.ndarray`, `pandas.DataFrame` und `tf.data.Dataset`. Die Daten sollten zweidimensional sein und können sowohl numerische als auch kategorische Werte enthalten. Für Regressionsziele akzeptiert AutoKeras Vektoren numerischer Werte, wobei sowohl `numpy.ndarray` als auch `pandas.DataFrame` oder `pandas.Series` verwendet werden können. Diese Flexibilität macht AutoKeras zu einem vielseitigen Werkzeug in der AutoML-Landschaft, besonders wenn es um die Verarbeitung und Analyse strukturierter Daten geht. [62]

2.2.3 Auto-sklearn

Auto-sklearn ist ein fortschrittliches automatisiertes Maschinenlernsystem, das auf `scikit-learn` aufbaut und als Drop-in-Ersatz für einen `scikit-learn` Schätzer dient. Es wurde entwickelt, um die ständig wachsende Nachfrage nach Maschinenlernsystemen zu bedienen, die von Nicht-Experten direkt verwendet werden können. Ein solches System muss in der Lage sein, automatisch einen geeigneten Algorithmus und Vorverarbeitungsschritte für einen neuen Datensatz auszuwählen sowie deren jeweilige Hyperparameter einzustellen. Auto-sklearn nutzt hierfür effiziente Bayesianische Optimierungsmethoden und baut auf früheren Arbeiten im Bereich des automatisierten Maschinenlernens (AutoML) auf. Es beinhaltet 15 Klassifikatoren, 14 Feature-Vorverarbeitungsmethoden und 4 Daten-Vorverarbeitungsmethoden, was zu einem strukturierten Hypothesenraum mit 110 Hyperparametern führt. Diese Komponenten ermöglichen Auto-sklearn, aus vergangenen Leistungen auf ähnlichen Datensätzen zu lernen und Ensembles aus den während der Optimierung bewerteten Modellen zu konstruieren, was eine wesentliche Verbesserung gegenüber existierenden AutoML-Methoden darstellt. Auto-sklearn hat sich nicht nur in der ersten Phase der ChaLearn AutoML Challenge durchgesetzt, sondern zeigt auch in einer umfassenden Analyse über mehr als 100 diverse Datensätze eine deutliche Überlegenheit gegenüber dem bisherigen Stand der Technik in AutoML. [19]

Auto-sklearn behandelt AutoML als ein kombiniertes Algorithmusauswahl- und Hyperparameteroptimierungsproblem (CASH), das effizient als ein einziges, strukturiertes, gemeinsames Optimierungsproblem angegangen werden kann. Durch die Nutzung von Meta-Lernen und automatisierter Ensemble-Konstruktion ermöglicht Auto-sklearn die Nutzung aller während der Bayesianischen Optimierung gefundenen Klassifikatoren. Auto-sklearn stellt damit eine robuste und effiziente Lösung im AutoML-Bereich dar, die aufgrund ihrer Flexibilität und Leistungsfähigkeit eine wertvolle Ressource für Anwender bietet, die Zugang zu leistungsfähigem maschinellen Lernen ohne tiefgreifendes Fachwissen suchen. [19]

2.2.4 PyCaret

PyCaret ist eine Open-Source, Low-Code Machine-Learning-Bibliothek in Python, die darauf abzielt, Machine-Learning-Workflows zu automatisieren. Als End-to-End-Tool für Machine Learning und Modellmanagement beschleunigt PyCaret den Experimentzyklus exponentiell und steigert die Produktivität des Anwenders. Im Vergleich zu anderen Open-Source Machine-

Learning-Bibliotheken bietet PyCaret eine alternative Low-Code-Lösung, die es ermöglicht, Hunderte von Codezeilen durch nur wenige Zeilen zu ersetzen. Dadurch werden Experimente deutlich schneller und effizienter. PyCaret dient im Wesentlichen als Python-Wrapper um mehrere Machine-Learning-Bibliotheken und Frameworks, darunter `scikit-learn`, XGBoost, LightGBM, CatBoost, spaCy, Optuna, Hyperopt, Ray und einige mehr. [1]

Zu den Kernfunktionen von PyCaret gehören Datenbereitung, Modelltraining, Hyperparameter-Tuning, Analyse und Interpretierbarkeit, Modelauswahl, sowie Experimentprotokollierung, was es zu einem umfassenden Werkzeug für Data-Science-Projekte macht. Die Gestaltung und Einfachheit von PyCaret ist von der aufkommenden Rolle der Citizen Data Scientists inspiriert, einem Begriff, der erstmals von Gartner verwendet wurde. Citizen Data Scientists sind "Power-User", die sowohl einfache als auch mäßig anspruchsvolle analytische Aufgaben ausführen können, für die zuvor mehr technisches Fachwissen erforderlich war. Erfahrene Datenwissenschaftler sind oft schwer zu finden und teuer einzustellen, aber Citizen Data Scientists können eine effektive Möglichkeit sein, diese Lücke zu schließen und Datenwissenschaftsherausforderungen im Geschäftsumfeld zu bewältigen.

Eines der herausragenden Features von PyCaret ist die `tune_model` Funktion, die speziell für Regressionsaufgaben entwickelt wurde. Diese Funktion automatisiert die Auswahl und Feinabstimmung von Machine-Learning-Algorithmen, indem sie die neuesten Fortschritte in der Bayesianischen Optimierung und anderen Hyperparameter-Optimierungstechniken nutzt. PyCaret ist bereit für den Einsatz, was bedeutet, dass alle Schritte, die in einem ML-Experiment durchgeführt werden, mit Hilfe eines reproduzierbaren und für die Produktion garantierten Pipelines reproduziert werden können. Eine Pipeline kann in einem binären Dateiformat gespeichert werden, das über Umgebungen hinweg übertragbar ist, was die Bereitstellung und Wiederholbarkeit von ML-Experimenten erheblich vereinfacht. [1]

2.2.5 Ludwig

Ludwig, entwickelt von Uber Technologies Inc., ist eine Deep Learning Toolbox, die auf einem typbasierten, deklarativen Ansatz basiert. Ein besonders bemerkenswertes Feature von Ludwig ist die Unterstützung für AutoML, das es ermöglicht, ein abgestimmtes Deep Learning Modell zurückzugeben, gegeben ein Datensatz, die Zielvariable und ein Zeitbudget. Ludwig AutoML befindet sich derzeit noch in der experimentellen Phase und konzentriert sich auf tabellarische Datensätze. Es leitet die Typen der Eingabe- und Ausgabe-Features ab, wählt die Modellarchitektur aus und startet einen Ray Tune

Async HyperBand Suchauftrag über einen Satz von Hyperparametern und deren Bereichen, begrenzt durch das angegebene Zeitbudget. Ludwig AutoML liefert eine Reihe von Modellen, die durch die Suchläufe erzeugt wurden, sortiert von den besten bis zu den schlechtesten, zusammen mit einem Hyperparametersuchbericht, der manuell inspiziert oder durch verschiedene Ludwig-Visualisierungswerkzeuge nachbearbeitet werden kann. [39]

Nutzer können Ludwig AutoML auf verschiedene Weise prüfen und damit interagieren. Dies ermöglicht eine effiziente Entwicklung und Feinabstimmung von Deep Learning Modellen, insbesondere für Anwender, die mit tabellarischen Datensätzen arbeiten. Die Bereitstellung von Ludwig AutoML markiert einen wichtigen Schritt in Richtung zugänglicherem und effizienterem Machine Learning, indem sie komplexe Prozesse der Modellauswahl und Hyperparameteroptimierung automatisiert.

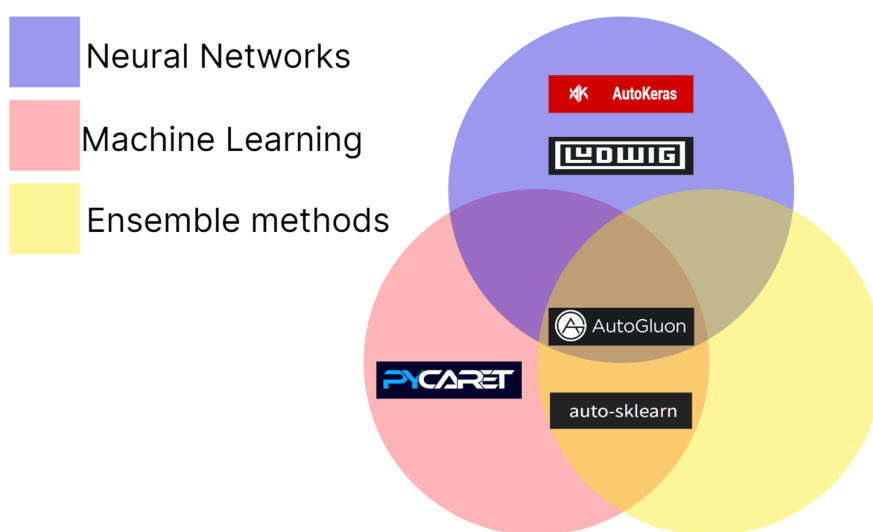


Figure 2.1: Übersicht der Bibliotheken

Durch die Kombination dieser fünf AutoML-Bibliotheken ist es möglich, ein umfassendes Spektrum an ML-Techniken abzudecken und so die Vorhersagegenauigkeit von Bandlücken in Polymeren zu maximieren. Diese Bibliotheken wurden sorgfältig ausgewählt, um die Forschungsziele zu unterstützen und gleichzeitig die Zugänglichkeit und Effizienz des Forschungsprozesses zu gewährleisten.

2.3 Datenvorbereitung

Die Datenvorbereitung ist ein entscheidender Schritt im Prozess des maschinellen Lernens, insbesondere in der Materialwissenschaft, wo die genauen Eigenschaften der Materialien für die Vorhersagemodelle von großer Bedeutung sind. Der verwendete Datensatz für dieses Projekt stammt aus einer Datenbank des Georgia Institute of Technology und ist über Khazana zugänglich. Der Datensatz enthält Informationen zu verschiedenen Polymeren, ihren Eigenschaften und besteht aus vier Spalten: dem Index, den SMILES-Strings der Polymere, der Eigenschaftsbezeichnung und dem Wert für die entsprechende Eigenschaft. Für die Zwecke dieses Projekts wurde der Datensatz nach der Bandlücke (Bandgap), gekennzeichnet als "Egc", gefiltert. Zusätzlich zu dieser fokussierten Auswahl wurde der Datensatz sorgfältig in Trainings- und Testdatensätze aufgeteilt, wobei 80% der Daten für das Training und die verbleibenden 20% für die Testphase reserviert wurden. [21]

Die SMILES-Strings bieten eine textbasierte Darstellung der chemischen Struktur der Polymere. Um diese Daten für den Regressions-Task nutzbar zu machen, ist eine Umwandlung der SMILES-Strings in Fingerabdrücke erforderlich. Dieser Prozess ermöglicht es, die chemische Struktur in einen numerischen Vektor umzuwandeln, der von maschinellen Lernmodellen verarbeitet werden kann.

Für die Umwandlung wurde die Bibliothek psmiles verwendet [50], die eine kanonische Darstellung der SMILES-Strings bietet und es ermöglicht, einzigartige Fingerabdrücke zu erstellen. Diese Umwandlung liefert eine maschinenlesbare Darstellung der chemischen Struktur der Polymere und ermöglicht eine effizientere Verarbeitung durch die AutoML-Modelle.

2.4 Modelltraining und Evaluation

Das Training der Modelle erfolgte unter Verwendung des AutoML-Ansatzes, der die Auswahl des Modells, das Hyperparameter-Tuning und das Training automatisiert. Das Ziel war es, eine effiziente und effektive Methode zur Vorhersage der Bandlücken-Eigenschaft zu finden. Als Leistungsmaße wurden der Determinationskoeffizient (R^2), der Mittlere Quadratische Fehler (MSE), der Mittlere Absolute Fehler (MAE), sowie zwei zusätzliche Metriken berücksichtigt: die Dauer des Trainings und die Einfachheit der Implementierung. Diese Metriken ermöglichen eine umfassende Bewertung der Modelleistung und der Praktikabilität des AutoML-Ansatzes.

Die Einfachheit der Umsetzung reflektiert das Kernziel von AutoML, Machine Learning so zugänglich und benutzerfreundlich wie möglich zu gestalten. Dies ist besonders relevant für Anwendungen, bei denen Zeit und Ressourcen für manuelles Hyperparameter-Tuning und Modellauswahl begrenzt sind.

Chapter 3

Entwicklung des Webtools

3.1 Ziel des Webtools

Das primäre Ziel der entwickelten Webseite ist es, die Barrieren für den Einstieg in die Welt des maschinellen Lernens (ML) signifikant zu reduzieren. Insbesondere richtet sich das Angebot an Nutzer, die zwar ein tiefgehendes Interesse an der Anwendung und Nutzung von ML-Modellen haben, jedoch über begrenzte Erfahrungen im Bereich der Programmierung und Datenwissenschaft verfügen. Die Kernphilosophie hinter diesem Projekt basiert auf den Leitprinzipien von Automated Machine Learning (AutoML), welches darauf abzielt, den Prozess des maschinellen Lernens zu demokratisieren, indem es die Komplexität der Modellauswahl, der Hyperparameter-Optimierung und der Validierung automatisiert. Diese Prinzipien reflektierend, strebt das Webtool danach, den Programmieraufwand für die Nutzer vollständig zu eliminieren und stattdessen eine intuitive, benutzerfreundliche Oberfläche anzubieten, die es ermöglicht, ohne Vorkenntnisse im Codieren effektive ML-Modelle zu trainieren und einzusetzen.

3.2 Aufbau und Entwicklung

Die Entwicklung des Webtools wurde mit Streamlit realisiert, einer Open-Source Python-Bibliothek, die die Erstellung und das Teilen von benutzerdefinierten Web-Apps für Machine Learning und Data Science vereinfacht. Streamlit ermöglicht es, in nur wenigen Minuten leistungsstarke Daten-Apps zu bauen und zu deployen, indem es eine effiziente Verknüpfung von Frontend und Backend in einem einzigen Ort bietet. Dies erle-

ichtet insbesondere die Entwicklung von Anwendungen, bei denen Datenvisualisierung und interaktive Steuerung des Machine-Learning-Prozesses im Vordergrund stehen [61]. Die Webseite ist in fünf Hauptbereiche untergliedert, um eine klare und logische Navigation zu gewährleisten und den Nutzern einen schnellen Zugriff auf alle notwendigen Funktionen zu ermöglichen. Die Umsetzung dieser benutzerfreundlichen Struktur wurde mithilfe der ‘streamlit.tabs’ Funktion von Streamlit realisiert, einer leistungsfähigen Komponente, die es Entwicklern ermöglicht, mehrere Registerkarten innerhalb einer einzigen Streamlit-Anwendung zu erstellen. Nutzer können einfach zwischen verschiedenen Themenbereichen wechseln, ohne die Orientierung zu verlieren.

3.2.1 My Project Tab

Der **My Project** (siehe 3.1) Tab dient als zentrale Anlaufstelle für Nutzer, die sich eingehend mit dem AutoML-Projekt beschäftigen möchten. Hier wird nicht nur ein Überblick über das Projekt gegeben, sondern auch detailliert auf die Performance und die Ergebnisse der verschiedenen getesteten AutoML-Bibliotheken eingegangen. Mithilfe der `streamlit.table` Methode werden diese Informationen übersichtlich in Tabellenform dargestellt, was einen direkten Vergleich der Leistungsfähigkeit der einzelnen Bibliotheken ermöglicht.

Darüber hinaus ist dieser Tab in fünf weitere Registerkarten unterteilt, die jeweils einer spezifischen Bibliothek gewidmet sind: *AutoGluon*, *AutoSklearn*, *AutoKeras*, *PyCaret* und *Ludwig*. Diese Aufteilung wurde durch die Nutzung der `streamlit.tabs` Methode realisiert, wodurch eine klare und benutzerfreundliche Strukturierung des Inhalts gewährleistet wird. Innerhalb jeder dieser Registerkarten findet sich eine ausführliche Beschreibung des verwendeten Codes. Die `streamlit.code` Methode sorgt dabei für eine ansprechende Formatierung und ermöglicht es den Nutzern, den Code direkt zu kopieren.

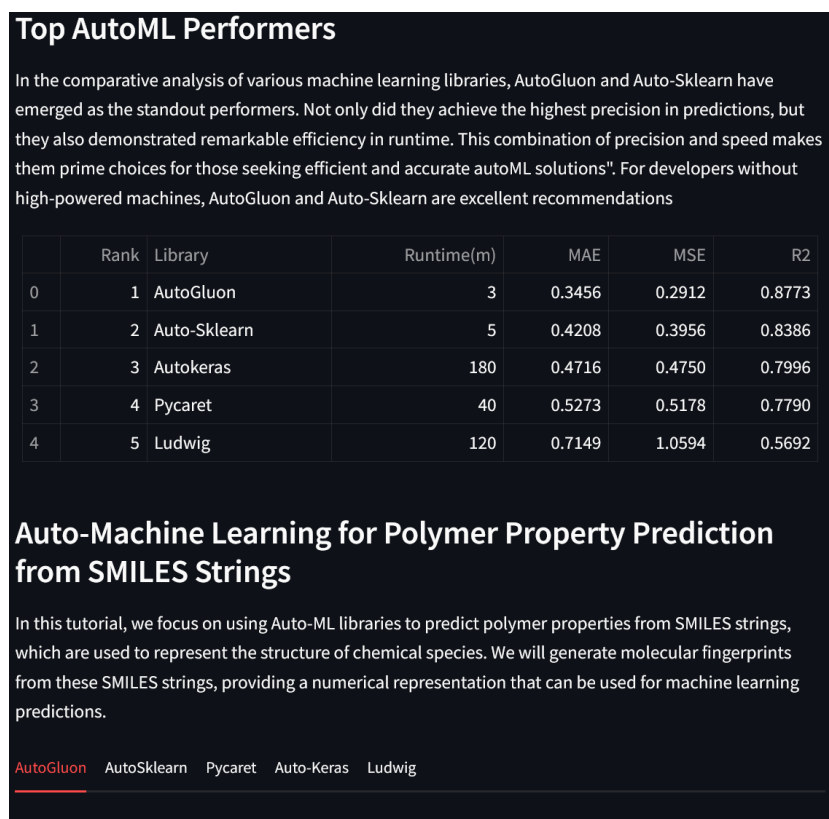


Figure 3.1: My Project tab

3.2.2 Polymere Tab

Der **Polymere Tab** (siehe 3.2) widmet sich voll und ganz dem Thema Polymere und ihrer Bedeutung sowohl in der Materialwissenschaft als auch in der Informatik. Innerhalb dieses Tabs werden die Grundlagen über Polymere erläutert, die auch für Laien verständlich ist. Hauptaugenmerk liegt auf der Darstellung, wie Polymere im Kontext des Machine Learnings verwendet werden, um Vorhersagen für bestimmte Eigenschaften zu treffen. Die Erklärungen und Inhalte werden hauptsächlich durch Textelemente vermittelt, die mit der `streamlit.write` Methode erstellt wurden, was eine klare und leicht verständliche Informationsvermittlung ermöglicht. Zur visuellen Unterstützung und um die Inhalte greifbarer zu machen, wurden Bilder eingesetzt, die mit der `streamlit.image` Methode in die Webseite eingebettet wurden. Diese Bilder dienen nicht nur der ästhetischen Aufwertung, sondern auch der Veranschaulichung komplexer Sachverhalte, wie der Struktur von Polymeren und den Prozessen ihrer Analyse durch ML-Modelle.

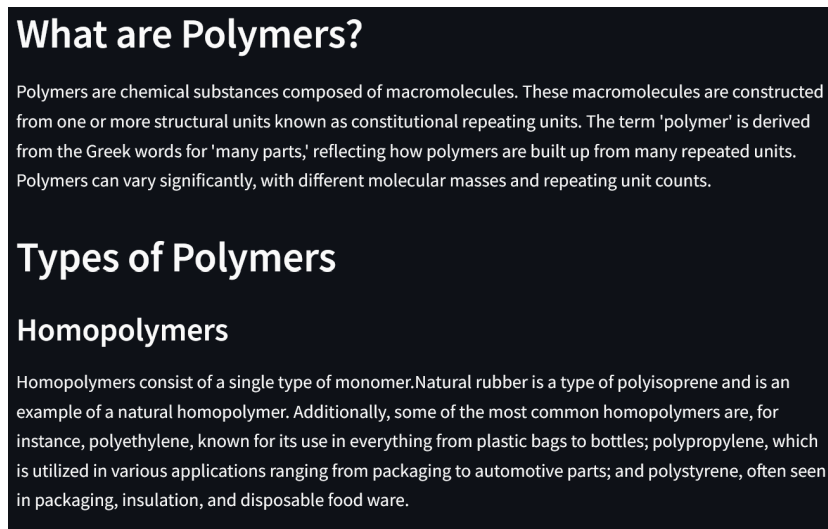


Figure 3.2: Polymere Tab

3.2.3 Automated Machine Learning Tab

Der **Automated Machine Learning Tab** (siehe 3.3) zielt darauf ab, Nutzern ein tiefgehendes Verständnis der grundlegenden Prinzipien und Konzepte des maschinellen Lernens (ML) sowie des Automated Machine Learning (AutoML) zu vermitteln. Diese Sektion ist speziell darauf ausgerichtet, auch Nicht-Experten, die möglicherweise keine Vorkenntnisse in diesen Bereichen haben, einen zugänglichen und leicht verständlichen Einstieg zu bieten. Durch die Erklärung, was ML und AutoML sind und welche revolutionären Möglichkeiten sie bieten, wird ein Fundament geschaffen, auf dem Interessierte aufbauen können, um tiefer in das Feld einzutauchen. Zur Unterstützung des Lernprozesses und zur Förderung des Verständnisses werden in diesem Tab künstlich generierte Bilder eingesetzt, die mit Technologien wie ChatGPT und DALL·E erstellt wurden. Die Verwendung von Bildern dient nicht nur der ästhetischen Bereicherung, sondern auch der Veranschaulichung abstrakter Ideen und der Förderung eines intuitiven Verständnisses der Materie.

3.2.4 Train Your Model Tab

Der **Train Your Model Tab** (siehe 3.4) stellt die Hauptkomponente der Webseite dar und ist speziell darauf ausgelegt, den Nutzern einen praktischen Einstieg in das Training von ML-Modellen mittels AutoML zu ermöglichen, ohne dass sie eine einzige Zeile Code schreiben müssen. Dieser interaktive

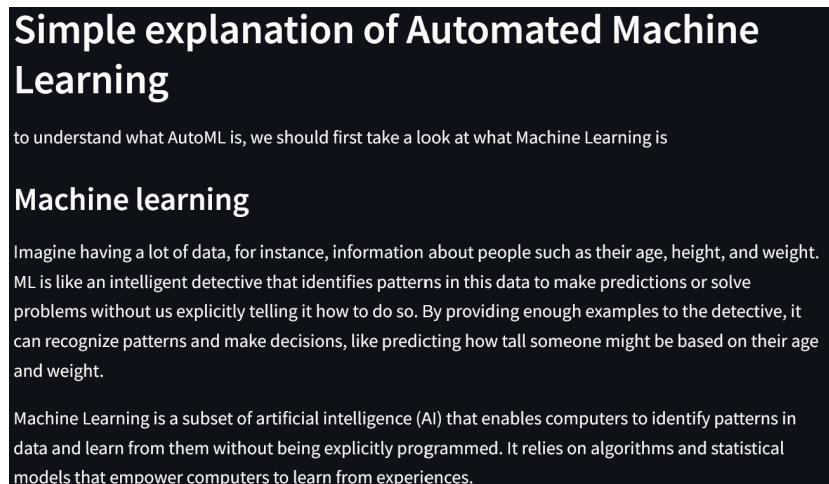


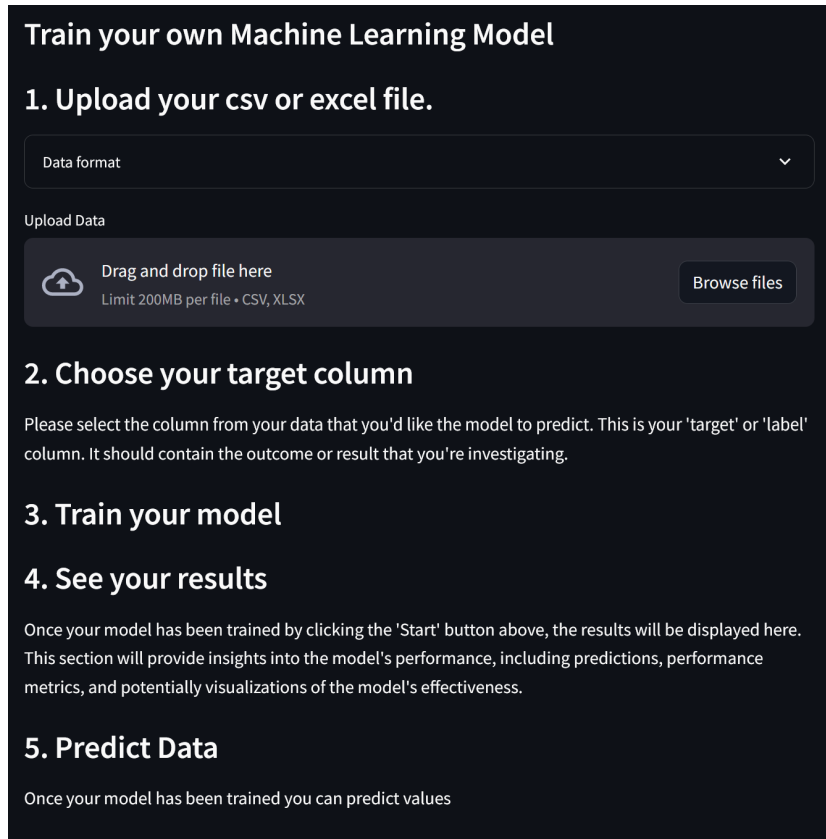
Figure 3.3: Automated Machine Learning tab

Prozess beginnt mit der Möglichkeit, über den `streamlit.file_uploader` eine CSV- oder XLSX-Datei hochzuladen. Durch bedingtes Rendering, basierend auf einfacher boolescher Logik, werden die weiteren Schritte des Prozesses nach und nach offenbart, wodurch die Nutzer schrittweise und intuitiv durch den gesamten Prozess geführt werden.

Im nächsten Schritt werden die Nutzer aufgefordert, mittels der `streamlit.selectbox` die Spalte auszuwählen, die sie vorhersagen möchten. Dies fördert das Verständnis für die Zielvariable und die Art der Vorhersage, die durchgeführt werden soll. Nach der Auswahl der zu prognostizierenden Spalte ermöglicht ein `streamlit.button` den Start des Trainingsprozesses.

Nach Abschluss des Trainings werden die Ergebnisse und eine Erklärung zu allen Bewertungsmetriken präsentiert. Sollten die Ergebnisse hinter den Erwartungen zurückbleiben, werden mögliche Ursachen dafür aufgezeigt und dem Nutzer Hilfestellungen für Verbesserungen gegeben.

Im letzten Schritt bietet der Tab die Möglichkeit, einen eigenen Datensatz hochzuladen, der vorhergesagt werden soll. Die Ergebnisse werden anschließend in Form einer Tabelle mit der `streamlit.table` Methode dargestellt. Nutzer haben die Möglichkeit, ihre Ergebnisse über den `streamlit.download_button` als XLSX-Datei herunterzuladen, was einen direkten Mehrwert schafft und die praktische Anwendung der gelernten Inhalte fördert.



Train your own Machine Learning Model

1. Upload your csv or excel file.

Data format ▾

Upload Data

Drag and drop file here
Limit 200MB per file • CSV, XLSX

Browse files

2. Choose your target column

Please select the column from your data that you'd like the model to predict. This is your 'target' or 'label' column. It should contain the outcome or result that you're investigating.

3. Train your model

4. See your results

Once your model has been trained by clicking the 'Start' button above, the results will be displayed here. This section will provide insights into the model's performance, including predictions, performance metrics, and potentially visualizations of the model's effectiveness.

5. Predict Data

Once your model has been trained you can predict values

Figure 3.4: Train Your Model Tab

3.2.5 About Me Tab

Der abschließende Bereich der Webseite bietet persönliche Informationen über den Entwickler (siehe 3.5). Neben einer kurzen Vorstellung sind verschiedene Kontaktmöglichkeiten angegeben, die es Interessierten erlauben, in direkten Austausch mit dem Entwickler zu treten. Dieser persönliche Ansatz fördert die Interaktion und ermöglicht den Nutzern, Feedback oder Fragen einfach zu kommunizieren.

In dieser Sektion wurde der strukturierte Aufbau und die inhaltliche Ausgestaltung der einzelnen Tabs der entwickelten Webanwendung ausführlich beschrieben. Jeder Tab erfüllt eine spezifische Funktion innerhalb des Gesamtkonzepts der Anwendung, die darauf abzielt, die Prinzipien und die Anwendung von Automated Machine Learning (AutoML) einer breiteren Nutzergruppe zugänglich zu machen. Von der Einführung in die Grundlagen des maschinellen Lernens und AutoML, über die spezifische Anwendung auf Polymere, bis hin zum praktischen Training von Modellen, vermittelt die

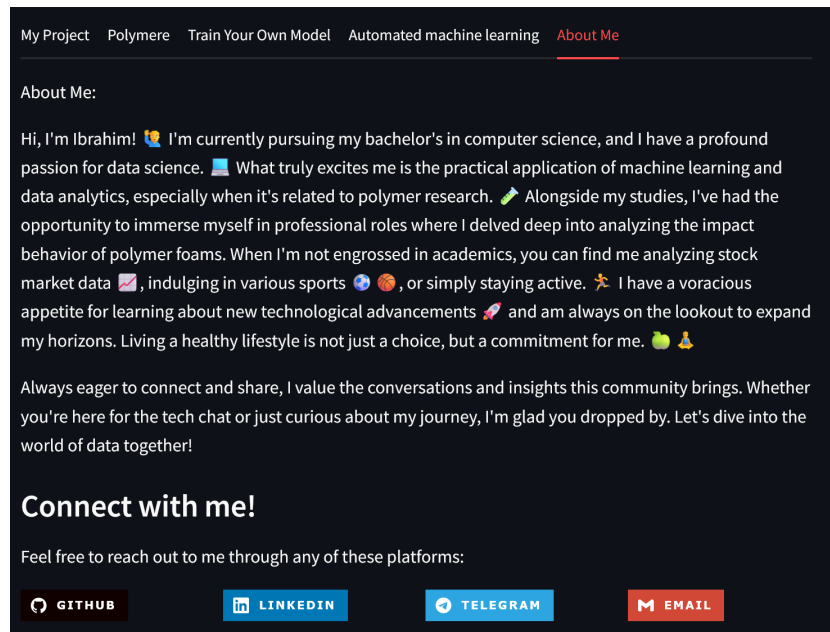


Figure 3.5: About Me Tab

Webanwendung ein umfassendes Verständnis für die Materie und ermöglicht es auch Nutzern ohne tiefgehende technische Vorkenntnisse, aktiv mit ML-Modellen zu arbeiten und Vorhersagen zu treffen.

Das Hauptziel dieser interaktiven Lernumgebung ist es, die Barriere für den Einstieg in das maschinelle Lernen zu senken und die Anwendung von AutoML-Techniken zu demokratisieren. Durch die Bereitstellung einer intuitiven und benutzerfreundlichen Oberfläche, die den Nutzern schrittweise Anleitungen und direktes Feedback gibt, trägt die Anwendung dazu bei, das Interesse und das Verständnis für AutoML zu fördern. Die Implementierung von Features wie Datei-Uploads, Auswahlmöglichkeiten über Dropdown-Menüs, Training und Evaluation von Modellen sowie die Möglichkeit, Ergebnisse herunterzuladen und mit eigenen Daten zu experimentieren, unterstreicht den praktischen Wert der Anwendung.

Abschließend lässt sich festhalten, dass die Gestaltung der Tabs und die Integration verschiedener Streamlit-Funktionen nicht nur die technische Umsetzung eines AutoML-Tools darstellen, sondern auch einen didaktischen Ansatz verfolgen. Dieser Ansatz ermöglicht es den Nutzern, die Potenziale und Herausforderungen von AutoML durch aktive Teilnahme zu erkunden, was die Webanwendung zu einem wertvollen Instrument für Bildungszwecke im Bereich des maschinellen Lernens macht.

Chapter 4

Durchführung der Nutzerstudie

Die User-Studie zielte darauf ab, die Benutzerfreundlichkeit und Zugänglichkeit des entwickelten Webtools für eine spezifische Zielgruppe zu evaluieren: Personen, die mit Datensätzen arbeiten, jedoch wenig bis gar keine Erfahrung im Bereich des maschinellen Lernens (ML) haben. Basierend auf den Empfehlungen von Nielsen Jakobs wurden für den Prozess fünf Personen ausgewählt, da frühere Forschungen gezeigt haben, dass das Testen mit nicht mehr als fünf Benutzern die effektivste Methode ist, um Usability-Probleme effizient zu identifizieren und die User Experience zu verbessern [43].

Das primäre Ziel der Studie war es, die Zugänglichkeit des Tools für die Zielgruppe zu verstehen und potenzielle Usability-Probleme zu identifizieren. Dies umfasste die Überprüfung, ob die Nutzer in der Lage waren, ML-Modelle selbstständig zu trainieren und Vorhersagen zu treffen, und ob die Konzepte des maschinellen Lernens und des automatisierten maschinellen Lernens (AutoML) klar vermittelt wurden. Für die Studie wurde die Think-Aloud-Methode in Kombination mit direkten Beobachtungen und Interviews eingesetzt, um Einblicke in die Gedankenprozesse der Nutzer zu erhalten.

4.1 User Tasks

In der durchgeführten Studie wurden die Teilnehmer mit sechs spezifischen Aufgaben konfrontiert, die darauf ausgelegt waren, verschiedene Aspekte des Webtools und dessen Benutzerführung zu testen. Um tiefergehende Einblicke in die Benutzererfahrung zu gewinnen und spezifische Herausforderungen sowie Problembereiche zu identifizieren, wurde die Think-Aloud-Methode

angewandt. Dieser Ansatz ermöglicht es, wertvolle Insights darüber zu sammeln, an welchen Stellen die Nutzer auf Hindernisse stoßen, welche Funktionen möglicherweise nicht intuitiv genug gestaltet sind oder wo zusätzliche Anleitungen und Informationen erforderlich sein könnten, um die Benutzerführung zu optimieren.

Die Studienteilnehmer wurden zu Beginn des Interviews sorgfältig eingeführt, wobei der Fokus darauf lag, den Zweck des Interviews klar zu kommunizieren und die Erwartungen an die Teilnehmer zu definieren. Die Einführung diente dazu, den Teilnehmern einen umfassenden Überblick über das Forschungsprojekt zu geben, die Ziele der Studie zu erläutern und die Bedeutung ihrer Beiträge hervorzuheben. Es wurde betont, dass ihre Einsichten und Erfahrungen während des Tests des Webtools von entscheidender Bedeutung sind, um die Benutzerfreundlichkeit und Funktionalität der Anwendung zu bewerten und zu verbessern.

In dieser Phase wurde den Teilnehmern auch erklärt, was von ihnen erwartet wird: Sie sollten das Webtool nutzen, während sie ihre Gedanken und Aktionen laut äußern, gemäß der Think-Aloud-Methode. Diese Methode wurde gewählt, um ein tieferes Verständnis dafür zu gewinnen, wie die Nutzer mit der Anwendung interagieren, welche Schritte für sie intuitiv sind und an welchen Punkten sie möglicherweise Schwierigkeiten haben oder Unsicherheiten auftreten. Den Teilnehmern wurde versichert, dass es keine "richtigen" oder "falschen" Aktionen gibt und dass ihre offenen und ehrlichen Feedbacks für die Studie von großem Wert sind.

Verständnis von ML und AutoML: Die erste Aufgabe im Rahmen des Interviews konzentrierte sich darauf, das grundlegende Verständnis der Nutzer bezüglich Machine Learning (ML) und Automated Machine Learning (AutoML) zu evaluieren. Zu Beginn wurden die Teilnehmer gebeten, den AutoML-Tab sorgfältig durchzulesen, um sich mit den Konzepten und der allgemeinen Funktionsweise von ML und AutoML vertraut zu machen. Diese einführende Aufgabe diente nicht nur dazu, das Interesse und die Motivation der Nutzer zu wecken, sondern auch als Grundlage, um zu überprüfen, inwieweit Personen ohne Vorerfahrung im Bereich des maschinellen Lernens die dargebotenen Informationen verstehen und wiedergeben können.

Anschließend wurden die Teilnehmer aufgefordert, in eigenen Worten zu erklären, was sie unter ML und AutoML verstehen. Dieser Schritt hatte zum Ziel, Einblicke in das intuitive Verständnis der Konzepte zu gewinnen und festzustellen, wie effektiv der Tab die Kernideen vermittelt. Insbesondere wurde darauf geachtet, ob die Nutzer die Automatisierung von Prozessen innerhalb des ML-Lebenszyklus durch AutoML und dessen Vorteile für An-

wender ohne technischen Hintergrund erfassen konnten.

Training mit dem Boston Housing Dataset: Die Aufgabe rund um das Training mit dem Boston Housing Dataset war sorgfältig darauf ausgerichtet, die Intuitivität der Benutzerführung durch den Trainingsprozess im Webtool zu bewerten. Konkret sollten die Nutzer den Datensatz innerhalb der Anwendung nutzen, um ein Modell zu trainieren, wobei die Spalte "Tax" als Vorhersageziel ausgewählt werden musste. Diese spezifische Auswahl diente nicht nur dazu, die Funktionalität des Tools unter realen Bedingungen zu testen, sondern auch um zu verstehen, wie Nutzer mit der Aufgabenstellung umgehen, insbesondere in Bezug auf die Identifizierung und Auswahl der Zielvariable für die Vorhersage. Der Datensatz wurde dabei auf die Hälfte der Datenpunkte reduziert, um ein ungenaues Modell zu erhalten. Der verwendete Datensatz ist frei verfügbar und kann unter [36] gefunden werden.

Diese Phase des Interviews bot tiefe Einblicke in die Benutzererfahrung, insbesondere hinsichtlich der Aspekte Navigation, Verständlichkeit der Interface-Elemente und der generellen Benutzerführung durch den Trainingsprozess. Von besonderem Interesse war, wie Nutzer die Interaktion mit dem Tool wahrnahmen, wenn es darum ging, aus einer Liste von Spalten die richtige Zielvariable auszuwählen. Herausforderungen oder Unsicherheiten bei diesem Schritt konnten wichtige Hinweise auf Verbesserungsmöglichkeiten in der Gestaltung der Benutzeroberfläche oder der bereitgestellten Anleitungen liefern.

Evaluation der Ergebnisse: Die Aufgabe zur Evaluation der Ergebnisse stellte für die Nutzer eine besondere Herausforderung dar. Sie zielte darauf ab, die Qualität eines trainierten Modells zu beurteilen, indem sie die vorgelegten Ergebnisse und die zugehörigen Bewertungsmetriken – Mean Squared Error (MSE), Mean Absolute Error (MAE) und das Bestimmtheitsmaß (R^2) – analysierten. Nach Abschluss des Trainingsprozesses stellte das Webtool den Nutzern detaillierte Informationen und Erklärungen zu diesen Metriken zur Verfügung, um ihnen die Beurteilung zu erleichtern, ob die Ergebnisse ihres Modells als gut oder schlecht zu bewerten waren.

Die Nutzer wurden dazu aufgefordert, nicht nur die Qualität der Modellergebnisse zu bewerten, sondern auch zu reflektieren und zu begründen, warum die Ergebnisse so ausgefallen waren. Durch das Webtool wurden zusätzlich mögliche Gründe für schlecht trainierte Modelle genannt, um die Nutzer in ihrer Analyse zu unterstützen. Der Datensatz für diese Aufgabe wurde bewusst so gewählt, dass tendenziell schlechte Ergebnisse erzielt wurden. Dies hatte das Ziel, herauszufinden, ob die Nutzer in der Lage sind, ein suboptimales Modellergebnis zu erkennen, und ob sie die zugrundeliegenden Gründe

für diese Leistung identifizieren können.

Training mit dem Polymer Properties Dataset: Die Aufgabe des Trainings mit dem Polymer Properties Dataset war konzeptionell ähnlich zur vorherigen Herausforderung gestaltet, allerdings mit einem entscheidenden Unterschied: Diesmal wurde den Nutzern ein qualitativ hochwertiger Datensatz zur Verfügung gestellt. Ziel dieser Aufgabe war es, zu beobachten, ob und wie die Nutzer die Güte der Ergebnisse anhand der bereitgestellten Bewertungsmetriken – einschließlich Mean Squared Error (MSE), Mean Absolute Error (MAE) und dem Bestimmtheitsmaß (R^2) – beurteilen konnten, insbesondere im Kontext eines Datensatzes, der zu besseren Modellergebnissen führt.

Nach dem Training des Modells mit dem Polymer Properties Dataset wurden den Nutzern, ähnlich wie in der vorherigen Aufgabe, Informationen zu den Bewertungsmetriken präsentiert. Zusätzlich erhielten sie Erklärungen zu diesen Metriken, um die Beurteilung der Modellergebnisse zu erleichtern. Im Gegensatz zur vorherigen Aufgabe, die darauf ausgelegt war, die Fähigkeit der Nutzer zu testen, schlechte Ergebnisse zu erkennen und zu analysieren, zielte diese Aufgabe darauf ab, herauszufinden, ob die Nutzer in der Lage waren, gute Modellergebnisse als solche zu identifizieren und die Qualität des Modells basierend auf einem qualitativ hochwertigen Datensatz positiv zu bewerten.

Vorhersage und Download der Ergebnisse: Bei dieser spezifischen Aufgabe stand der Vorhersageprozess im Zentrum der Betrachtung. Der Fokus lag darauf zu evaluieren, inwieweit die Nutzer die verschiedenen Schritte des Prozesses intuitiv nachvollziehen konnten, vom Einlesen des Datensatzes bis hin zum Download der vorhergesagten Ergebnisse. Diese Phase des Tests diente dazu, die Benutzerfreundlichkeit und Intuitivität des Tools in Bezug auf eine zentrale Funktion des Machine Learning zu überprüfen.

Den Nutzern wurde ein spezifischer Datensatz zur Verfügung gestellt, den sie innerhalb des Webtools verwenden sollten, um Vorhersagen zu generieren. Die Aufgabe umfasste mehrere Schritte: Zunächst mussten die Teilnehmer den Datensatz über eine Upload-Funktion in das Tool einbringen. Anschließend wurden sie durch den Prozess geführt, eine Vorhersage zu erstellen, was die Auswahl relevanter Parameter und die Initiierung des Vorhersageprozesses einschloss. Nach der Generierung der Vorhersagen durch das Tool lag der nächste Schritt in der Bewertung und Interpretation der Ergebnisse durch die Nutzer. Ein entscheidender Aspekt dieser Aufgabe war die Überprüfung, ob und wie die Teilnehmer die generierten Vorhersagen herunterladen konnten. Die Möglichkeit, Ergebnisse in einer handhabbaren Form

zu exportieren, ist ein wesentlicher Bestandteil des praktischen Einsatzes von Machine Learning-Modellen, da sie die Weiterverwendung und Analyse der Daten ermöglicht. Die Download-Funktion wurde somit als kritischer Punkt für die Benutzerfreundlichkeit des Tools betrachtet.

In der durchgeführten Studie wurden die Teilnehmer mit sechs spezifischen Aufgaben konfrontiert, die sorgfältig darauf ausgerichtet waren, verschiedene Aspekte des Webtools und dessen Benutzerführung umfassend zu testen. Die Auswahl und Gestaltung der Aufgaben wurden mit der Absicht vorgenommen, ein breites Spektrum an Nutzungsszenarien abzudecken, von der Einführung in die Grundkonzepte von ML und AutoML, über den Prozess des Trainierens von Modellen mit unterschiedlich beschaffenen Datensätzen, bis hin zur Evaluation der Modellergebnisse und dem Download der vorhergesagten Daten. Diese umfassende Herangehensweise zielte darauf ab, alle möglichen Anwendungsfälle innerhalb des Webtools zu beleuchten, um so ein ganzheitliches Verständnis für die Stärken und Schwachstellen der Anwendung zu erlangen.

Die bewusste Entscheidung, ein breites Spektrum an Aufgaben zu integrieren, hatte zum Ziel, eine möglichst vollständige Einsicht in die Benutzererfahrung zu erlangen. Dadurch konnte nicht nur festgestellt werden, wie Anwender mit bekannten oder erwarteten Situationen umgehen, sondern es bot auch die Möglichkeit, zu beobachten, wie sie auf unerwartete Herausforderungen oder weniger intuitive Aspekte des Tools reagieren. Die Erkenntnisse aus dieser umfangreichen Evaluierung liefern entscheidende Ansatzpunkte für die Weiterentwicklung des Webtools, um es noch benutzerfreundlicher zu gestalten und die Zugänglichkeit von Machine Learning und AutoML für ein breiteres Publikum zu fördern.

4.2 Semi-Strukturierte Online-Befragung

Nach Abschluss der Aufgaben wurden die Nutzer gebeten, an einer semi-strukturierten Online-Befragung teilzunehmen. Beispiele für die gestellten Fragen sind in den Abbildungen 4.1 und 4.2 zu sehen. Diese Befragung war darauf ausgelegt, detailliertes Feedback zur Benutzererfahrung und zur Benutzbarkeit der Anwendung zu sammeln. Semi-strukturierte Interviews kombinieren vordefinierte Fragen mit der Flexibilität, Antworten ausführlich zu erkunden. Sie folgen standardisierten Themen, erlauben jedoch vielfältige Fragen in einem Mix aus offenen und geschlossenen Formaten. Dieser Ansatz liefert reichhaltige qualitative Daten, während er sich auf die Forschungsziele konzentriert.

Die Fragen basierten auf dem Likert-Format und wurden aus der System Usability Scale (SUS) entnommen, die aus 10 Likert-Items mit jeweils 5 Punkten besteht. Die Erstellung des Fragebogens erfolgte über **Typeform**, eine Plattform, die das einfache Erstellen von benutzerdefinierten Fragen ermöglicht.

Folgende Abbildungen zeigen Beispiele der im Fragebogen gestellten Fragen:

1 → It was easy for you to understand the evaluation results?
Description (optional)
0 1 2 3 4 5
Strongly disagree Neutral Strongly agree

2 → Did you find the system unnecessarily complex?
Description (optional)
0 1 2 3 4 5
Strongly disagree Neutral Strongly agree

3 → Do you think that you would need the support of a technical person to be able to use this system?
Description (optional)
Yes No

4 → Most people would learn to use this system very quickly
Description (optional)
0 1 2 3 4 5
Strongly disagree Neutral Strongly agree

5 → Did you need to learn a lot of things before you could get going with this system?
Description (optional)
Yes No

6 → The various functions were well integrated into the system.
Description (optional)
0 1 2 3 4 5
Strongly disagree Neutral Strongly agree

Figure 4.1: Beispiel einer Frage aus dem semi-strukturierten Fragebogen.

6 → Did you need to learn a lot of things before you could get going with this system?
Description (optional)
Y Yes
N No

7 → Evaluating the results was challenging.
Description (optional)
0 1 2 3 4 5
Strongly disagree Neutral Strongly agree

Figure 4.2: Ein weiteres Beispiel einer Frage aus dem Fragebogen.

Am Ende der Befragung wurde den Nutzern die Möglichkeit gegeben, zusätzliches Feedback zu ihrer Erfahrung mit dem AutoML-Tool zu teilen: *"Gibt es noch etwas, das Sie über Ihre Erfahrung mit unserem AutoML-Tool mitteilen möchten?"* Dies ermöglichte es, wertvolle Einsichten in die Nutzererfahrung zu gewinnen, die über die standardisierten Fragen hinausgingen.

Chapter 5

Ergebnisse

5.1 Leistung der AutoML-Modelle

Einleitung: Die Leistungsfähigkeit von Automated Machine Learning bei der Vorhersage von Bandlücken in Polymeren ist ein zentrales Thema dieser Arbeit. In diesem Kapitel werden die Ergebnisse verschiedener AutoML-Bibliotheken präsentiert und diskutiert, die im Rahmen dieser Studie getestet wurden. Die Leistung der Modelle wird anhand mehrerer Metriken bewertet, darunter Mean Absolute Error (MAE), Mean Squared Error (MSE) und R2-Score. Zusätzlich wird die Laufzeit der Modelle auf einer NVIDIA A100-SXM4-40GB GPU gemessen, die speziell für Hochleistungsrechenaufgaben und KI-Training konzipiert ist.

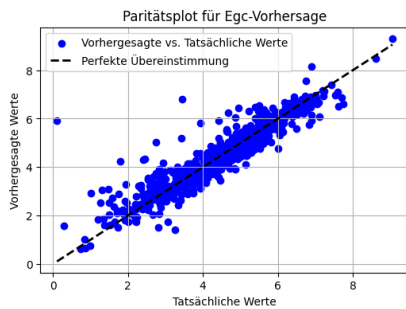
Tabelle: Leistung der AutoML-Modelle

Rank	Library	Runtime (m)	MAE	MSE	R2
1	AutoGluon	3	0.3456	0.2912	0.8773
2	Auto-Sklearn	5	0.4208	0.3956	0.8386
3	Autokeras	180	0.4716	0.4750	0.7996
4	Pycaret	40	0.5273	0.5178	0.7790
5	Ludwig	120	0.7149	1.0594	0.5692

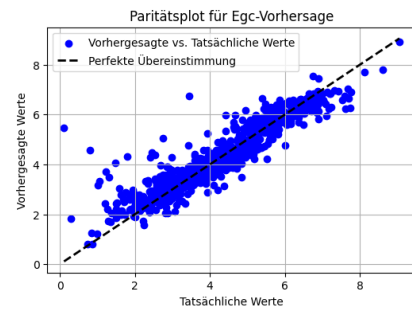
Table 5.1: Leistung der AutoML-Modelle

Die Ergebnisse zeigen, dass AutoGluon die beste Performance erzielt hat, sowohl in Bezug auf die Vorhersagegenauigkeit als auch auf die Laufzeit, wie auch im Paritätsplot deutlich wird (siehe Abbildung 5.1). Es war mit Abstand der schnellste und erreichte die niedrigsten Fehlerwerte. Besonders

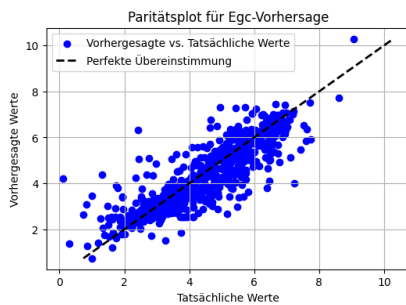
hervorzuheben ist die Benutzerfreundlichkeit von AutoGluon, da es keine komplizierten Konfigurationsschritte erforderte und problemlos verwendet werden konnte. Im Gegensatz dazu benötigten Bibliotheken wie Autokeras und Ludwig deutlich mehr Zeit und hatten häufiger Probleme mit der GPU-Nutzung. Diese Ergebnisse deuten darauf hin, dass die Effizienz und Benutzerfreundlichkeit von AutoML-Bibliotheken entscheidende Faktoren für ihren Einsatz in der Praxis sind, insbesondere für Anfänger ohne Zugang zu leistungsstarken Rechenressourcen.



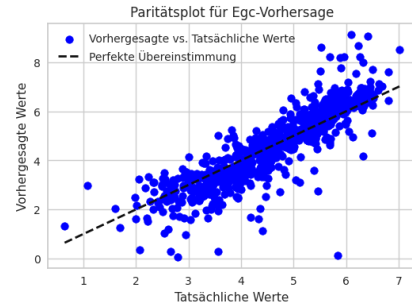
(a) AutoGluon



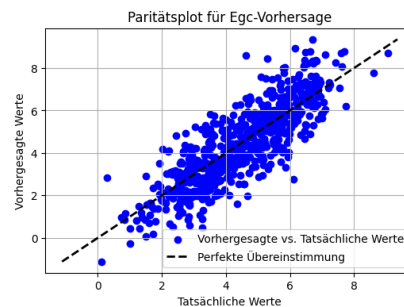
(b) Auto-Sklearn



(c) Autokeras



(d) Pycaret



(e) Ludwig

Figure 5.1: Paritätsplots der AutoML-Modelle für die Vorhersage der elektrischen Bandbreite

5.2 Ergebnisse der Nutzerstudie

Die Durchführung der Nutzerstudie ergab, dass die Nutzertasks erfolgreich abgelaufen sind und die Think-Aloud-Methode wertvolle Einblicke in die Nutzererfahrung lieferte. Dadurch konnte genau erfasst werden, wo die Nutzer Schwierigkeiten hatten und wo nicht.

Bei der ersten Aufgabe, dem Lesen und Verständnis der Erklärungen zu ML und AutoML auf der Webseite, wurde festgestellt, dass alle Nutzer nach dem Lesen ein gutes Verständnis davon hatten. Jedoch hatten alle Nutzer Schwierigkeiten mit bestimmten Fachbegriffen wie Hyperparameter-Optimierung, Feature Engineering und Modellensembles. Als Lösungsvorschlag wurde empfohlen, diese Begriffe genauer zu erklären oder sogar zu entfernen.

Beim Training mit dem Boston Housing Dataset zeigten alle Nutzer keine Probleme und bestanden die Aufgabe ohne Komplikationen. Es gab jedoch bei einigen Nutzern Verwirrung, da sie nicht verstanden haben, ob der Trainingsprozess stattgefunden hat.

Die Evaluation der Ergebnisse zeigte, dass einige Nutzer Schwierigkeiten hatten, die Probleme des Datensatzes zu identifizieren, warum das Training keine guten Ergebnisse liefern konnte, selbst nachdem sie die möglichen Probleme über Datensätze auf der Webseite gelesen hatten. Es wurde festgestellt, dass nur weil die Nutzer einen Datensatz hochladen, nicht deren Eigenschaften kennen, speziell in Hinblick auf die wichtigen Eigenschaften im Machine Learning. Als Lösung wurde vorgeschlagen, zusätzliche Informationen über den Datensatz nach dem Upload anzuzeigen, um den Nutzern zu helfen, die Ergebnisse besser zu verstehen.

Beim Training mit dem Polymer Properties Dataset hatten alle Nutzer keine Probleme und konnten die Schritte problemlos ausführen. Die Erläuterungen zu den Leistungsmaßen halfen den Nutzern, die Ergebnisse zu verstehen und zu bewerten.

Bei der letzten Aufgabe, der Vorhersage und dem Download der Ergebnisse, konnten alle Nutzer die Möglichkeit zum Hochladen eines zusätzlichen Datensatzes erfolgreich und ohne Probleme nutzen. Die alternative Methode, die Eigenschaften des Datensatzes in ein Eingabefeld einzugeben, erwies sich als problematisch, da es zu vielen Missverständnissen kam. Die Nutzer empfahlen, dieses Feature auszulassen und stattdessen das Hochladen eines Datensatzes für Vorhersagen zu nutzen.

5.3 Evaluation der Semi-Strukturierten Online-Befragung

Likert-Skala-Fragen:

- *It was easy for you to understand the evaluation results based on the potential issues with categorical data.*

- **Strongly disagree:** 0%
- **Disagree:** 40%
- **Neutral:** 20%
- **Agree:** 40%

Hier zeigt sich, dass einige Nutzer Schwierigkeiten hatten, die Ergebnisse zu verstehen.

- *Did you find the system unnecessarily complex?*

- **Strongly disagree:** 20%
- **Disagree:** 80%
- **Neutral:** 0%
- **Agree:** 0%
- **Strongly agree:** 0%

Hier kann man erkennen, dass das System im Allgemeinen als einfach zu benutzen empfunden wurde.

- *The various functions were well integrated into the system.*

- **Strongly disagree:** 0%
- **Disagree:** 0%
- **Neutral:** 0%
- **Agree:** 20%
- **Strongly agree:** 80%

Hier zeigt sich ebenfalls, dass die Nutzer gut mit dem System zurechtkamen.

- *Most people would learn to use this system very quickly.*

- **Strongly disagree:** 0%
- **Disagree:** 0%
- **Neutral:** 0%
- **Agree:** 60%
- **Strongly agree:** 40%

Auch hier zeigt sich, dass das System nicht komplex ist und leicht zu verstehen ist, insbesondere für Anfänger.

- *Evaluating the results based on the error metrics was challenging.*

- **Strongly disagree:** 20%
- **Disagree:** 40%
- **Neutral:** 20%
- **Agree:** 20%
- **Strongly agree:** 0%

Hier zeigt sich, dass die Erklärungen über die Fehlermetriken helfen konnten zu erkennen, wie gut das System funktioniert.

Ja/Nein-Fragen:

- *Did you need to learn a lot of things before you could get going with this system?*

- **Ja:** 20%
- **Nein:** 80%

Dies unterstreicht, dass das System auch ohne Vorkenntnisse gut zu benutzen ist, besonders für ML-Anfänger.

- *Do you think that you would need the support of a technical person to be able to use this system?*

- **Nein:** 100%

Dies unterstreicht erneut die Benutzerfreundlichkeit des Systems.

Is there anything else you would like to share about your experience using our AutoML tool?

1. *Kommentar 1:* “I liked your section, where you pointed at common problems why my model performed badly (e.g. you dont have enough data, ...) It would be nice to also have some information about the input data (e.g. the number of rows, how your data looks, like a preview) Also it wasn't so clear, that the model was training, since the "start training" button was still enabled. It could be nice to make it more clear that the model is training (disable the button or add a progress bar)”
2. *Kommentar 2:* “A Progressbar while the data is being trained would be good to have. Make it more clear how you can see what's wrong with the dataset because of bad training results.”
3. *Kommentar 3:* “Evaluation was hard. Some kind of progress bar would nice. And the Preview was confusing”
4. *Kommentar 4:* “Simplify User Interface a bit and make the overall appearance more intuitive (e.g. where do i find the meaning of `r2`, `mean_error_squared`, etc.)”

Durch die erhaltenen Rückmeldungen konnte ich die Herausforderungen genauer identifizieren. Es gestaltet sich schwierig zu ergründen, warum ein bestimmter Datensatz nicht wie erwartet funktioniert. Zudem ist der Prozess des Modelltrainings nicht immer transparent für den Nutzer, was zu Verwirrung führen kann. Darüber hinaus wurde betont, wie wichtig es ist, den hochgeladenen Datensatz und seine Eigenschaften ausführlich zu erläutern, um dem Nutzer ein besseres Verständnis zu ermöglichen.

Zusammenfassung

Die Ergebnisse dieses Kapitels bieten einen umfassenden Einblick in die Leistung der AutoML-Modelle sowie die Nutzererfahrung mit dem System. AutoGluon erwies sich als führende Bibliothek mit der besten Vorhersagegenauigkeit und Benutzerfreundlichkeit. Die Nutzerstudie zeigte, dass die Think-Aloud-Methode wertvolle Einblicke lieferte, aber auch Herausforderungen bei der Erklärung von ML-Konzepten und der Identifizierung von Problemen mit Datensätzen aufzeigte. Die Likert-Skala- und Ja/Nein-Fragen deuteten darauf hin, dass das System im Allgemeinen als benutzerfreundlich wahrgenommen wurde, aber Verbesserungspotenzial in Bezug auf die Transparenz des Modelltrainings und die Bereitstellung von Informa-

tionen über die Eingabedaten aufwies. Die Kommentare der Nutzer lieferten wertvolle Anregungen zur Verbesserung des Systems, insbesondere hinsichtlich einer klareren Darstellung des Trainingsfortschritts und einer vereinfachten Benutzeroberfläche. Insgesamt bieten die Ergebnisse dieses Kapitels einen umfassenden Überblick über die Leistung und Nutzererfahrung mit dem AutoML-Tool.

Chapter 6

Diskussionen

In der vorliegenden Bachelorarbeit wurde die Anwendung von Automated Machine Learning (AutoML) zur Vorhersage der Bandlücken-Eigenschaften von Polymeren untersucht, wobei ein besonderes Augenmerk auf die Benutzerfreundlichkeit und Effizienz der AutoML-Tools gelegt wurde. Insbesondere konnte durch den Einsatz von AutoGluon ein bemerkenswertes Modell mit einem R^2 -Wert von 0.8773 erstellt werden, dessen Trainingszeit mit nur 3 Minuten die Kernanforderungen von AutoML erfüllt: Schnelligkeit, Einfachheit und Zugänglichkeit auch für Anfänger im Bereich des maschinellen Lernens. Die Untersuchung hat gezeigt, dass ML-Bibliotheken und Ensemble-Methoden im Vergleich zu Deep-Learning-Ansätzen deutliche Vorteile bieten. Während mit Ensemble-Methoden effiziente und präzise Modelle generiert werden konnten, waren Deep-Learning-Bibliotheken wie AutoKeras und Ludwig nicht nur fehleranfälliger und somit für Programmieranfänger weniger zugänglich, sondern benötigten auch erheblich längere Trainingszeiten, die ohne entsprechend leistungsfähige Hardware kaum realisierbar sind. Die Analyse und Anwendung von Deep-Learning-Methoden im Kontext von Automated Machine Learning (AutoML) offenbarten signifikante Herausforderungen, die eng mit den derzeitigen technologischen Grenzen und der Zugänglichkeit für Anwender ohne spezialisierte Hardware verbunden sind. Insbesondere wurden AutoKeras und Ludwig, beides Deep-Learning-Bibliotheken, im Rahmen dieser Arbeit untersucht. Obwohl diese Tools das Versprechen bieten, komplexe neuronale Netzwerkmodelle automatisch zu generieren und zu optimieren, zeigte sich, dass der damit verbundene Rechenaufwand und die langen Trainingszeiten erhebliche Hindernisse darstellen.

Der Kern dieser Problematik liegt in der Natur neuronaler Netzwerke

selbst. Um präzise Vorhersagen zu treffen, müssen neuronale Netzwerke auf umfangreichen Daten trainiert werden. Dieser Trainingsprozess erfordert beträchtliche Rechenressourcen, insbesondere Grafikprozessoren (GPUs), die in der Lage sind, komplexe mathematische Berechnungen schnell durchzuführen. Während traditionelle maschinelle Lernmethoden und Ensemble-Modelle auf Standard-CPUs effizient trainiert werden können, benötigen Deep-Learning-Modelle zur Beschleunigung ihres Trainingsprozesses spezialisierte Hardware in Form von hochleistungsfähigen GPUs. Dies macht den Einsatz von Deep-Learning-Methoden für Anwender ohne Zugang zu derartiger Hardware nicht nur schwierig, sondern in vielen Fällen praktisch unmöglich. Darüber hinaus führen die Komplexität und die zahlreichen Hyperparameter, die für das Training neuronaler Netzwerke erforderlich sind, zu weiteren Schwierigkeiten. Obwohl AutoML-Tools wie AutoKeras und Ludwig darauf abzielen, diesen Prozess zu vereinfachen, indem sie die Auswahl und Optimierung der Hyperparameter automatisieren, bleibt der Zeitaufwand für das Training und die Feinabstimmung der Modelle erheblich. Beispielsweise resultierten in dieser Studie die Trainingszeiten von 180 Minuten für AutoKeras und 120 Minuten für Ludwig, ohne dass die erzielten Ergebnisse die erhebliche Investition in Zeit und Ressourcen rechtfertigen konnten.

Ein weiterer wesentlicher Aspekt dieser Arbeit war die Entwicklung und Evaluierung einer Web-Anwendung basierend auf AutoGluon, welche es Nutzern ermöglicht, ohne Programmierkenntnisse ML-Modelle zu trainieren und anzuwenden. Nutzerstudien offenbarten sowohl Stärken als auch Schwachstellen der Anwendung. Positiv hervorgehoben wurde die einfache und verständliche Handhabung. Basierend auf dem Feedback der Nutzer wurden jedoch Anpassungen vorgenommen, um die Benutzerfreundlichkeit weiter zu verbessern. Dazu gehörten unter anderem eine klarere Kommunikation des Trainingsstatus des Modells, die Vereinfachung oder Eliminierung komplexer Fachbegriffe sowie die Bereitstellung von Informationen zum hochgeladenen Datensatz, um Nutzern eine bessere Evaluierung der Modellergebnisse zu ermöglichen. Zusammenfassend hat diese Arbeit nicht nur die Potenziale und Herausforderungen von AutoML für die Vorhersage spezifischer Materialeigenschaften aufgezeigt, sondern auch einen praktischen Beitrag zur Demokratisierung des maschinellen Lernens geleistet, indem sie einen Zugang für Anwender ohne tiefgreifende technische Kenntnisse bietet. Die Ergebnisse unterstreichen die Bedeutung weiterer Forschung und Entwicklung, insbesondere im Bereich der Rechenleistung und der Automatisierungstechniken, um die Effizienz und Anwendbarkeit neuronaler Netzwerke zu verbessern und die Reichweite von AutoML-Tools weiter auszubauen.

Chapter 7

Schlussfolgerungen und Ausblick

Schlussfolgerungen

Abschließend lässt sich festhalten, dass die Anwendung von AutoML-Techniken in der vorliegenden Arbeit erheblich dazu beigetragen hat, die Vorhersage von Bandlücken-Eigenschaften zu verbessern und dabei die Kernanforderungen von AutoML – Einfachheit, Geschwindigkeit und Zugänglichkeit auch für ML-Anfänger – erfolgreich zu erfüllen. Mit AutoML konnte ein präzises Modell mit einem R^2 -Wert von 0.8773 und einer bemerkenswert kurzen Laufzeit von nur 3 Minuten entwickelt werden. Dies bestätigt die Effizienz von AutoML-Lösungen im Umgang mit komplexen datenwissenschaftlichen Herausforderungen.

Die Erkenntnisse dieser Arbeit verdeutlichen, dass Ensemble-Methoden und traditionelle ML-Bibliotheken im Vergleich zu spezialisierten AutoML-Bibliotheken für Deep Learning deutlich bessere Ergebnisse erzielen. Während Deep Learning zweifellos ein mächtiges Instrument im Arsenal der Datenwissenschaft darstellt, offenbarte die Untersuchung die Herausforderungen und Einschränkungen beim automatisierten Training neuronaler Netze. Insbesondere der hohe Rechenaufwand und die Notwendigkeit leistungsfähiger Hardware stellen signifikante Hürden dar, die die Zugänglichkeit und Praktikabilität von Deep Learning mittels AutoML einschränken. Diese Beobachtung unterstreicht die Notwendigkeit, dass stärkere GPUs und leistungsfähigere Rechenressourcen im Massenmarkt zum Standard werden müssen, um Machine Learning und insbesondere Deep Learning einer breiteren Nutzerschaft zugänglich zu machen. Die Fortschritte in der Hardwaretechnologie und deren Verbreitung werden entscheidend sein, um die Au-

tomatisierung und Anwendung von Deep-Learning-Modellen in der Zukunft zu vereinfachen und effektiver zu gestalten.

Die Entwicklung eines webbasierten Tools auf Basis von Autogluon als Teil dieser Arbeit hat zusätzlich die Zugänglichkeit von Machine Learning für Anwender ohne Programmierkenntnisse revolutioniert. Durch Nutzerstudien und das Feedback der Anwender konnten wichtige Aspekte der Benutzerfreundlichkeit identifiziert und verbessert werden, was zu einer intuitiveren und zugänglicheren Plattform führte. Die Implementierung von Verbesserungen basierend auf Nutzerrückmeldungen – wie die Verdeutlichung des Trainingsstatus des Modells und die Vereinfachung fachspezifischer Termini – hat wesentlich dazu beigetragen, den Prozess des Machine Learnings weiter zu demokratisieren.

Diese Arbeit hat nicht nur die Machbarkeit und Effizienz von AutoML-Tools zur Vorhersage physikalischer Eigenschaften aufgezeigt, sondern auch wertvolle Erkenntnisse über die Grenzen und Herausforderungen der aktuellen Technologie geliefert. Die Weiterentwicklung der Hardware, die Optimierung von AutoML-Algorithmen für Deep Learning und die fortlaufende Verbesserung der Benutzererfahrung werden Schlüsselrollen spielen, um die Leistungsfähigkeit und Zugänglichkeit von Machine Learning weiter voranzutreiben. Die Zukunft verspricht eine noch engere Verschmelzung von fortschrittlichen ML-Techniken und benutzerfreundlichen Anwendungen, die das Potenzial von Machine Learning voll ausschöpfen und es einer noch breiteren Öffentlichkeit zugänglich machen.

Ausblick

Die Weiterentwicklung von AutoML-Techniken steht im Mittelpunkt zukünftiger Forschungsarbeiten, mit dem Ziel, die Algorithmen zu verfeinern und noch präzisere Vorhersagemodelle zu ermöglichen. Besonders in der Vorhersage komplexer physikalischer Eigenschaften, wie den Bandlücken, ist die Genauigkeit entscheidend. Parallel dazu ist die Hardware-Innovation unerlässlich, da stärkere GPUs und leistungsfähigere Rechenressourcen die Effizienz von AutoML-Prozessen, insbesondere im Deep Learning, verbessern werden.

Die Verbesserung der Benutzererfahrung durch die Entwicklung intuitiverer und zugänglicherer ML-Tools wird weiterhin eine wichtige Rolle spielen. Diese Tools sollten nicht nur einfach zu bedienen sein, sondern auch proaktives Feedback liefern, um den Nutzern zu helfen, die Qualität ihrer Daten-

sätze und Modelle zu verbessern.

Insgesamt wird die Zukunft von AutoML durch eine Kombination aus technologischer Innovation und benutzerzentriertem Design geprägt sein. Diese Entwicklungen versprechen, AutoML zu einem unverzichtbaren Werkzeug in der Datenwissenschaft zu machen, das nicht nur für Experten, sondern auch für Anwender ohne tiefgreifende technische Kenntnisse zugänglich ist. Die fortlaufende Integration von AutoML in verschiedene Anwendungsbereiche wird die Art und Weise, wie wir mit Daten umgehen, revolutionieren und neue Horizonte in der Datenanalyse und -anwendung eröffnen.

References

- [1] 16101305 Abdullah et al.: *Performance Analysis of Intrusion Detection Systems Using the PyCaret Machine Learning Library on the UNSW-NB15 Dataset*. (2022). URL: <https://dspace.bracu.ac.bd/xmlui/handle/10361/15701>.
- [2] Harry R. Allcock, Frederick W. Lampe, and James E. Mark: *Contemporary Polymer Chemistry*. 3rd ed. Pearson Education, (2003), p. 546. ISBN: 978-0-13-065056-6.
- [3] M. Anguera et al.: Indirect Observation in Everyday Contexts: Concepts and Methodological Guidelines within a Mixed Methods Framework. In: *Frontiers in Psychology* 9 (2018). DOI: 10.3389/fpsyg.2018.00013.
- [4] Wolfgang Asselborn, Manfred Jäckel, and Karl T. Risch: *Chemie heute S II Gesamtband*. 2019. Auflage. Schroedel westermann, (2019), 368 ff. ISBN: 978-3-507-10652-9.
- [5] Debra J. Audus and Juan J. de Pablo: Polymer Informatics: Opportunities and Challenges. In: *ACS Macro Letters* 6.10 (2017). PMID: 29201535, pp. 1078–1082. DOI: 10.1021/acsmacrolett.7b00228. eprint: <https://doi.org/10.1021/acsmacrolett.7b00228>. URL: <https://doi.org/10.1021/acsmacrolett.7b00228>.
- [6] A. Author and B. Another: Glossary of basic terms in polymer science. In: *Pure and Applied Chemistry* 68.12 (1996), pp. 2287–2311. DOI: 10.1351/pac199668122287.
- [7] *auto-sklearn*. <https://automl.github.io/auto-sklearn/master/index.html>. Zugriff: Datum des Zugriffs.
- [8] *AutoGluon Tabular Feature Engineering*. <https://auto.gluon.ai/dev/tutorials/tabular/tabular-feature-engineering.html>. Zugriff: Datum des Zugriffs.
- [9] *AutoGluon Tabular: Robust and Accurate Machine Learning with Automated Feature Engineering and Model Tuning*. <https://auto.gluon>.

- ai/dev/tutorials/tabular/tabular-essentials.html. Zugriff: Datum des Zugriffs.
- [10] Aaron Bangor, Philip T. Kortum, and James T. Miller: An Empirical Evaluation of the System Usability Scale. In: *International Journal of Human-Computer Interaction* 24.6 (2008), pp. 574–594. DOI: 10.1080/10447310802205776. URL: <https://doi.org/10.1080/10447310802205776>.
 - [11] Denise Thwaites Bee and D. Murdoch-Eaton: Questionnaire design: the good, the bad and the pitfalls. In: *Archives of Disease in Childhood: Education & Practice Edition* 101 (2016), pp. 210–212. DOI: 10.1136/archdischild-2015-309450.
 - [12] Dietrich Braun, Harald Cherdron, and Helmut Ritter: *Polymer Synthesis: Theory and Practice: Fundamentals, Methods, Experiments*. Springer Science Business Media, (2001). ISBN: 978-3-540-41697-5.
 - [13] Lihua Chen et al.: Polymer informatics: Current status and critical next steps. In: *Materials Science and Engineering: R: Reports* 144 (2021), p. 100595. ISSN: 0927-796X. DOI: <https://doi.org/10.1016/j.mser.2020.100595>. URL: <https://www.sciencedirect.com/science/article/pii/S0927796X2030053X>.
 - [14] Davide Chicco, Matthijs Warrens, and Giuseppe Jurman: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. In: *PeerJ Computer Science* 7 (2021). DOI: 10.7717/peerj-cs.623.
 - [15] Barbara DiCicco-Bloom and Benjamin F. Crabtree: The qualitative research interview. In: *Medical Education* 40 (2006), pp. 314–321. DOI: 10.1111/j.1365-2929.2006.02418.x.
 - [16] Huan Doan Tran et al.: Machine-learning predictions of polymer properties with Polymer Genome. In: *Journal of Applied Physics* 128.17 (2020).
 - [17] David W. Eccles and Guler Aarsal: The think aloud method: what is it and how do I use it? In: *Qualitative Research in Sport, Exercise and Health* 9 (2017), pp. 514–531. DOI: 10.1080/2159676X.2017.1331501.
 - [18] A. Falk and J. Heckman: Lab Experiments Are a Major Source of Knowledge in the Social Sciences. In: *Science* 326.5952 (2009), pp. 535–538. DOI: 10.1126/science.1168244.
 - [19] Matthias Feurer et al.: Efficient and Robust Automated Machine Learning. In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc. (2015), pp. 2962–2970. URL: <https://proceedings.neurips.cc/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf>.

- [20] U. Gedde and M. Hedenqvist: *Morphology of Semicrystalline Polymers*. Graduate Texts in Physics, (2019). DOI: 10.1007/978-3-030-29794-7_7.
- [21] Georgia Institute of Technology: *Polymer Properties Dataset*. <https://khazana.gatech.edu/dataset/>. (2021).
- [22] Glossary of basic terms in polymer science. In: *Pure and Applied Chemistry* 68 (1996). IUPAC Recommendations 1996, pp. 2287–2299.
- [23] Mohammad Amin Hariri-Ardebili, Parsa Mahdavi, and Farhad Pourkamali-Anaraki: Benchmarking AutoML solutions for concrete strength prediction: Reliability, uncertainty, and dilemma. In: *Construction and Building Materials* 423 (2024), p. 135782. ISSN: 0950-0618. DOI: 10.1016/j.conbuildmat.2024.135782. URL: <https://www.sciencedirect.com/science/article/pii/S0950061824009231>.
- [24] A. F. Holleman, E. Wiberg, and N. Wiberg: *Lehrbuch der Anorganischen Chemie*. 101st ed. Berlin: Walter de Gruyter, (1995), p. 1313. ISBN: 3-11-012641-9.
- [25] A. F. Holleman, E. Wiberg, and N. Wiberg: *Lehrbuch der Anorganischen Chemie*. 101st ed. Berlin: Walter de Gruyter, (1995), p. 1313. ISBN: 3-11-012641-9.
- [26] homopolymer. In: (2019). DOI: doi:10.1351/goldbook.H02854. URL: <https://doi.org/10.1351/goldbook.H02854>.
- [27] T.D. Huan, R. Batra, J. Chapman, et al.: A universal strategy for the creation of machine learning-based atomistic force fields. In: *npj Computational Materials* 3 (2017), p. 37.
- [28] Bing Huang and O. Anatole von Lilienfeld: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. In: *Journal of Chemical Information and Modeling* 56.6 (2016), pp. 1004–1016.
- [29] Haifeng Jin et al.: AutoKeras: An AutoML Library for Deep Learning. In: *Journal of Machine Learning Research* 24.6 (2023), pp. 1–6. URL: <http://jmlr.org/papers/v24/20-1355.html>.
- [30] A. Jørgensen: Thinking-aloud in user interface design: a method promoting cognitive ergonomics. In: *Ergonomics* 33 (1990), pp. 501–507. DOI: 10.1080/00140139008927157.
- [31] Shubhra (Santu) Karmaker et al.: AutoML to Date and Beyond: Challenges and Opportunities. In: *ACM Computing Surveys (CSUR)* 54 (2020), pp. 1–36. DOI: 10.1145/3470918.
- [32] W. Klonowski: Probabilistic theory of structural and physical characteristics of crosslinked polymer systems. In: *Rheologica Acta* 18 (1979), pp. 442–450. DOI: 10.1007/BF01736949.

- [33] Miroslav Kubat: *An Introduction to Machine Learning*. 2nd ed. Springer, (2017), pp. 117–195.
- [34] Christopher Kuenneth and Rampi Ramprasad: polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. In: *Nature Communications* 14.1 (2023), p. 4099.
- [35] *Kunststoffchemie für Ingenieure*. 3rd ed. München: Carl Hanser, (2011), p. 84.
- [36] MANIMALA: *Boston House Prices*. <https://www.kaggle.com/datasets/vikrishnan/boston-house-prices>.
- [37] W. McCubbin and I. Gurney: Conduction in Paraffinic Polymers. In: *Journal of Chemical Physics* 43 (1965), pp. 983–987. DOI: 10.1063/1.1696881.
- [38] John B. O. Mitchell: Machine learning methods in chemoinformatics. In: *Wiley Interdisciplinary Reviews. Computational Molecular Science* 4 (2014), pp. 468–481. DOI: 10.1002/wcms.1183.
- [39] Piero Molino, Yaroslav Dudin, and Sai Sumanth Miryala: *Ludwig: a type-based declarative deep learning toolbox*. (2019). eprint: arXiv:1909.07930.
- [40] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining: *Introduction to Linear Regression Analysis*. 6th ed. John Wiley & Sons, (2021).
- [41] T. Mueller, A. G. Kusne, and R. Ramprasad: Machine Learning in Materials Science: Recent Progress and Emerging Applications. In: *Reviews in Computational Chemistry* 29 (2016), p. 186.
- [42] Anne Mulhall: In the field: notes on observation in qualitative research. In: *Journal of Advanced Nursing* 41.3 (2003), pp. 306–313. DOI: 10.1046/J.1365-2648.2003.02514.X.
- [43] Jakob Nielsen and Thomas K. Landauer: A mathematical model of the finding of usability problems. In: *Proceedings of ACM INTERCHI'93 Conference*. Amsterdam, The Netherlands, (1993), pp. 206–213.
- [44] I. Omar et al.: Automated Prediction of Crack Propagation Using H2O AutoML. In: *Sensors* 23 (2023), p. 8419. DOI: 10.3390/s23208419.
- [45] PAC: Definitions of terms relating to the structure and processing of sols, gels, networks, and inorganic-organic hybrid materials (IUPAC Recommendations 2007). In: *Pure and Applied Chemistry* 79 (2007), pp. 1801–1817.
- [46] Gabriel A. Pinheiro et al.: Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset. In: *The Journal of Physical Chemistry A* 124.47 (2020). PMID: 33174750, pp. 9854–9866. DOI: 10.1021/acs.

- jpca.0c05969. eprint: <https://doi.org/10.1021/acs.jpca.0c05969>. URL: <https://doi.org/10.1021/acs.jpca.0c05969>.
- [47] *Polystyrol-Struktur*. <https://de.wikipedia.org/wiki/Datei:Polyethylen.png>.
- [48] S. C. Price et al.: Fluorine substituted conjugated polymer of medium band gap yields 7% efficiency in polymer-fullerene solar cells. In: *Journal of the American Chemical Society* 133.12 (2011), pp. 4625–4631.
- [49] D. G. Pruitt: Field Experiments on Social Conflict. In: *International Negotiation* 10 (2005), pp. 33–50. DOI: 10.1163/1571806054741173.
- [50] *PSMILES Documentation*. <https://psmiles.readthedocs.io/en/latest/>. Zugriff: 16.03.2024.
- [51] Rampi Ramprasad et al.: Machine learning in materials informatics: recent applications and prospects. In: *npj Computational Materials* 3.1 (2017), p. 54.
- [52] A. Repp et al.: Direct Observation: Factors Affecting the Accuracy of Observers. In: *Exceptional Children* 55 (1988), pp. 29–36. DOI: 10.1177/001440298805500103.
- [53] S. Roopa and M. Rani: Questionnaire Designing for a Survey. In: *Journal of Indian Orthodontic Society* 46 (2012), pp. 273–277.
- [54] Matthias Rupp et al.: Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. In: *Phys. Rev. Lett.* 108 (2012), p. 058301.
- [55] Imrus Salehin et al.: AutoML: A systematic review on automated machine learning with neural architecture search. In: *Journal of Information and Intelligence* 2.1 (2024), pp. 52–81. DOI: 10.1016/j.jiixd.2023.10.002. URL: <https://www.sciencedirect.com/science/article/pii/S2949715923000604>.
- [56] Abu Sayed: Types of polymer: Requirements of fibre forming polymer. In: *Textile Apex* (2014).
- [57] Sigurd Schacht and Carsten Lanquillon, eds.: *Blockchain und maschinelles Lernen: Wie das maschinelle Lernen und die Distributed-Ledger-Technologie voneinander profitieren*. Verlagsort: Verlagsname, (2020).
- [58] K. T. Schütt et al.: How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. In: *Phys. Rev. B* 89 (2014), p. 205118.
- [59] Jiale Shi et al.: Transfer Learning Facilitates the Prediction of Polymer-Surface Adhesion Strength. In: *Journal of Chemical Theory and Computation* 19.14 (2023). PMID: 37068204, pp. 4631–4640. DOI: 10.1021/acs.jctc.2c01314. eprint: <https://doi.org/10.1021/acs.jctc.2c01314>. URL: <https://doi.org/10.1021/acs.jctc.2c01314>.

- [60] Leslie Howard Sperling: *Introduction to Physical Polymer Science*. Hoboken, N.J.: Wiley, (2006).
- [61] *Streamlit Documentation*. <https://docs.streamlit.io/>. Zugriff: Datum des Zugriffs.
- [62] *Structured Data Regression - AutoKeras*. https://autokeras.com/tutorial/structured_data_regression/. Zugriff: Datum des Zugriffs.
- [63] *Styrol-Butadien-Kautschuk*. <https://upload.wikimedia.org/wikipedia/commons/thumb/f/f8/Styrol-Butadien-Kautschuk.svg/316px-Styrol-Butadien-Kautschuk.svg.png>.
- [64] Bernd Tieke: *Makromolekulare Chemie*. 3rd ed. Weinheim: Wiley-VCH, (2014), p. 295.
- [65] Leszek A. Utracki: *Compatibilization of polymer blends*. Vol. 80. 6. S. 1012. (2002), pp. 1008–1016.
- [66] G. Vinogradov et al.: The viscoelastic properties of linear polymers in the state of flow and the transition to the high elastic state. Review. In: *Polymer Science U.S.S.R.* 20 (1978), pp. 2701–2717. DOI: 10.1016/0032-3950(78)90450-1.
- [67] B. Voit and A. Lederer: Hyperbranched and highly branched polymer architectures—synthetic strategies and major characterization aspects. In: *Chemical reviews* 109.11 (2009), pp. 5924–5973. DOI: 10.1021/cr900068q.
- [68] David Weininger: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. In: *Journal of Chemical Information and Computer Sciences* 28.1 (1988), pp. 31–36. DOI: 10.1021/ci00057a005.
- [69] Sanford Weisberg: *Applied linear regression*. Vol. 528. John Wiley & Sons, (2005).
- [70] Hironao Yamada et al.: Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. In: *ACS Central Science* 5.10 (2019). PMID: 31660440, pp. 1717–1730. DOI: 10.1021/acscentsci.9b00804. eprint: <https://doi.org/10.1021/acscentsci.9b00804>. URL: <https://doi.org/10.1021/acscentsci.9b00804>.

