

# Masterthesis

im Fachgebiet Medieninformatik

## Entwicklung neuer Verfahren zur Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen

vorgelegt von: Alexander Galperin

Studiengebiet: Medieninformatik

Matrikelnummer: 891920

Erstgutachter: Prof. Dr. Florian Huber

Zweitgutachter: Prof. Dr. Dennis Müller

©2024

Dieses Werk einschließlich seiner Teile ist **urheberrechtlich geschützt**. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Autors unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen sowie die Einspeicherung und Verarbeitung in elektronischen Systemen.

# Inhaltsverzeichnis

<b>1 Einleitung</b>	<b>1</b>
<b>2 Theoretischer Hintergrund</b>	<b>4</b>
2.1 Hochdimensionale Datenräume . . . . .	4
2.2 Distanzmetrik . . . . .	6
2.3 Bestehende Methoden und Algorithmen . . . . .	7
2.3.1 k-Nearest Neighbors . . . . .	10
2.3.2 Local Outlier Factor . . . . .	10
2.3.3 Stochastic Outlier Selection . . . . .	10
2.4 Aktueller Forschungsstand . . . . .	11
2.5 Überblick über Datenpunkte und ihre Seltenheit . . . . .	13
<b>3 Methodik</b>	<b>16</b>
3.1 Einsatz von ChatGPT . . . . .	16
3.2 Datenquellen und Datensammlung . . . . .	16
3.2.1 MNIST . . . . .	17
3.2.2 Madrid Tripadvisor Rezensionen . . . . .	18
3.2.3 Audio MNIST . . . . .	19
3.2.4 Molekulare Daten . . . . .	20
3.3 Defintion von Seltenheit . . . . .	22
3.4 Datenvorbereitung . . . . .	23
<b>4 Entwicklung</b>	<b>25</b>
4.1 Entwicklungsumgebung und Pakete . . . . .	25
4.2 Nächste-Nachbarn-Methode . . . . .	28

4.3	Flußsuch-Methode . . . . .	31
4.4	Ausreißer-Methoden . . . . .	34
<b>5</b>	<b>Hypothese und Ergebnisse</b>	<b>36</b>
5.1	Hypothesen . . . . .	36
5.2	Experimente . . . . .	38
5.3	Resultate . . . . .	41
5.3.1	MNIST und Audio MNIST . . . . .	41
5.3.2	Madrid Tripadvisor Rezensionen . . . . .	57
5.3.3	Molekulare Daten . . . . .	59
<b>6</b>	<b>Diskussion und Ausblick</b>	<b>60</b>
6.1	Diskussion der Ergebnisse . . . . .	60
6.2	Ausblick . . . . .	62
	<b>Abkürzungsverzeichnis</b>	<b>64</b>
	<b>Algorithmusverzeichnis</b>	<b>66</b>
	<b>Abbildungsverzeichnis</b>	<b>69</b>
	<b>Literaturverzeichnis</b>	<b>77</b>

## **Kurzfassung**

Diese Masterarbeit untersucht die Entwicklung neuer Methoden zur Identifizierung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen, eine Notwendigkeit, die mit dem rasanten Anstieg der Datenmenge in verschiedenen Wissenschafts- und Technologiefeldern zunimmt. Angesichts des „Fluchs der Dimensionalität“, der die Datenanalyse in solchen komplexen Räumen erschwert, zielt diese Arbeit darauf ab, innovative Ansätze zu entwickeln und zu bewerten. Diese Methoden zeigen nicht nur ihre praktische Anwendbarkeit, sondern auch ihr Potenzial, analytische Grenzen zu erweitern. Durch den Vergleich dieser spezialisierten Methoden mit traditionellen Ausreißererkennungsverfahren in umfangreichen Experimenten werden ihre Effektivität und spezifische Stärken sowie Schwächen beleuchtet. Die Ergebnisse zeigen, dass die neuen Methoden einen pragmatischen Ansatz für die Seltenheitsbewertung bieten und tragen wesentlich zum Verständnis der Analyse in hochdimensionalen Daten bei.

## **Abstract**

This master's thesis examines the development of new methods for identifying the rarity of data points in high-dimensional datasets, a necessity that is growing with the rapid increase in data volume across various science and technology fields. Given the "curse of dimensionality" that complicates data analysis in such complex spaces, this work aims to develop and evaluate innovative approaches. These methods not only demonstrate their practical applicability but also their potential to extend analytical boundaries. By comparing these specialized methods with traditional outlier detection techniques in extensive experiments, their effectiveness and specific strengths and weaknesses are highlighted. The results show that the new methods offer a pragmatic approach to rarity assessment and significantly contribute to the understanding of analysis in high-dimensional data.

# 1 Einleitung

Die vorliegende Masterarbeit beschäftigt sich mit der Entwicklung neuer Verfahren zur Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen, einem Gebiet, das angesichts der stetig wachsenden Datens Mengen in verschiedenen wissenschaftlichen und technologischen Bereichen von erheblicher Bedeutung ist (Cukier u. Mayer-Schoenberger, 2013). Die Analyse hochdimensionaler Daten stellt eine Reihe von Herausforderungen dar, die als „Fluch der Dimensionalität“ bekannt sind (Donoho u. a., 2000). Dies bezieht sich auf die zunehmenden Schwierigkeiten bei der Suche, Approximation und Integration in hochdimensionalen Räumen. Diese Probleme erscheinen unlösbar, wenn die Dimensionen der Daten zunehmen. Eine Dimension in einem Datensatz ist eine bestimmte Eigenschaft oder Variable, die die Daten beschreibt oder repräsentiert (Oracle, 2002). In der Ära der Big Data, in der wir uns befinden, werden immer mehr Daten aus verschiedenen Quellen gesammelt, wie zum Beispiel Hyperspektrale Bilder, Internetportale, Finanzdaten und DNA-Mikroarrays (Donoho u. a., 2000). Diese Daten sind oft hochdimensional, was bedeutet, dass sie aus vielen Beobachtungen und diversen Variablen bestehen.

Das Hauptziel dieser Arbeit ist die Entwicklung und Evaluierung fortschrittlicher Methoden, die spezialisiert darauf sind, die Seltenheit von Datenpunkten in hochdimensionalen Datensätzen effektiv und verlässlich zu erfassen. Diese Methoden sollen so konzipiert werden, dass sie weit über rein theoretische Konzepte hinausgehen und eine umfassende praktische Anwendbarkeit aufzeigen. Ihr Einsatzspektrum ist breit gefächert und nicht auf einen spezifischen Datentyp beschränkt, sondern zielt darauf ab, universell auf verschie-

denste Daten angewendet werden zu können.

Die Betrachtung von Naturstoffen dient in diesem Kontext als inspirierende Motivation und illustriert eine der potenziellen Anwendungen. Naturstoffe, verstanden als Biomaterialien, die aus der natürlichen Umgebung gewonnen werden (Burkel u. a., 2012), bieten ein Beispiel, wo Methoden entwickelt werden müssen, um seltene Molekülstrukturen zu identifizieren, die für biomedizinische Anwendungen relevant sind. Diese Anwendung ist jedoch nur eine von vielen, da die entwickelten Techniken universell konzipiert sind, um Muster und Anomalien in jeglichen hochdimensionalen Datensets zu erkennen, unabhängig von ihrem spezifischen Ursprung oder ihrer Domäne. Die Problematik der seltenen Datenpunkte in hochdimensionalen Datensätzen und deren Identifikation ist eng mit den Herausforderungen verknüpft, die Machine Learning in der Forschung zu seltenen Krankheiten gegenübersteht (Banerjee u. a., 2023). Ein wesentliches Problem dabei ist die geringe Anzahl von Daten oder Proben für spezifische Krankheiten, was die Anwendung von Machine Learning-Techniken erschwert. Diese Schwierigkeit wird durch die Definition seltener Krankheiten unterstrichen, die per se nur wenige Patient\*innen oder Proben für eine spezifische Erkrankung aufweisen (Banerjee u. a., 2023). Die geringe Datenmenge führt zu einer Vielzahl von Problemen, darunter die unzureichende Repräsentation jeder Klasse im Datensatz, was dazu führt, dass relevante Variabilität innerhalb dieser Klassen nicht vollständig erfasst wird. Die Motivation, solche seltenen Punkte zu identifizieren, speist sich somit aus dem Bedürfnis, die Genauigkeit und Effektivität von Machine Learning-Modellen zu verbessern und neue Erkenntnisse in Bereichen zu gewinnen, in denen Daten intrinsisch begrenzt oder schwer zu sammeln sind.

Darüber hinaus widmet sich die Motivation und die konkrete Anwendung der entwickelten Verfahren zur Bewertung der Seltenheit von Datenpunkten auch der Betrachtung spezifischer, herausfordernder Situationen, wie sie in der Forschung und Praxis auftreten. Ein markantes Beispiel hierfür ist die frühe Phase neu auftretender Infektionskrankheiten, exemplarisch dargestellt am Fall von COVID-19 (Feng u. a., 2023). In solchen Initialphasen sind

die Fälle oft limitiert, und das Auftreten der Krankheit kann in der Bevölkerung als seltenes Ereignis betrachtet werden, da sie sich noch nicht weit verbreitet hat. Die präzise Vorhersage und frühzeitige Identifizierung solcher seltenen Ereignisse kann entscheidend sein, um frühzeitige und gezielte Interventionen zu ermöglichen, die das Risiko einer Ausbreitung minimieren und die Entwicklung von Behandlungsstrategien beschleunigen (Feng u. a., 2023).

Weiterhin unterstreicht die Betrachtung der Auswirkungen des Klimawandels auf das Auftreten neuer Krankheiten, wie hitzebedingte Erkrankungen oder neu auftretende durch Vektoren übertragene Krankheiten (Feng u. a., 2023), die Bedeutung dieser Arbeit. In ihren Anfangsstadien ist es von größter Wichtigkeit, effektive Vorhersagemodelle zu besitzen, um präventive Maßnahmen zu gestalten und Risiken zu mindern. Darüber hinaus ist die Anwendbarkeit der entwickelten Methoden auf seltene Formen von Krebs oder spezielle medizinische Zustände wie Neonataler Diabetes mellitus gegeben (Feng u. a., 2023). Hier ermöglicht die genaue Vorhersage und Identifizierung seltener Ereignisse eine frühzeitige Diagnose und Behandlung, was die Lebensqualität betroffener Patienten erheblich verbessern kann.

Die Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen ist eine wichtige Aufgabe, da sie dazu beitragen kann, Muster und Trends in den Daten zu erkennen, die sonst möglicherweise übersehen würden. Allerdings ist diese Aufgabe aufgrund des oben genannten „Fluch der Dimensionalität“ nicht trivial. In dieser Arbeit werden neue Verfahren zur Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen entwickelt und untersucht, um diese Herausforderungen zu bewältigen.

## 2 Theoretischer Hintergrund

Dieses Kapitel legt das Fundament für ein tiefes Verständnis der Herausforderungen und Chancen, die hochdimensionale Datensätze mit sich bringen. Eine solide theoretische Basis stellt nicht nur den Kontext für die anschließenden Diskussionen bereit, sondern hebt auch die Bedeutung der Entwicklung neuer Verfahren zur Bewertung der Seltenheit in diesen Datenräumen hervor.

### 2.1 Hochdimensionale Datenräume

Hochdimensionalität in Datensätzen bezeichnet eine Situation, in der jeder Datenpunkt durch eine große Anzahl von Attributen oder Merkmalen beschrieben wird (Liu u. a., 2017). Diese Eigenschaft findet sich häufig in Daten aus Bereichen wie der Wirtschaft, Biologie, Chemie und Physik. Die Herausforderungen, die mit hochdimensionalen Daten einhergehen, sind vielfältig und komplex und umfassen mehrere Schlüsselaspekte (Liu u. a., 2017).

Einer der Hauptaspekte ist der „Fluch der Dimensionalität“, ein Begriff, der die zunehmende Komplexität und die Herausforderungen beschreibt, die mit jeder zusätzlichen Dimension in den Daten entstehen (Bellman u. Kalaba, 1959). Die Verteilung der Daten in hochdimensionalen Räumen unterscheidet sich von niedrigdimensionalen Räumen. In hochdimensionalen Räumen tendieren Daten dazu, an den Rändern des Raumes verteilt zu sein, was bedeutet, dass das Innere des Raumes relativ leer ist (Bellman u. Kalaba, 1959). Diese Besonderheit hat Auswirkungen auf statistische Modelle und Algorithmen, die auf die Daten angewendet werden. Dies erschwert es, statistisch

signifikante Muster zu erkennen und führt zu einer Distanzverzerrung, bei der die Unterschiede zwischen den nächsten und fernsten Nachbarn tendenziell geringer werden (Bellman u. Kalaba, 1959).

Ein zentrales Konzept im Umgang mit hochdimensionalen Daten ist die Dimensionsreduktion (Van Der Maaten u. a., 2009). Sie zielt darauf ab, eine sinnvolle niedrigdimensionale Darstellung eines gegebenen hochdimensionalen Datensatzes zu erreichen. Laut Waggoner (2021) bietet die Dimensionsreduktion die Möglichkeit, komplexe, hochdimensionale Datensätze einfacher und handhabbarer zu machen, indem sie die Anzahl der Variablen (Dimensionen) verringert. Es gibt verschiedene Techniken zur Dimensionsreduktion, darunter die Hauptkomponentenanalyse (PCA), lokal lineare Einbettung, t-Distributed Stochastic Neighbor Embedding (t-SNE), einheitliche Mannigfaltigkeitsapproximation und Projektion, selbstorganisierende Karten und tiefe Autoencoder (Waggoner, 2021). Diese Techniken helfen dabei, die Komplexität von hochdimensionalen Daten zu bewältigen.

In hochdimensionalen Räumen ist das Konzept der Distanz nicht so intuitiv wie in niedrigeren Dimensionen. Eine der Herausforderungen ist das Phänomen der Distanzkonzentration (Angiulli, 2018). Die Distanzkonzentration bezieht sich auf die Tendenz, dass Distanzen in hochdimensionalen Räumen fast ununterscheidbar werden. Dieses Phänomen kann die Qualität und Leistung von maschinellem Lernen, Data Mining und Information Retrieval Techniken erheblich beeinflussen (Angiulli, 2018). Dieses Phänomen der Distanzkonzentration tritt in einer Vielzahl von Datenverteilungen auf, einschließlich zentral verteilter und gruppierter Daten (Angiulli, 2018). Es ist wichtig zu beachten, dass die Auswirkungen der Distanzkonzentration von der Art der Datendistribution abhängen. Es tritt bei unabhängigen und identisch verteilten (iid) Datendistributionen mit endlichen Momenten sowie bei Daten mit korrelierten Merkmalen auf. Es gibt jedoch auch Datendistributionen, die keine Distanzkonzentration aufweisen, wie lineare latente Variablenmodelle (Angiulli, 2018).

Das Thema der Arbeit, die Entwicklung neuer Verfahren zur Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen, geht diese Herausforderungen direkt an. Es geht darum, Methoden zu entwickeln, die in der Lage sind, die Seltenheit von Datenpunkten in diesen komplexen und oft spärlich besiedelten Räumen effektiv zu beurteilen.

## 2.2 Distanzmetrik

Distanzmetriken, die auf mathematischen Formeln basieren, nehmen zwei Punkte im Eingaberaum eines Problems und berechnen eine positive Zahl, die nicht nur Auskunft darüber gibt, wie nah oder fern sich diese Punkte von einander befinden (Draghici, 2012), sondern auch als Maß für ihre Ähnlichkeit dient (Cha, 2007). Bei der Auswahl einer Distanzmetrik müssen bestimmte Kriterien erfüllt sein. Ein wesentliches Kriterium ist die Symmetrie, was bedeutet, dass die Distanz von Punkt A zu Punkt B gleich der Distanz von Punkt B zu Punkt A sein muss. Des Weiteren muss die Distanz stets positiv sein und nur dann den Wert Null annehmen, wenn die beiden Punkte identisch sind (Chen u. a., 2009). Dies wird als Positivität bezeichnet. Schließlich muss die Distanz auch die Dreiecksungleichung erfüllen, was besagt, dass die direkte Distanz zwischen zwei Punkten immer kürzer oder gleich der Summe der Distanzen über einen dritten Punkt sein muss (Chen u. a., 2009).

Die Wahl der richtigen Distanzmetrik in der Datenanalyse hängt von verschiedenen Faktoren ab. Dazu gehören die Art der Daten (z.B. unterschiedliche Maßeinheiten), die Dimensionalität der Daten (insbesondere in hochdimensionalen Räumen), die Verteilung der Daten (z.B. Ausreißer) und die spezifischen Anforderungen der Analyse (z.B. Clustering oder Klassifizierung) (Draghici, 2012). Je nach Situation kann eine unterschiedliche Distanzmetrik erforderlich sein, um die besten Ergebnisse zu erzielen. So konnte beispielsweise in der Studie von Rusdiana u. a. (2021) die Minkowski-Distanz als optimale Wahl für die Kundensegmentierung identifiziert werden, da ihre Anpassungsfähigkeit es ermöglicht, die Distanzmetrik präzise auf die spezi-

fischen Eigenschaften der Daten und die Anforderungen der Analyse abzustimmen.

## 2.3 Bestehende Methoden und Algorithmen

Ein Ansatz zur Zuweisung von Seltenheitswerten zu Datenpunkten in hochdimensionalen Datensätzen besteht darin, Methoden zur Anomalieerkennung zu nutzen. Anomalieerkennung ist ein wichtiges Forschungsfeld, das sich über verschiedene Bereiche und Disziplinen erstreckt und sich mit der Herausforderung befasst, Muster in Daten zu identifizieren, die von erwartetem oder normalem Verhalten abweichen (Chandola u. a., 2009). Diese Abweichungen, oft als Anomalien, Ausreißer oder Ausnahmen bezeichnet, sind bedeutsam, da sie kritische und handlungsrelevante Erkenntnisse in einer breiten Palette von Anwendungen liefern können, einschließlich Betrugserkennung, Cybersicherheit und Gesundheitswesen (Chandola u. a., 2009). Ihre Anwendung ist breit gefächert und vielseitig und befasst sich mit entscheidenden Fragen in verschiedenen Bereichen, stellt aber auch einzigartige Herausforderungen aufgrund der subjektiven und sich entwickelnden Natur dessen dar, was als Anomalie gilt.

Die Seltenheitsbewertung kann mithilfe der Entscheidungswerte von Anomalieerkennungsmethoden durchgeführt werden. Diese Werte sind beispielsweise in der Python-Bibliothek *PYOD* verfügbar. Die Python Outlier Detection (*PYOD*) Bibliothek ist eine vielseitige und skalierbare Ressource für die Erkennung von Ausreißern oder Anomalien in hochdimensionalen Datensätzen (Zhao u. a., 2019). Sie enthält eine breite Palette von über 50 Erkennungsalgorithmen, angefangen bei traditionellen Ansätzen wie Local Outlier Factor (LOF) (Breunig u. a., 2000) bis hin zu den neuesten Techniken wie Empirical-Cumulative-distribution-based Outlier Detection (ECOD) (Li u. a., 2023) und Deep Isolation Forest (DIF) (Xu u. a., 2023). Diese Entscheidungswerte repräsentieren die Ausreißerwerte innerhalb der Trainingsdaten. Ein höherer Wert deutet auf eine größere Abnormalität hin (Zhao u. a., 2019). Folglich neigen Ausreißer dazu, höhere Entscheidungswerte zu haben. Diese

Werte können genutzt werden, um die Seltenheit von Datenpunkten zu bewerten. Die Werte können im Bereich von 0 bis 1 normalisiert werden, wobei 0 für häufig vorkommende Punkte steht und 1 für Ausreißer oder, falls vorhanden, äußerst seltene Punkte. Mit dieser Methode lässt sich die relative Seltenheit von Datenpunkten in hochdimensionalen Datensätzen annährend erfassen und nutzen (Zhao u. a., 2019).

„ADBench: Anomaly Detection Benchmark“ stellt eine umfassende Studie dar, die die Performance von mehr als 30 verschiedenen Anomalieerkennungsalgorithmen anhand von 57 verschiedenen Datensätzen untersucht und bewertet (Han u. a., 2022b). Diese umfassende Untersuchung bietet wertvolle Einblicke in die Auswirkungen von Überwachung und unterschiedlichen Arten von Anomalien auf die Effektivität der Anomalieerkennung. Durch die Ergebnisse dieser Studie können geeignete Methoden zur Bewertung der Seltenheit von Daten identifiziert werden.

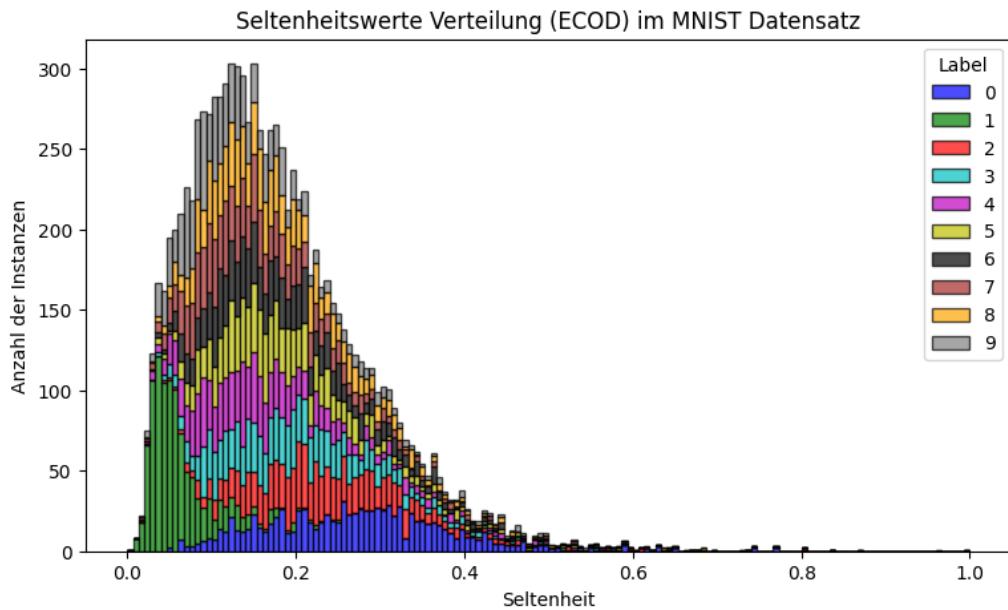


Abbildung 2.1: Seltenheitswerte von MNIST Daten (siehe Kapitel 3.2) anhand von ECOD

Wenn man sich auf bewährte Algorithmen wie LOF (Breunig u. a., 2000), Isolation Forest (Liu u. a., 2008), Subspace Outlier Detection (Sathe u. Aggarwal, 2016) oder ECOD (Li u. a., 2023) stützt, die in der Studie als leistungsfähig in der Erkennung von Ausreißern identifiziert wurden, kann man diese Algorithmen nutzen, um Seltenheitswerte für Daten in hochdimensionalen Datensätzen zu erfassen, wie in Abbildung 2.1 gezeigt wird. Dies ermöglicht eine Einschätzung der relativen Seltenheit von Datenpunkten in komplexen Datensätzen mit vielen Dimensionen. Die Anwendung dieser etablierten Algorithmen kann somit dazu beitragen, einen Maßstab für Seltenheit zu schaffen, ohne eine vorherige explizite Definition für Seltenheit festlegen zu müssen.

Andererseit kann die Wahrnehmung von Seltenheit von Fall zu Fall variieren und hängt von den spezifischen Anforderungen und Zielen eines Projekts ab (siehe 2.4). Was in einem Szenario als selten angesehen wird, mag in einem anderen Szenario als häufig gelten. Ohne klare Definition kann die Interpretation von Ausreißerergebnissen subjektiv sein und zu Fehlinterpretationen führen. In der Biologie kann etwas als selten eingestuft werden, basierend auf seiner Häufigkeit und Verbreitung Gaston (1994), was sich von anderen Feldern wie der Finanzanalyse unterscheidet, wo die Kriterien für Seltenheit anders definiert sein können. Ohne klare Kontextdefinitionen ist es schwierig, die Seltenheit angemessen zu bewerten. Es ist auch möglich, dass Ausreißer-Algorithmen Datenpunkte identifizieren, die tatsächlich selten sind, aber auch solche, die aufgrund unbekannter Informationen als selten erscheinen. Diese unbekannten Faktoren könnten die Seltenheit beeinflussen, aber die Methoden sind nicht in der Lage, dies zu erfassen.

Folglich können Ausreißererkennungsmethoden keine zuverlässigen und konsistenten Ergebnisse liefern, da sie auf vordefinierten Funktionen zur Ausreißererkennung basieren (Zhao u. a., 2019), welche unter Umständen nicht die beabsichtigte Seltenheit im gegebenen Kontext angemessen abbilden. Dies führt zu einer potenziell ungenauen Bewertung der Seltenheit von Daten.

### **2.3.1 k-Nearest Neighbors**

Der k-Nearest Neighbors (KNN) Algorithmus ist ein grundlegendes Verfahren in der Datenanalyse, das zur Erkennung von Ausreißern in Datensätzen verwendet wird. Der Kerngedanke des KNN-Verfahrens ist es, die Distanz eines Datenpunktes zu seinen  $k$  nächsten Nachbarn zu berechnen (Dang u. a., 2015). Die Ausreißerwertung basiert darauf, wie unähnlich ein Punkt im Vergleich zu seinen Nachbarn ist. Ein hoher Grad an Unähnlichkeit deutet darauf hin, dass der Punkt ein potenzieller Ausreißer ist. Die Auswahl des Parameters  $k$  ist entscheidend, da er die Anzahl der berücksichtigten Nachbarn bestimmt und somit die Sensitivität des Algorithmus gegenüber Ausreißern steuert. Ein zu kleines  $k$  macht den Algorithmus anfällig für Rauschen, während ein zu großes  $k$  dazu führen kann, dass Ausreißer nicht erkannt werden, da sie in die lokale Nachbarschaft der vielen anderen Punkte eingebettet sind.

### **2.3.2 Local Outlier Factor**

Die Local Outlier Factor (LOF)-Methode erweitert die Idee der  $k$ -nächsten Nachbarn, indem sie nicht nur die Distanz, sondern auch die lokale Dichteverteilung der Datenpunkte berücksichtigt (Breunig u. a., 2000). Sie quantifiziert die Ausreißereigenschaft eines Datenpunktes, indem sie das Verhältnis der lokalen Dichte dieses Punktes mit denen seiner Nachbarn vergleicht. Ein LOF-Wert größer als 1 deutet darauf hin, dass der Datenpunkt eine niedrigere Dichte aufweist als seine Nachbarn, was ihn zu einem potenziellen Ausreißer macht. Diese Methode ist besonders effektiv in Datensätzen, in denen Ausreißer in Regionen niedriger Dichte liegen, während normale Datenpunkte in dichteren Regionen gruppiert sind.

### **2.3.3 Stochastic Outlier Selection**

Der Stochastic Outlier Selection (SOS) Algorithmus modelliert Datenpunkte und ihre Beziehungen durch einen stochastischen Nachbargraphen, wobei Datenpunkte als Knoten und die Wahrscheinlichkeit, dass ein Punkt einen anderen als seinen Nachbarn wählt, als Kanten dargestellt werden (Janssens,

2013). Die Bindungswahrscheinlichkeiten, die aus den Affinitäten zwischen den Datenpunkten abgeleitet werden, bestimmen die Kanten des Graphen. Ein Datenpunkt wird als Ausreißer identifiziert, wenn er keine eingehenden Kanten hat, was bedeutet, dass er von anderen Punkten isoliert ist. Durch die Betrachtung vieler solcher Graphen und deren Bindungsverhältnisse kann die Ausreißerwahrscheinlichkeit für jeden Datenpunkt geschätzt werden, basierend auf der Häufigkeit, mit der er in diesen Graphen als isoliert erscheint. Diese Methode ermöglicht eine differenzierte Betrachtung der Datenstruktur, um Ausreißer zuverlässig zu identifizieren.

## 2.4 Aktueller Forschungsstand

Die Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen ist ein komplexes und herausforderndes Problem, das in der aktuellen Forschung intensiv untersucht wird. Es gibt nur wenige Ansätze und Methoden, die zur Lösung dieses Problems vorgeschlagen wurden.

Eine bemerkenswerte Arbeit in diesem Bereich ist die Studie „Rarity Score: A New Metric to Evaluate the Uncommonness of Synthesized Images“. In dieser Arbeit schlagen die Autor\*innen eine neue Bewertungsmetrik vor, den „Rarity Score“, um die individuelle Seltenheit jedes von generativen Modellen synthetisierten Bildes zu messen (Han u. a., 2022a). Sie zeigen empirisch, dass häufige Proben nahe beieinander liegen und seltene Proben weit voneinander entfernt sind in den nächsten Nachbardistanzen des Merkmalsraums. Dieser Ansatz bietet eine neue Perspektive auf das Problem der Seltenheitsbewertung und eröffnet neue Möglichkeiten für die Entwicklung von Algorithmen und Techniken in diesem Bereich (Han u. a., 2022a).

Die Methode zur Berechnung des Rarity Score in diesem Paper basiert auf der Beobachtung, dass gemeinsame Muster in den generierten Bildern dazu neigen, nahe beieinander zu liegen, während seltene Muster weit voneinander entfernt sind. Diese Beobachtung wird genutzt, um einen Score mit k-Nearest Neighbors (siehe 2.3.1) zu berechnen, der die Seltenheit jedes generierten Bildes quantifiziert. Der Rarity Score wird dann verwendet, um zu

demonstrieren, inwieweit verschiedene generative Modelle seltene Bilder erzeugen können.

Ein weiterer wichtiger Beitrag zur Forschung in diesem Bereich ist die Arbeit „Pixel Rarity Score: Rarity Learned Directly From Pixel Data“. In diesem Paper diskutieren die Autoren eine pixelbasierte Seltenheit, die auf Rohpixeldaten basiert, um Seltenheit zu bestimmen (Lommers u. a., 2023). Dieser Algorithmus analysiert die Pixel in Pixelbildern, die als Non-Fungible Tokens (NFTs) vorliegen, und weist jedem Pixel basierend auf seiner Seltenheit einen Wert zu. Die Bewertung der Pixel erfolgt durch einen Algorithmus, der jedem Pixel innerhalb eines NFT eine Seltenheitsbewertung zuordnet, basierend auf der empirischen Wahrscheinlichkeitsdichte der Rohpixelwerte in der Stichprobe. Für jedes Pixel wird die Wahrscheinlichkeit seines Auftretens im Vergleich zur gesamten Stichprobe berechnet. Seltene Pixel, die weniger häufig im Datensatz vorkommen, erhalten eine höhere Seltenheitsbewertung. Dieser Ansatz ist besonders interessant, da er eine direkte Methode zur Bestimmung der Seltenheit auf der Grundlage von Rohdaten bietet, was in vielen Anwendungsfällen von Vorteil sein kann.

Außerdem ist die Studie „Rarity: Discovering rare cell populations from single-cell imaging data“ in Bezug auf die Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen ein wichtiger Beitrag. Die Autor\*innen dieser Studie haben eine innovative Methode namens „Rarity“ entwickelt, die auf einem Bayesschen latenten Variablenmodell basiert (Märtens u. a., 2022). Diese Methode ermöglicht es, seltene Zellpopulationen in Einzelzell-Bilddaten zu identifizieren und bietet eine neue Perspektive auf das Problem der Seltenheitsbewertung.

Die Methode „Rarity“ von Märtens u. a. (2022) verbessert die Sensitivität für seltene Zellpopulationen und ermöglicht gleichzeitig die Kontrolle und Interpretation potenzieller falsch-positiver Entdeckungen. Dieser Ansatz klassifiziert Zellen nach bestimmten Mustern, die entweder vorhanden oder nicht vorhanden sind. Jede Zelle oder Zellgruppe wird durch ein einzigartiges Muster gekennzeichnet. Die Seltenheit wird ermittelt, indem man nach einmali-

gen Musterkombinationen sucht, die seltene Zelltypen repräsentieren könnten. Die Methode wurde an verschiedenen bildgebenden Massenzytometrie-Datensätzen getestet und betont die Herausforderungen, die mit der Identifizierung seltener Zelltypen verbunden sind. Es wird festgestellt, dass herkömmliche unüberwachte Ansätze dazu neigen, solche seltenen Unterpopulationen zu übersehen.

Trotz der Fortschritte in der Forschung zur Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen, gibt es derzeit noch keine universelle Methode, die einen Seltenheitswert für jegliche Art von Datensätzen (wie Bilder, Audio, Text usw.) liefern kann. Die meisten der vorhandenen Methoden, wie oben beschrieben, sind spezifisch für bestimmte Arten von Daten konzipiert und können nicht ohne Weiteres auf andere Datentypen angewendet werden.

Dies unterstreicht die Komplexität des Problems und die Notwendigkeit weiterer Forschung in diesem Bereich. Es besteht ein dringender Bedarf an Methoden, die in der Lage sind, die Seltenheit von Datenpunkten in einer Vielzahl von Datensätzen zu bewerten, und es ist wahrscheinlich, dass dies ein wichtiger Schwerpunkt zukünftiger Forschungsarbeiten in diesem Bereich sein wird.

## **2.5 Überblick über Datenpunkte und ihre Seltenheit**

Seltenheit ist ein zentrales Konzept in verschiedenen wissenschaftlichen Disziplinen, einschließlich der Statistik (King u. Zeng, 2002), Ökologie (Ellison u. Agrawal, 2005) und Datenwissenschaft (Märtens u. a., 2022). Sie spielt eine wichtige Rolle bei der Bewertung der Bedeutung und des Einflusses von Beobachtungen innerhalb eines gegebenen Datensatzes. In der Datenanalyse ist die Identifizierung seltener Datenpunkte von entscheidender Bedeutung, da diese oft wertvolle Einsichten oder Anomalien repräsentieren können (Weiss,

2004). Ein Datenpunkt kann als eine einzelne Beobachtung oder Messung innerhalb eines Datensatzes verstanden werden. Diese Datenpunkte können verschiedene Merkmale oder Eigenschaften aufweisen, abhängig von ihrem Kontext und ihrer Natur. Die Seltenheit eines Datenpunkts wird oft durch seine Abweichung von der Mehrheit oder der Norm des Datensatzes definiert. Dabei kann diese Seltenheit entweder absolut oder relativ sein (Weiss, 2004).

Die Unterscheidung zwischen absoluter und relativer Seltenheit ist ein zentraler Aspekt in der Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen. Absolute Seltenheit bezieht sich auf die tatsächliche Anzahl, wie oft ein spezifisches Ereignis oder eine Eigenschaft innerhalb eines Datensatzes auftritt. Ein typisches Beispiel hierfür ist ein Datenpunkt, der nur einmal in einem Datensatz von 1.000 Punkten auftritt. Diese Form der Seltenheit ist unmittelbar und direkt messbar (Weiss, 2004).

Im Gegensatz dazu bezieht sich die relative Seltenheit auf die Häufigkeit eines Ereignisses oder einer Eigenschaft im Verhältnis zur Gesamtzahl der Ereignisse oder Eigenschaften in einem Datensatz. Wenn beispielsweise ein Datenpunkt nur einmal unter 1.000 Punkten auftritt, beträgt seine relative Seltenheit 0,001 oder 0,1%. Diese Art der Seltenheit ist besonders hilfreich, um die Bedeutung eines Datenpunktes im Kontext des gesamten Datensatzes zu bewerten. Sie ermöglicht es, Datenpunkte in Bezug auf ihre Seltenheit innerhalb des Datensatzes zu klassifizieren und zu priorisieren (Weiss, 2004).

In dieser Arbeit liegt der Fokus aufgrund der genannten Vorteile auf der relativen Seltenheit. Jeder Datenpunkt in einem hochdimensionalen Datensatz wird hinsichtlich seiner Seltenheit bewertet, indem ihm ein spezifischer Seltenheitswert zugewiesen wird. Diese Vorgehensweise zielt darauf ab, Ausreißer als die seltensten Punkte innerhalb des Datensatzes zu identifizieren. Die Zuweisung eines Seltenheitswertes an jeden Datenpunkt erfordert eine methodische Herangehensweise, die die spezifischen Merkmale und Strukturen des Datensatzes berücksichtigt. Indem die relative Seltenheit von Datenpunkten in den Vordergrund gestellt wird, können subtile Muster und Beziehungen innerhalb der Daten aufgedeckt werden, die bei einer ausschließlichen

Betrachtung der absoluten Seltenheit möglicherweise unerkannt bleiben würden.

Die Ermittlung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen erfordert eine differenzierte und methodisch fundierte Herangehensweise. Hierbei ist es essentiell, eine spezifische, auf das jeweilige Projekt zugeschnittene Definition von Seltenheit zu etablieren. Dies ermöglicht eine angepasste Bewertung der Datenpunkte, die nicht nur die allgemeine Seltenheit berücksichtigt, sondern auch die besonderen Eigenschaften und Dimensionen des Datensatzes einbezieht. Im nachfolgenden Kapitel Methodik wird die Definition von Seltenheit, wie sie in dieser Arbeit verwendet wird, eingehend erläutert.

# **3 Methodik**

Die methodische Herangehensweise dieser Masterarbeit bildet das Kernstück der Forschung und legt den Grundstein für die Entwicklung neuer Verfahren zur Bewertung der Seltenheit von Datenpunkten. Dieses Kapitel umfasst eine detaillierte Beschreibung der verwendeten Datenquellen, Datenvorbereitungsprozesse sowie der konzeptionellen Überlegungen zur Definition von Seltenheit.

## **3.1 Einsatz von ChatGPT**

ChatGPT (OpenAI, 2024), ein hochentwickeltes KI-basiertes Sprachmodell, wurde gezielt eingesetzt, um den Entwurf der Masterarbeit zu optimieren. Dieser Prozess beinhaltete die Neugestaltung von Textabschnitten, Verbesserung der Kohärenz und logischen Struktur sowie die Grammatik- und Rechtschreibprüfung. Zusätzlich wurde ChatGPT zur Unterstützung beim Programmieren verwendet, indem es bei der Code-Entwicklung, Fehlerbehebung und Optimierung von Algorithmen assistierte. Diese methodische Herangehensweise gewährleistet, dass die Arbeit nicht nur inhaltlich präzise, sondern auch sprachlich einwandfrei, gut strukturiert und technisch fundiert ist.

## **3.2 Datenquellen und Datensammlung**

Die Identifizierung und Analyse der Seltenheit erfordert einen sorgfältigen Umgang mit Datenquellen und deren Sammlung. In dieser Arbeit werden diverse Datensätze ausgewählt, die eine breite Palette von Anwendungsfällen

abdecken und somit eine umfassende Grundlage für die Untersuchung bieten. Von handgeschriebenen Ziffern im MNIST-Datensatz über Kundenbewertungen von Restaurants bis hin zu Audioaufnahmen gesprochener Zahlen und molekularen Daten, reflektieren die verwendeten Datensätze die Vielfalt und Komplexität hochdimensionaler Daten in unterschiedlichen Kontexten.

### 3.2.1 MNIST

Der Modified National Institute of Standards and Technology (MNIST) Datensatz, entwickelt von Lecun u. a. (1998) und häufig in der Forschungsliteratur zitiert, dient als Benchmark für Bilderkennungsalgorithmen in der Informatik und Data Science. Ursprünglich aus der NIST-Datenbank zusammengestellt, umfasst er 70.000 Bilder handgeschriebener Ziffern, aufgeteilt in 60.000 Trainings- und 10.000 Testbilder. Jedes Bild ist ein 28x28 Pixel großes, graustufiges Bild einer Ziffer zwischen 0 und 9. Diese Struktur führt zu einer hohen Dimensionalität der Daten, die für das Studium der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen ideal ist.

In der Forschung wurde der MNIST-Datensatz auch verwendet, um die Leistung von Convolutional Neural Networks (CNNs) zu bewerten, speziell in der Bilderkennung und -detektion. Diese Modelle erreichten eine Genauigkeit von bis zu 99,6 % auf dem MNIST-Datensatz, was seine Eignung als Test- und Vergleichsbasis für solche Algorithmen unterstreicht (Chauhan u. a., 2018).

Der MNIST-Datensatz eignet sich besonders gut für die Untersuchung der Seltenheit von Datenpunkten, da die Variationen innerhalb einer Ziffernkategorie sowie zwischen den Kategorien ausreichend divers sind. Einige Ziffern, wie beispielsweise die „1“, weisen tendenziell weniger Variabilität auf als komplexere Ziffern wie „8“ oder „9“. Diese Eigenschaft ermöglicht es, die Wirksamkeit von Seltenheitsbewertungsverfahren zu testen, indem man analysiert, wie gut diese Verfahren in der Lage sind, selten auftretende Varianten innerhalb einer häufig vorkommenden Ziffer zu identifizieren.

Zusätzlich ist die Seltenheit bei MNIST leichter zu verstehen als bei anderen Datensätzen, da er visuell zugänglich und minimalistisch gestaltet ist. Die Ziffern sind klar erkennbar und die Einfachheit der Bilder erleichtert die visuelle Identifikation von Seltenheiten. Diese Eigenschaften machen MNIST ideal für das Studium von Seltenheitsphänomenen, da Abweichungen und Anomalien in den Daten visuell hervorgehoben werden können. Dies ermöglicht es, die Leistungsfähigkeit und Genauigkeit von Seltenheitsbewertungsalgorithmen unter kontrollierten, aber intuitiv erfassbaren Bedingungen zu testen.

### **3.2.2 Madrid Tripadvisor Rezensionen**

Der „TripAdvisor Dataset for Dyadic Context Analysis“ Datensatz (López-Riobóo Botana u. a., 2022) beinhaltet Bewertungen von Restaurants in sechs Metropolen: London, New York, New Delhi, Paris, Barcelona und Madrid, und bietet eine reichhaltige Quelle für die Analyse von Nutzerinteraktionen. Er besteht aus sechs CSV-Dateien mit numerischen, kategorialen und Textmerkmalen, die sich für vielfältige Forschungsansätze, wie Empfehlungssysteme, Sentimentanalysen und Textpersonalisierung, eignen.

Die gesammelten Daten sind in sechs CSV-Dateien organisiert und enthalten numerische, kategoriale und Textmerkmale wie die Anzahl der Bewertungen, Benutzer-ID, Restaurantname, Bewertungsscore, Review-Titel, vollständiger Review-Text, Veröffentlichungsdatum der Bewertung und die Stadt des Restaurants. Diese Daten wurden exklusiv in englischer Sprache gesammelt und die Teilnehmerdaten wurden anonymisiert. Die Informationen wurden mittels einer Web-Scraper-Software gesammelt, die sowohl das Scrapy Python-Framework als auch das Selenium Webdriver-Testtool nutzt, wie in López-Riobóo Botana u. a. (2022) beschrieben.

Für diese Arbeit werden ausschließlich die vollständigen Rezensionen von Madrid verwendet, da sie eine gute Repräsentation von Textdaten bieten. Die Fokussierung auf die full\_review Spalte des Datensatzes bietet eine prä-

zise Basis für die Untersuchung der Seltenheit von Datenpunkten. In dieser Spalte sind umfassende Textdaten enthalten, die eine Vielzahl von Nutzermeinungen und sprachlichen Ausdrücken abdecken. Durch die Analyse dieser Textdaten auf seltene oder ungewöhnliche Ausdrucksweisen können wertvolle Erkenntnisse über unkonventionelle Meinungen oder Sprachmuster gewonnen werden. Die Untersuchung dieser spezifischen Spalte ermöglicht es, ein tieferes Verständnis dafür zu entwickeln, wie Seltenheit in umfangreichen, textbasierten Datensätzen identifiziert und bewertet werden kann.

### 3.2.3 Audio MNIST

Audio MNIST ist ein spezialisierter Datensatz, der für die Analyse und Anwendung von Deep-Learning-Techniken im Audio-Bereich entwickelt wurde, insbesondere für die Klassifizierung gesprochener Zahlen. Er wurde von Becker u. a. (2023) erstellt und in ihrem Artikel „Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals“ verwendet. Der Datensatz umfasst 30.000 Audio-Beispiele, die jeweils als .wav-Datei gespeichert sind. Diese Beispiele repräsentieren gesprochene Ziffern von 0 bis 9.

Dieser Datensatz ist so strukturiert, dass er sowohl hinsichtlich der Labels/Ziffern als auch der Sprecherinnen und Sprecher ausgeglichen ist. Er enthält Aufnahmen von 60 Sprecher\*innen, wobei jede Person 500 Aufnahmen für verschiedene Ziffern beisteuert, was 50 Aufnahmen pro Ziffer pro Person entspricht. Die Aufnahmen werden so bearbeitet, dass sie möglichst wenig Stille am Anfang und Ende aufweisen, was ihre Eignung für effiziente maschinelle Lernanwendungen erhöht. Die Audio-Dateien sind im wav-Format mit einer Abtastrate von 44.1kHz aufgezeichnet.

Die Relevanz des Audio MNIST-Datensatzes für die Arbeit liegt in seiner spezifischen Struktur und Vielfalt. Audio MNIST bietet eine reiche Sammlung von Audio-Beispielen, die sich ideal für die Erforschung und Implementierung von Techniken zur Bewertung der Seltenheit von Datenpunkten eignen, insbesondere in Bezug auf Audiodaten. Die Daten sind hochdimensional, da

sie aus einer großen Anzahl von Audio-Samples bestehen, die verschiedene Sprecher\*innen, Aussprachen und Aufnahmebedingungen beinhalten. Dadurch ergeben sich vielfältige Möglichkeiten, Methoden zu entwickeln und zu testen, die auf das Erkennen von Mustern und die Identifizierung seltener Merkmale in den Daten abzielen.

Zusammenfassend ist Audio MNIST eine bedeutende Ressource im Bereich der Audio-Signalverarbeitung und des maschinellen Lernens und bietet einen reichhaltigen Datensatz für das Training und Testen von Modellen, die auf die Klassifizierung gesprochener Ziffern fokussiert sind. Es wurde beispielsweise in der Studie von Becker u. a. (2018) verwendet, um zu zeigen, wie man die Entscheidungsfindung von Deep-Learning-Modellen besser verstehen und erklären kann. Die Nutzung von Audio MNIST könnte dazu beitragen, neue Perspektiven und Ansätze in der Analyse und Klassifizierung von Audio-Daten zu entwickeln, insbesondere im Hinblick auf die Erkennung und Bewertung von Seltenheiten in den Daten.

### 3.2.4 Molekulare Daten

Die molekularen Daten, bereitgestellt durch die Global Natural Products Social Molecular Networking (GNPS) Plattform (Wang u. a., 2016), repräsentieren eine umfangreiche Sammlung chemischer Verbindungen, die aus natürlichen Quellen gewonnen wurden. Diese Datenbank ist eine bedeutende Ressource für die chemische und biochemische Forschung, insbesondere für die Analyse und Klassifikation von Naturstoffen. Die GNPS-Plattform fördert auch die Zusammenarbeit und den Datenaustausch innerhalb der wissenschaftlichen Gemeinschaft durch die Bereitstellung von Werkzeugen zur Datenanalyse und -teilung. Dies trägt zur Beschleunigung der Entdeckung und Charakterisierung neuer Naturstoffe bei und erweitert unser Verständnis der chemischen Vielfalt in der Natur (Wang u. a., 2016).

Anwendungsbeispiele der GNPS-Plattform illustrieren ihren vielseitigen Nutzen in der Forschung. So ermöglicht GNPS beispielsweise die Entdeckung neuer bioaktiver Verbindungen durch die Analyse von Metabolomdaten, die

Identifizierung von Biomarkern für Krankheiten oder die Aufklärung der Biosynthesewege von Naturstoffen (Wang u. a., 2016). Ein konkretes Beispiel ist die Nutzung von GNPS für die globale Analyse des biosynthetischen chemischen Raums mariner Prokaryoten, indem sie massenspektrometrische Daten aus verschiedenen Studien zusammenführte und mithilfe von molekularen Netzwerken visualisierte (Wei u. a., 2023). Dies ermöglichte die Identifizierung neuer Naturstoffe und biosynthetischer Stoffwechselwege, die für die Entwicklung neuer Medikamente und Anwendungen in der Biotechnologie relevant sein könnten.

Die Daten enthalten nicht nur Informationen über Eigenschaften und Struktur der Verbindungen, wie Simplified Molecular Input Line Entry System (SMILES) und International Chemical Identifier (InChI) Codes (O’Boyle, 2012), sondern auch Metadaten, die beispielsweise den Typ der Probe, das Jahr der Analyse und die Methode der Sammlung umfassen (Wang u. a., 2016). Diese Identifikatoren sind entscheidend für die präzise Darstellung der molekularen Strukturen, was eine fundamentale Basis für deren Analyse und das Verständnis ihrer Funktionen und Wechselwirkungen bietet.

Der Einsatz von dem Programm *ClassyFire* ermöglichte eine effiziente Strukturierung und Kategorisierung der auf der GNPS-Plattform verfügbaren Daten (Djoumbou Feunang u. a., 2016). Diese Technologie bedient sich einer ausgeklügelten, auf Strukturdaten basierenden chemischen Klassifizierung (ChemOnt), um chemische Substanzen automatisch in über 4800 Kategorien einzurichten, die auf den chemischen Strukturen und Merkmalen der Moleküle basieren, was ein wesentlicher Schritt in der Datenvorbereitung ist. Diese innovative chemische Klassifizierung umfasst verschiedene Ebenen – von ’kingdom’ bis ’subclass’ –, die alle durch klare, algorithmisch erfassbare Regeln charakterisiert sind. Zudem wird für jede dieser Kategorien eine einheitliche, auf Konsens basierende Benennung verwendet und eine Beschreibung basierend auf den typischen strukturellen Merkmalen der enthaltenen Substanzen bereitgestellt. Mit dem Zugang zum *ClassyFire*-Webdienst und der Ruby API können Forschende effizient bestimmte chemische Stoffe oder

Stoffklassen identifizieren und analysieren. Für die Kategorisierung der Daten wird entweder ein SMILES- oder ein InChI-String als Eingabe erwartet.

### 3.3 Defintion von Seltenheit

Die Herausforderung, eine allgemeingültige Definition von Seltenheit für die Zwecke dieser Forschungsarbeit zu etablieren, wurde bereits in Kapitel 2.5 dargelegt. Diese Schwierigkeit ergibt sich primär aus der Tatsache, dass Seltenheit ein kontextabhängiges Phänomen ist. Ein universeller Ansatz zur Definition von Seltenheit erscheint daher nicht praktikabel, insbesondere angesichts der variierenden Formate und Eigenschaften verschiedener Datensätze. Um dieser Problematik zu begegnen, basiert die vorliegende Arbeit auf der Prämisse, dass eine Vorverarbeitung der Daten stattgefunden hat, bevor eine Bewertung der Seltenheit vorgenommen wird.

Ein zentraler Aspekt dieser Vorverarbeitung ist die Annahme, dass eine Ähnlichkeitsmatrix vorhanden ist, die für jedes Datenpunktpaar mithilfe einer Distanzmetrik (siehe Abschnitt 2.2) deren Ähnlichkeit abbildet. Eine Ähnlichkeitsmatrix stellt eine zweidimensionale Tabelle dar, in der die (Un)Ähnlichkeit zwischen zwei Elementen a und b einer gegebenen Sequenz gemessen wird (Rafii u. Pardo, 2012). Die Existenz einer solchen Matrix impliziert, dass die Definition der Ähnlichkeit in diesem spezifischen Kontext durch die vorangegangene Datenverarbeitung indirekt bestimmt wird. So eine Datenrepräsentation erleichtert die Quantifizierung der Seltenheit eines Datenpunktes relativ zu anderen Punkten im Datensatz. Folglich konzentriert sich die Entwicklung neuer Verfahren zur Bewertung der Seltenheit in dieser Arbeit auf die Nutzung dieser Ähnlichkeitswerte.

Die Konzeption dieser Methoden beruht auf dem Verständnis, dass die Ähnlichkeitsmatrix nicht nur ein Werkzeug zur Messung von Distanzen oder Ähnlichkeiten zwischen Datenpunkten ist, sondern auch eine fundamentale Basis für die Einschätzung ihrer Seltenheit bietet. Indem sie die relationalen Eigenschaften zwischen den Datenpunkten hervorhebt, ermöglicht die Matrix eine

differenzierte Betrachtung von Seltenheit, die über einfache frequenzbasierte Ansätze hinausgeht. Diese Herangehensweise vereinfacht die Entwicklung der Seltenheitsbewertung, da die Definition der Seltenheit damit schon gegeben ist.

## 3.4 Datenvorbereitung

Vor der Entwicklung der Methoden zur Seltenheitsbewertung wird eine umfassende Datenpräparation durchgeführt, um die Basis für die weiterführende Analyse und Modellierung zu schaffen. Diese Vorbereitung beinhaltet vielfältige Techniken und Ansätze, die entsprechend den speziellen Anforderungen der verschiedenen Datentypen angepasst sind. Die verwendeten Werkzeuge werden im darauffolgenden Kapitel 4.1 aufgelistet.

Bei der Vorbereitung von Bilddaten, wie beispielweise im MNIST Datensatz, wird ein Deep Learning Ansatz verfolgt, der darauf abzielt, durch die Verwendung von neuronalen Netzwerken, die speziell für die Bildanalyse konzipiert sind, tiefergehende Merkmale der Daten zu extrahieren (LeCun u. a., 2015). Diese Netzwerke sind darauf spezialisiert, relevante Muster und Strukturen innerhalb der Bilddaten zu identifizieren und lernen selbstständig, die Daten in eine kompakte, aussagekräftige Repräsentation zu überführen. Diese Repräsentation, oft in Form von hochdimensionalen Vektoren, ermöglicht eine effiziente Ähnlichkeitsbewertung und Klassifizierung der Daten.

Für Textdaten wird ein moderner Ansatz zur Verarbeitung und Analyse angewendet, der es ermöglicht, die Texte in numerische Repräsentationen zu transformieren (Honnibal u. a., 2020). Diese numerischen Vektoren fassen die semantische Bedeutung der Wörter oder Textabschnitte zusammen und ermöglichen es, komplexe sprachbasierte Analysen durchzuführen. Der Einsatz fortschrittlicher Verarbeitungsmethoden erleichtert die Erkennung und Nutzung linguistischer Muster und Charakteristika, die für die Untersuchung der Datensätze entscheidend sind.

Im Bereich der Audiodatenanalyse werden Techniken eingesetzt, die darauf abzielen, charakteristische Merkmale aus den Audiosignalen zu extrahieren (McFee u. a., 2023). Diese Merkmale, wie etwa spezifische Klangmuster oder Frequenzspektren, bieten eine fundierte Basis für die vergleichende Analyse von Audiodaten. Durch die Transformation der Audiodaten in eine standardisierte Form wird es möglich, unterschiedliche Aufnahmen objektiv zu vergleichen und ihre Einzigartigkeit oder Ähnlichkeit auf Basis der extrahierten Merkmale zu bewerten.

In der Arbeit wird bei der Verarbeitung von Molekulardaten direkt auf vorverarbeitete Ähnlichkeitsinformationen zurückgegriffen, die in einer Pickle-Datei bereitgestellt werden. Eine Pickle-Datei ist ein spezifisches Dateiformat in Python, das es ermöglicht, Python-Objekte zu serialisieren und zu deserialisieren (Van Rossum, 2020). Neben den Ähnlichkeitsinformationen beinhaltet die Datenbasis ebenfalls Strukturinformationen und Klassifizierungen der Moleküle, die bereits im Vorfeld mithilfe von *ClassyFire* erstellt und bereitgestellt wurden, um eine umfassende Analyse der chemischen Eigenschaften und Kategorien zu unterstützen.

# 4 Entwicklung

Die Berechnung des Seltenheitswerts wird so entwickelt, dass unabhängig von der gewählten Distanzmetrik konsistente Seltenheitswerte erzeugt werden können. Dies ermöglicht eine flexible Anwendung verschiedener Metriken, während gleichzeitig gewährleistet wird, dass die resultierenden Seltenheitswerte eine zuverlässige und vergleichbare Quantifizierung der Einzigartigkeit von Datenpunkten innerhalb eines hochdimensionalen Datensatzes bieten.

## 4.1 Entwicklungsumgebung und Pakete

Python 3.11.2 dient als Basis für die programmtechnische Umsetzung, charakterisiert durch seine Effizienz und Anpassungsfähigkeit in der wissenschaftlichen Berechnung und Datenanalyse (Ranjan u. a., 2023). Die Integration spezifischer Pakete ermöglicht eine umfassende Unterstützung verschiedener Aspekte der Datenverarbeitung und Modellierung.

Für die Handhabung und Manipulation von Arrays wird *numpy* in der Version 1.25.1 verwendet. Dieses Paket bietet umfassende Funktionen für die Arbeit mit großen, mehrdimensionalen Arrays und Matrizen sowie eine Vielzahl von mathematischen Operationen zur Bearbeitung dieser Datenstrukturen (Harris u. a., 2020).

*TensorFlow* in der Version 2.14.0, ein umfassendes Framework für maschinelles Lernen und neuronale Netzwerke (Abadi u. a., 2016), wird speziell für den Aufbau von CNNs mit der *Keras API* eingesetzt. Dies ermöglicht eine

effiziente Entwicklung, Training und Validierung komplexer Modelle (Chollet u. a., 2015). Die Verwendung von Keras vereinfacht die Modellierung durch eine hochgradig modulare und intuitive Schnittstelle, die den Aufbau von unterschiedlichen Netzwerkarchitekturen erleichtert.

*Librosa*, ein spezialisiertes Python-Paket für die Musik- und Audioanalyse, spielt eine wesentliche Rolle in der Arbeit, die akustische Daten verarbeiten. Es bietet umfangreiche Funktionen für die Audioverarbeitung, Feature-Extraktion und die Erstellung von Audiobezogenen Visualisierungen (McFee u. a., 2023).

Die Analyse und Berechnung von Distanzen zwischen Datenpunkten, ein kritischer Schritt bei der Erstellung von Ähnlichkeitsmatrizen, erfolgt durch die Verwendung von *scipy* in der Version 1.11.3 (Virtanen u. a., 2020). Dieses Paket stellt fortschrittliche mathematische Funktionen zur Verfügung, die für die Berechnung der paarweisen Abstände in hochdimensionalen Räumen erforderlich sind, und unterstützt damit die Konstruktion von Ähnlichkeitsmatrizen, die für die Bewertung der Seltenheit von Datenpunkten unerlässlich sind.

Für die Vektorisierung von Textdaten wird zusätzlich *spaCy* in der Version 3.7.2 verwendet, ein fortschrittliches Paket für die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) (Honnibal u. a., 2020). *spaCy* bietet effiziente und leicht zugängliche Methoden für die Tokenisierung, das Tagging, das Parsen sowie die Entitätserkennung und ermöglicht es, umfangreiche Textdaten in dichte Vektoren zu transformieren, die für maschinelles Lernen und tiefgehende linguistische Analysen genutzt werden können (Honnibal u. a., 2020). Diese Vektorisierung ist entscheidend, um Textdaten in einem Format zu repräsentieren, das für die anschließende Distanzberechnung geeignet ist.

Zur Visualisierung von Daten und Ergebnissen wird auf *matplotlib* 3.8.0 zurückgegriffen, das eine breite Palette von Werkzeugen für die Erstellung von

Grafiken bietet, die sowohl statisch als auch interaktiv sein können (Hunter, 2007). Diese Werkzeuge sind entscheidend für die Darstellung und Analyse der Forschungsergebnisse.

Für die Reduktion der Dimensionalität und die Visualisierung der Daten wird *scikit-learn* 1.3.1 eingesetzt (Pedregosa u. a., 2011). Dieses Tool ist besonders leistungsfähig, um komplexe Datensätze in einer niedrigeren Dimension darzustellen und zu analysieren, was ein tieferes Verständnis der Datenstrukturen und deren Seltenheitsmerkmale ermöglicht.

Zusätzlich wird für die spezialisierten Aufgaben der Ausreißerdetektion die *PYOD* 1.1.1 Bibliothek verwendet (Zhao u. a., 2019). Diese Bibliothek ist führend in der Implementierung verschiedener Algorithmen für die Erkennung von Ausreißern und ermöglicht eine effiziente und effektive Identifizierung von Anomalien in Daten, was für die zuverlässige Bewertung von Seltenheitswerten entscheidend ist.

Darüber hinaus wird die *RDKit*-Bibliothek in der Version 2023.9.4 zur Visualisierung und Berechnung von Fingerabdrücken molekularer Daten eingesetzt (Landrum u. a., 2024). *RDKit* ist ein essenzielles Werkzeug in der chemoinformatischen Datenverarbeitung, das leistungsfähige Funktionen für die Erstellung, Manipulation und Analyse molekularer Strukturen bietet. Insbesondere wird es für die Generierung von Molekülfingerabdrücken verwendet, die kritisch für die Analyse und Vergleich von molekularen Datensätzen sind.

Es ist wichtig zu betonen, dass für die Kernmethoden der Bewertung der Seltenheit von Datenpunkten lediglich *numpy* erforderlich ist. Dieses Paket ist essenziell, während die anderen Pakete vorrangig in der Entwicklungs- und Prototyping-Phase zur Unterstützung eingesetzt werden. Für die Seltenheitsbewertung mit Hilfe von Ausreißermethoden wird *PYOD* benötigt. Zudem wurden im Entwicklungsprozess auch kleinere, weniger zentrale Pakete genutzt. Diese dienten ebenfalls der Vereinfachung und boten unterstützende

Funktionen, die die Entwicklung und das Prototyping erleichterten.

## 4.2 Nächste-Nachbarn-Methode

In diesem Abschnitt wird der Algorithmus zur Berechnung von Seltenheitswerten basierend auf Distanzen und einer vorgegebenen Anzahl an Nachbarn beschrieben, der fortan als Nächste-Nachbarn-Methode (NNM) bezeichnet wird. Der Gedanke, sowohl lokale als auch globale Seltenheit zu bewerten, führt zu einem differenzierten Verständnis der Datenstruktur. Lokale Seltenheit bezieht sich auf die Unmittelbarkeit und Isolation eines Punktes in Bezug auf seine nächsten Nachbarn. Ein Punkt kann lokal selten sein, wenn er weit entfernt von seinen nächsten Nachbarn liegt, was auf einen potenziellen Ausreißer hindeuten könnte. Globale Seltenheit hingegen betrachtet, wie selten ein Punkt im Vergleich zur Gesamtheit aller Datenpunkte ist. Dies beinhaltet die Bewertung der Dichte und Verteilung der Punkte im gesamten Datensatz. Durch die Kombination beider Perspektiven – die Einschätzung der Nähe eines Punktes zu seinen Nachbarn und die allgemeine Verteilung der Punkte – ermöglicht der NNM eine umfassende Analyse von Seltenheitswerten. Dieser Ansatz erlaubt es, nuancierte Einsichten in die Daten zu gewinnen, indem er nicht nur offensichtliche Ausreißer identifiziert, sondern auch subtilere Muster von Seltenheit und Dichte aufdeckt.

Der Algorithmus durchläuft mehrere Phasen, um die Seltenheitswerte für jeden Punkt in einem Datensatz zu bestimmen. Zuerst wird überprüft, ob die Eingabeliste der Distanzmatrix leer ist, um sicherzustellen, dass Berechnungen durchführbar sind. Anschließend wird jeder Wert der Distanzmatrix validiert, dass es sich um numerische Werte in Form von ganzen Zahlen, Fließkommazahlen oder NumPy-spezifischen numerischen Typen handelt. Für die Distanzmatrix wird dann die Anzahl der zu berücksichtigenden Nachbarn auf die tatsächlich vorhandene Anzahl von Datenpunkten begrenzt, um Fehler durch eine zu hohe Nachbarzahl zu vermeiden. Die Distanzen jedes Punktes zu seinen Nachbarn werden sortiert, und die durchschnittliche Distanz zu

den nächsten  $n$  Nachbarn wird berechnet. Diese durchschnittlichen Distanzen werden anschließend normalisiert, um Seltenheitswerte zwischen 0 und 1 zu erhalten, wobei 0 für die am wenigsten seltenen und 1 für die seltensten Punkte steht.

Sollten alle Punkte dieselbe durchschnittliche Distanz aufweisen, was bedeutet, dass alle Punkte als gleich selten oder gleich häufig angesehen werden, wird jedem Punkt ein Seltenheitswert von 0 zugewiesen. Andernfalls werden die Seltenheitswerte so skaliert, dass sie in den Bereich von 0 bis 1 fallen, wobei höhere Werte eine größere Seltenheit anzeigen.

Diese Methode liefert eine Liste von Seltenheitswerten für jeden Punkt im Datensatz, welche zur Identifizierung von Ausreißern oder zur Bewertung der Verteilung von Datenpunkten in einem multidimensionalen Raum genutzt werden können. Der Algorithmus ist so gestaltet, dass er effizient mit großen Datensätzen umgehen kann und dabei präzise Einsichten in die Seltenheit von Datenpunkten bietet.

---

**Algorithmus 1** Nächste-Nachbarn-Methode

---

**Eingabe:** Eine Distanzmatrix

**Ausgabe:** Eine Liste von Seltenheitswerten

Seltenheitsscores  $\leftarrow$  leere Liste

**WENN** Distanzmatrizen sind leer **DANN**

**Rückgabe:** Fehler: „Keine Berechnung möglich, da keine Distanzen vorhanden sind“

**ENDE WENN**

**FÜR ALLE** Reihen in Distanzmatrix **MACHE**

**FÜR ALLE** Distanzen in Matrix **MACHE**

**WENN** Distanz ist nicht numerisch **DANN**

**Rückgabe:** Fehler: „Alle Distanzen müssen numerische Werte sein“

**ENDE WENN**

**ENDE FÜR**

gültigeNachbarn  $\leftarrow$  Minimum(AnzahlNachbarn, Anzahl der Elemente in Matrix - 1)

sortierteDistanzen  $\leftarrow$  Sortiere die ersten gültigeNachbarn Distanzen jeder Zeile

durchschnittlicheDistanzen  $\leftarrow$  Berechne den Durchschnitt der sortiertenDistanzen

minDistanz  $\leftarrow$  Minimum der durchschnittlichen Distanzen

maxDistanz  $\leftarrow$  Maximum der durchschnittlichen Distanzen

**WENN** maxDistanz gleich minDistanz **DANN**

Seltenheitswert  $\leftarrow$  Erstelle eine Liste von Nullen der Länge der durchschnittlichen Distanzen

**SONST**

Seltenheitswert  $\leftarrow$  Normalisiere die durchschnittlichen Distanzen zwischen 0 und 1

**ENDE WENN**

Füge Seltenheitswert zu Seltenheitswerten hinzu

**ENDE FÜR**

**Rückgabe:** Seltenheitswerte

---

## 4.3 Flußsuch-Methode

Zudem wird ein Algorithmus zur Berechnung von Seltenheitswerten entwickelt, der auf Distanzen, Fluss und der Anzahl der nächsten Knotenpunkte (Hubs) basiert, der fortan als Flow bezeichnet wird. Die Entwicklung des Flow-Algorithmus zur Berechnung von Seltenheitswerten beruht auf der Idee, die Struktur von Netzwerken und den Fluss innerhalb dieser Netzwerke als Metapher zur Identifizierung von Seltenheit in Datensätzen zu nutzen. Im Kern dieses Ansatzes steht die Vorstellung, dass Datenpunkte innerhalb eines Datensatzes ähnlich wie Knoten in einem Netzwerk funktionieren, wobei die Distanzen zwischen ihnen die Verbindungen oder „Flüsse“ zwischen den Knoten darstellen. Diese Perspektive ermöglicht es, die Konnektivität und die relationalen Eigenschaften der Datenpunkte in einer Weise zu analysieren, die über herkömmliche Distanzmaße hinausgeht.

Die Inspiration hinter dem Flow-Algorithmus leitet sich von der Beobachtung ab, dass in einem Netzwerk nicht alle Knoten gleich geschaffen sind. Einige Knoten dienen als zentrale Hubs, die viele Verbindungen aufweisen und eine wichtige Rolle in der Kommunikation oder im Fluss innerhalb des Netzwerks spielen. Andere Knoten sind möglicherweise peripherer und haben weniger Verbindungen, was sie isolierter macht. Diese Netzwerkdynamik lässt sich auf Datensätze übertragen, indem man die Distanzen zwischen den Datenpunkten als Flüsse betrachtet, die je nach Entfernung stärker oder schwächer sein können.

Der Algorithmus besteht aus mehreren Phasen: Zunächst erfolgt die Überprüfung der Gültigkeit und numerischen Werte der Distanzmatrix. Anschließend wird festgelegt, wie viele der nächsten Knotenpunkte berücksichtigt werden sollen, und überprüft, ob diese Anzahl die Größe der Distanzmatrix nicht überschreitet.

Für jede Distanz in der Matrix wird ein Flusswert berechnet, der mit einem Abklingparameter (decay) skaliert wird. Dieser Flusswert reflektiert die Konnektivität jedes Datenpunkts, wobei größere Distanzen zu einer exponentiellen Abnahme des Flusses führen. Die Distanzen werden dann sortiert, um

die Indizes in aufsteigender Reihenfolge zu erhalten. Für jeden Datenpunkt wird der Fluss zu den nächsten  $n\_next\_hubs$  Knotenpunkten aggregiert, indem die Flüsse summiert werden.

Die resultierenden Flusswerte werden anschließend normalisiert, um die Seltenheitswerte zu erhalten. Dies erfolgt durch Skalierung der Flusswerte auf ein Intervall von 0 bis 1, wobei der minimale Flusswert auf 0 und der maximale auf 1 gesetzt wird. Wenn alle Flusswerte identisch sind, wird der Seltenheitswert für alle Punkte auf 0 gesetzt, um zu kennzeichnen, dass alle Punkte gleich selten oder gleich häufig sind. Andernfalls wird jeder Flusswert so skaliert, dass er zwischen 0 (für den häufigsten Punkt) und 1 (für den seltensten Punkt) liegt.

Zum Abschluss gibt der Algorithmus die Liste der normalisierten Seltenheitswerte für die gesamte Datenmenge zurück. Diese Werte können dann genutzt werden, um seltene oder ungewöhnliche Datenpunkte innerhalb des Datensatzes zu identifizieren. Der Algorithmus ist darauf ausgelegt, effizient mit großen Datensätzen umzugehen, indem er die Berechnungen auf die wichtigsten Konnektivitätsmerkmale beschränkt und dabei präzise Einsichten in die Seltenheit der Datenpunkte liefert.

---

**Algorithmus 2** Flow Methode

---

**Eingabe:** Distanzmatrix, Anzahl nächster Knotenpunkte (`n_next_hubs`), Zerfallsparameter (`decay`, Standard: 10)

**Ausgabe:** Seltenheitswerte

**WENN** Distanzmatrix ist leer **DANN**

**Rückgabe:** Fehler: „Keine Berechnung möglich, da keine Distanzen vorhanden sind“

**ENDE WENN**

**WENN** `n_next_hubs` größer als Anzahl der Distanzen **DANN**

**Rückgabe:** Fehler: „`n_next_hubs` größer als Anzahl der Distanzen“

**ENDE WENN**

Initialisiere ein Nullen-Array für die Flussergebnisse

**FÜR** jeden Datenpunkt in der Distanzmatrix **MACHE**

sortierteIndizes  $\leftarrow$  Sortiere die Distanzen des Datenpunkts und erhalte die sortierten Indizes

nächsteKnotenIndizes  $\leftarrow$  Wähle die Indizes der `n_next_hubs` nächsten Knotenpunkte aus sortierteIndizes, beginnend beim zweiten Index

flussWerte  $\leftarrow$  Berechne die Flusswerte für die Distanzen an nächsteKnotenIndizes, wende  $e^{-\text{decay} \times \text{Distanz}}$  an

flussSumme  $\leftarrow$  Summiere die flussWerte, um das Flussergebnis für den Datenpunkt zu erhalten

speichere flussSumme im entsprechenden Array für Flussergebnisse

**ENDE FÜR**

Seltenheitswert  $\leftarrow$  Normalisiere die Flussergebnisse zwischen 0 und 1

**Anpassung:** Bei identischen Flusswerten setze alle Seltenheitswerte auf 0.

**Rückgabe:** Normalisierte Seltenheitswerte

---

## 4.4 Ausreißer-Methoden

Wie bereits im Abschnitt 2.3 beschrieben, bietet PyOD eine Vielzahl von Methoden zur Anomalieerkennung, die in diesem Kontext als Verfahren zur Bestimmung von Seltenheitswerten adaptiert werden. Besonderes Augenmerk liegt dabei auf distanzbasierten Methoden, da diese für die Verarbeitung der gegebenen Ähnlichkeitsmatrizen besonders geeignet sind. Ein Vorteil dieser Methoden ist in der Anomalieerkennung eher ein Nachteil, da sie nicht direkt Ausreißer erkennen, sondern lediglich eine Wahrscheinlichkeit dafür angeben, dass ein Punkt ein Ausreißer ist (Muhr u. a., 2023). Unter Berücksichtigung der Effizienz und Relevanz für hochdimensionale Daten wurden drei spezifische Methoden ausgewählt: K-Nearest Neighbors (KNN) (Ramaswamy u. a., 2000), LOF (Breunig u. a., 2000), Stochastic Outlier Selection (SOS) (Janssens, 2013).

Die Wahl dieser spezifischen Methoden liegt in ihrer herausragenden Kompetenz als distanzbasierte Verfahren begründet. Methoden wie KNN, LOF und SOS sind besonders effektiv im Umgang mit Ähnlichkeitsmatrizen, da sie in der Lage sind, die Beziehungen zwischen Datenpunkten durch Distanz- und Dichtemaße zu interpretieren. Diese Fähigkeit ermöglicht es ihnen, Anomalien und seltene Datenpunkte präzise zu identifizieren, indem sie die strukturellen Eigenschaften innerhalb der Ähnlichkeitsmatrizen nutzen, um tiefere Einblicke in die Daten zu gewähren.

Allerdings wurde die Methode KNN aufgrund ihrer hohen Rechenintensität und der damit verbundenen langsamen Ausführungszeit (Song u. a., 2022) von der weiteren Betrachtung ausgeschlossen. Die Rechenintensität von KNN steigt mit der Anzahl der Datenpunkte im Datensatz, was in hochdimensionalen und umfangreichen Datensätzen zu erheblichen Performanzproblemen führen kann. Im Gegensatz dazu bieten LOF und SOS einen ausgewogenen Ansatz zwischen Rechenzeit und Genauigkeit der Anomalieerkennung, was sie für die Anwendung in dieser Arbeit geeignet macht. Ergänzend lässt sich anführen, dass KNN in seiner Funktionsweise Parallelen zu NNM aufweist,

jedoch ist NNM in einer Weise implementiert worden, die eine effizientere Berechnung ermöglicht.

Die LOF-Methode basiert auf der Idee, die lokale Dichteabweichung eines Datenpunktes von seinen Nachbarn zu messen. Dabei wird die Anzahl der Nachbarn ( $n_{neighbors}$ ) festgelegt, um die lokale Dichte zu bestimmen (Breunig u. a., 2000). Zur Berechnung der Ausreißer-Wahrscheinlichkeiten *outlier\_scores* mittels LOF wird die Kosinus Distanz als Metrik genutzt, um die Unterschiede in der Dichte zwischen den Datenpunkten und ihren Nachbarn zu quantifizieren. Die Ausreißer-Werte werden anschließend auf einen Bereich von 0 bis 1 normalisiert, um die Seltenheitswerte zu generieren.

SOS verwendet die Perplexität als Maß für die effektive Anzahl von Nachbarn, ähnlich dem Parameter  $k$  im kNN-Algorithmus (Janssens, 2013). Die Wahl der Perplexität spielt eine entscheidende Rolle, da sie direkt die Empfindlichkeit der SOS-Methode bei der Identifikation von Ausreißern beeinflusst. Für SOS wird die euklidische Distanz verwendet, da Kosinus nicht wählbar ist. Zuletzt werden auch diese Werte von 0 bis 1 normalisiert.

# 5 Hypothese und Ergebnisse

Das Kapitel dient als Kernstück der wissenschaftlichen Untersuchung, indem es die theoretischen Annahmen der Studie mit den empirischen Befunden verknüpft. Dies ermöglicht, formulierte Hypothesen anhand von konkret gesammelten Daten zu überprüfen und somit die Validität der entwickelten Methoden zur Bewertung der Seltenheit in hochdimensionalen Datensätzen zu testen.

## 5.1 Hypothesen

Für diese Arbeit wurden der MNIST-Datensatz und der Audio MNIST-Datensatz ausgewählt, da sie exemplarische Szenarien bieten, in denen die Seltenheit von Datenpunkten auf intuitiv verständliche Weise dargestellt und untersucht werden kann. Es wird postuliert, dass ungewöhnlich oder atypisch geschriebene Ziffern innerhalb des MNIST Datensatzes seltener vorkommen als standardmäßig geschriebene Zahlen. Diese Seltenheit manifestiert sich in Abweichungen von der Norm, wie beispielsweise unkonventionellen Schleifen bei der Ziffer „8“ oder ungewöhnlichen Proportionen bei der „3“. Ähnlich verhält es sich mit dem Audio MNIST-Datensatz, der eine Sammlung von Sprachaufnahmen beinhaltet, in denen Zahlen von verschiedenen Sprecher\*innen ausgesprochen werden. In diesem Kontext wird angenommen, dass Aufnahmen, in denen Zahlen undeutlich artikuliert werden oder die Stimme des Sprechers oder der Sprecherin signifikant von der Mehrheit abweicht, als selten betrachtet werden können.

Neben dem MNIST- und dem Audio MNIST-Datensatz wurden ursprünglich auch ein Textdatensatz und ein Molekulardatensatz für die Untersuchung ausgewählt. Es stellte sich jedoch heraus, dass die Bewertung der Seltenheit in diesen Datensätzen nicht so unmittelbar nachvollziehbar ist wie in den visuellen und auditiven Beispielen. Dies liegt vor allem an der komplexeren Natur der Daten und der Schwierigkeit, Seltenheit in Form von Textvariationen oder molekularen Strukturen ohne umfassende domänenspezifische Kenntnisse zu identifizieren und zu quantifizieren.

Im Falle des Textdatensatzes ist die Herausforderung, Seltenheit zu definieren und zu messen, eng mit linguistischen und semantischen Aspekten verbunden. Während bestimmte Wörter oder Phrasen innerhalb eines Korpus als selten gelten können, hängt ihre Identifizierung stark vom Kontext, der Genre-spezifischen Verwendung und der Sprache selbst ab. Ähnlich komplex gestaltet sich die Situation bei dem Molekulardatensatz, wo Seltenheit durch eine Vielzahl von Faktoren bestimmt wird, einschließlich der chemischen Zusammensetzung, der räumlichen Struktur und der funktionellen Eigenschaften der Moleküle. Die Bewertung dieser Formen der Seltenheit erfordert spezialisierte Kenntnisse und Methoden, die über allgemeine Ansätze zur Datenanalyse hinausgehen.

Angesichts dieser Herausforderungen wurde entschieden, den Fokus der initialen Experimente auf den MNIST- und Audio MNIST-Datensatz zu legen, um die entwickelten Methoden zur Bewertung der Seltenheit in einem besser verständlichen Rahmen zu testen. Dieser Ansatz ermöglicht es, die Wirksamkeit der Methoden in klar definierten und intuitiv zugänglichen Kontexten zu validieren, bevor sie auf komplexere Datensätze angewandt werden. Nach Abschluss der Experimente mit diesen beiden Datensätzen ist jedoch geplant, die erprobten Methoden auf den Text- und Molekulardatensatz anzuwenden. Ziel ist es, die Techniken zur Erfassung und Bewertung von Seltenheitswerten so zu verfeinern, dass sie auch in diesen anspruchsvoller Domänen eingesetzt werden können. Durch diese schrittweise Erweiterung der Anwendungsbereiche soll nicht nur die Vielseitigkeit der entwickelten Methoden demonstriert,

sondern auch ein Beitrag zur Erforschung der Seltenheit in verschiedenen Datenformen geleistet werden.

Die Hypothese, dass sowohl im MNIST als auch im Audio MNIST-Datensatz die Seltenheit von Datenpunkten durch atypische Merkmale gekennzeichnet ist, bildet die Grundlage für eine Reihe von Experimenten. Neben dieser Hypothese, wird eine weitere Hypothese aufgestellt, die die dynamische Natur der Seltenheit in Datensätzen betrachtet. Diese zusätzliche Hypothese behauptet, dass die künstliche Verkleinerung eines Datensatzes – beispielsweise durch die Reduktion bestimmter Kategorien wie der Ziffer „2“ im MNIST-Datensatz oder der Aufnahmen von Stimme „2“ im Audio MNIST-Datensatz – die Seltenheit dieser spezifischen Datenpunkte innerhalb des Gesamtdatensatzes erhöhen würde. Diese Hypothese beruht auf der Annahme, dass die Frequenz und Verteilung von Datenpunkten maßgebliche Faktoren für die Bestimmung ihrer Seltenheit sind. Durch die experimentelle Reduktion der Häufigkeit bestimmter Datenpunkte sollen die Auswirkungen auf deren Seltenheitswert untersucht werden. Dies bietet eine direkte Methode zur Überprüfung der Wirksamkeit der entwickelten Bewertungsverfahren, indem beobachtet wird, ob und wie diese Methoden eine Veränderung in der Seltenheit der Datenpunkte erkennen können, die künstlich seltener gemacht wurden.

## 5.2 Experimente

Eine effektive Strategie zur Beschleunigung und Effizienzsteigerung dieser Analysen besteht in der initialen Reduktion der Datensätze. Dieser Prozess ermöglicht es, die Komplexität der Daten zu verringern, ohne dabei signifikante Informationen zu verlieren. Zur Visualisierung der Datenstruktur und zur Gewährleistung einer intuitiven Übersicht, wie in 2.1 erwähnt, wird t-SNE eingesetzt (siehe Abb. 5.1). Dieser Ansatz bietet einen tiefen Einblick in die Verteilung der Datenpunkte und erleichtert die Identifizierung von Mustern und Anomalien.

Konkret werden diese Methoden auf Teilsätze des MNIST- und des Audio

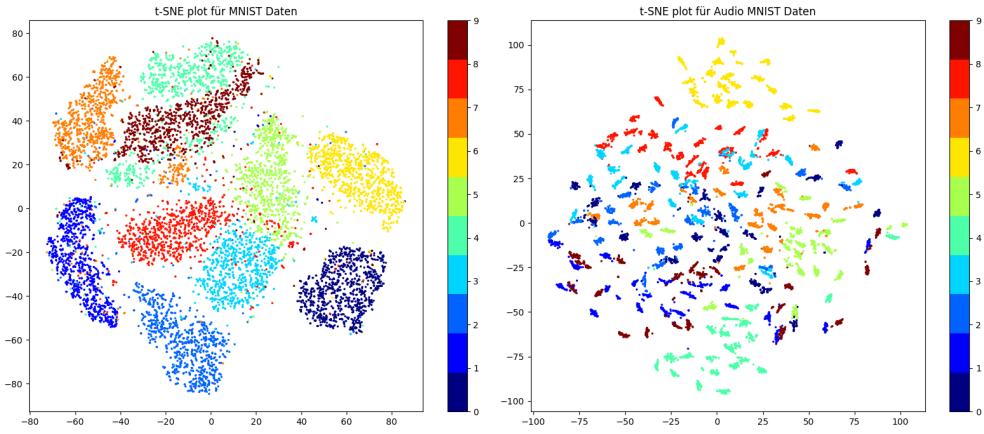


Abbildung 5.1: Links: t-SNE Darstellung der MNIST Teildaten gefärbt nach Ziffer. Rechts: t-SNE Darstellung der Audio MNIST Teildaten gefärbt nach Ziffer

MNIST-Datensatzes angewandt, um die Anwendbarkeit und Effektivität der entwickelten Verfahren zur Seltenheitsbewertung zu überprüfen. Nach der Anwendung dieser Methoden werden die resultierenden Seltenheitswerte sowohl im t-SNE-Plot als auch in Säulendiagrammen visualisiert. Diese Darstellung ermöglicht eine direkte Vergleichbarkeit der Ergebnisse und liefert wertvolle Erkenntnisse über die Verteilung der Seltenheitswerte innerhalb der Datensätze. Besonders die Betrachtung der als „selten“ identifizierten Datenpunkte bietet eine visuelle Bestätigung der Effektivität der Methoden und ermöglicht eine erste Evaluation der unterschiedlichen Ansätze.

In der Fortführung der Studie wird ein Ansatz verfolgt, der flexibel auf die Ergebnisse vorangegangener Visualisierungen reagiert. Statt feste Ziffern für die Experimente vorzugeben, werden die zu reduzierenden Ziffern basierend auf den Einsichten ausgewählt, die aus den initialen Datenvisualisierungen gewonnen wurden. Diese Methode stellt sicher, dass die Experimente auf den MNIST-Datensätzen maßgeschneidert und relevant sind. Konkret werden drei separate Experimente für den MNIST-Datensatz konzipiert, bei denen jeweils eine Ziffer reduziert wird. Die Auswahl dieser Ziffern orientiert sich an ihrer relativen Seltenheit, wie sie durch frühere Analysen und Visualisierun-

gen identifiziert wurde: Eine Ziffer, die als relativ selten gilt, eine als nicht selten betrachtete und eine, die einen mittleren Seltenheitsgrad aufweist.

Diese methodische Herangehensweise wird analog auf den Audio MNIST-Datensatz übertragen. Auch hier werden drei Experimente durchgeführt, die auf der vorherigen Analyse der Daten basieren. Die Experimente umfassen die Reduktion von Audiodateien für drei ausgewählte Ziffern, die analog zum visuellen Datensatz nach ihrer Seltenheit ausgewählt werden: eine für ihre Seltenheit bekannte Ziffer, eine üblicherweise als häufig eingestufte und eine im mittleren Seltenheitsbereich liegende.

Zusätzlich zu den Experimenten mit Ziffern werden beim Audio MNIST-Datensatz drei weitere Experimente durchgeführt, die sich auf die Reduktion von Audiodateien bestimmter Sprecherinnen und Sprecher konzentrieren. Diese Auswahl basiert ebenfalls auf der vorherigen Auswertung der Daten, wobei eine Person, deren Aufnahmen als selten gelten, eine Person mit häufigeren Aufnahmen und eine mit einem mittleren Seltenheitsgrad ausgewählt werden. Diese gezielten Experimente ermöglichen eine differenzierte Untersuchung der Seltenheitsdynamik, indem sie zeigen, wie die Veränderung der Datensatzgröße die Seltenheitswerte beeinflusst. Diese Vorgehensweise gewährleistet, dass die Experimente sowohl für den visuellen als auch für den auditiven Datensatz maßgeschneidert sind und relevante Einblicke in die Natur der Seltenheit innerhalb dieser spezifischen Datenkontexte liefern.

Durch den angepassten experimentellen Ansatz ist es möglich, die Effektivität der entwickelten Methoden zur Bewertung von Seltenheit umfassend zu testen. Die sorgfältig geplante Reduktion bestimmter Ziffern oder Sprecher\*innen in den Datensätzen dient als praxisnahe Experiment, um die Kernhypothese zu überprüfen: Es wird erwartet, dass eine Reduktion der Daten zu einem Anstieg der Seltenheit führt. Zur präzisen Quantifizierung dieses Effekts werden die Mittelwerte der Seltenheitswerte der spezifisch reduzierten Ziffern nach jeder Reduktion mit den Mittelwerten der gleichen Instanzen im ursprünglichen, nicht reduzierten Datensatz verglichen.

## 5.3 Resultate

Hier werden die Resultate der durchgeführten Experimente präsentiert und interpretiert, um die zuvor formulierten Hypothesen zu überprüfen.

### 5.3.1 MNIST und Audio MNIST

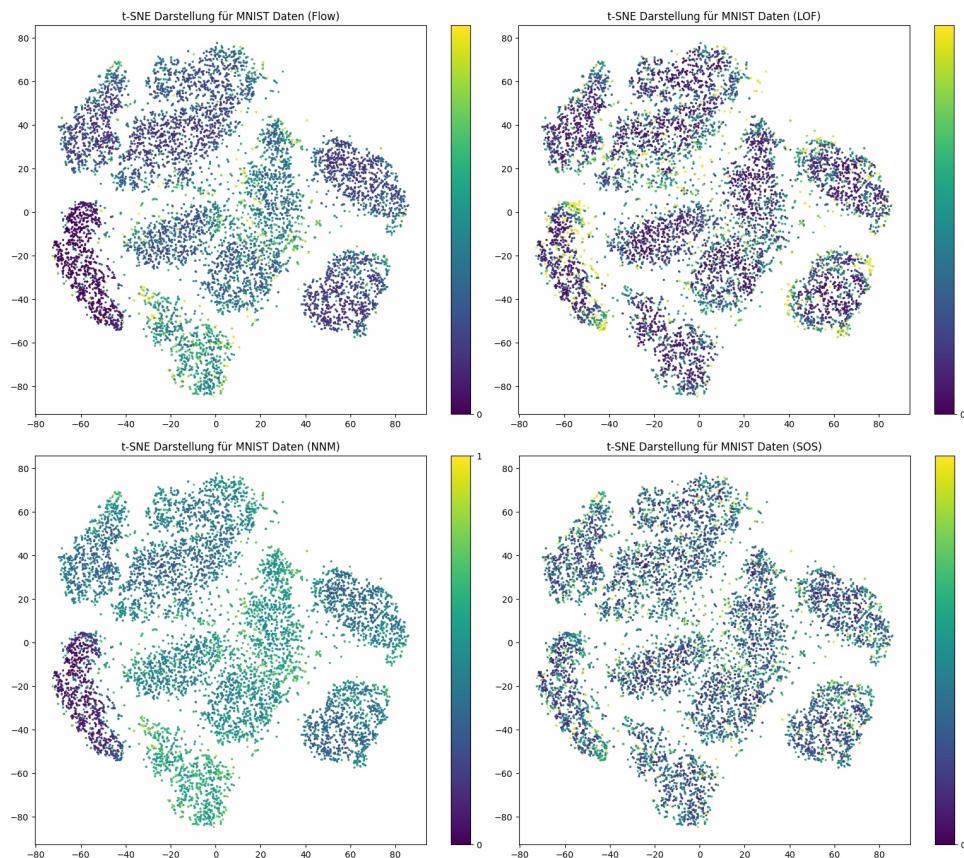


Abbildung 5.2: t-SNE Darstellungen der Seltenheitswerte von MNIST Daten mit ausgewählten Methoden (heller dargestellte Punkte entsprechen selteneren Werten)

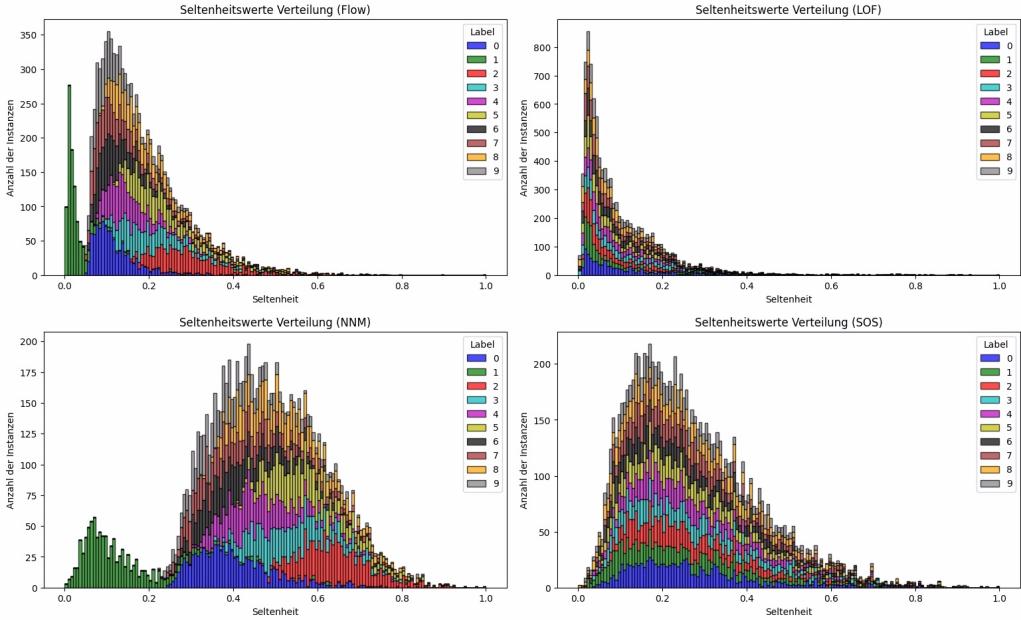


Abbildung 5.3: Histogramme der Seltenheitswerte von MNIST Daten mit ausgewählten Methoden

Die Analyse der t-SNE Darstellungen (siehe Abb. 5.2) und Histogramme (siehe Abb. 5.3) der Seltenheitswerte im Teildatensatz der MNIST Daten offenbart, dass die untersuchten Methoden in zwei Hauptkategorien unterteilt werden können. Einerseits gibt es Methoden, die auf lokaler Dichte basieren und Seltenheiten konsistent über alle Ziffern hinweg identifizieren, wie LOF und SOS. Andererseits zeigen Methoden wie NNM und Flow eine breitere Streuung der Seltenheitswerte, was auch auf die gewählten Parameter zurückzuführen ist. Insbesondere zeigt die Betrachtung der Histogramme der zweiten Methodenkategorie, dass die Ziffer „1“ häufig in den Bereich niedriger Seltenheitswerte fällt. Dies lässt sich dadurch erklären, dass Einsen meist wenig Variation aufweisen und oft als einfache Linie gezeichnet werden, was jedoch nicht ausschließt, dass vereinzelt ungewöhnliche Darstellungen existieren. Vielmehr bedeutet es, dass die Mehrheit der Einsen starke Ähnlichkeiten untereinander und mit anderen Ziffern aufweist.

Interessant ist auch, dass die Verteilung der Seltenheitswerte je nach Methode variiert. Insbesondere zeigt LOF kaum Werte in höheren Bereichen,

was für eine Methode zur Ausreißererkennung von Vorteil ist, da Daten außerhalb des dichten Bereichs wahrscheinlich als Ausreißer identifiziert werden können. Obwohl SOS sich in derselben Kategorie wie LOF befindet, zeigt das Histogramm eine bessere Verteilung über die gesamte Skala. Es ist wichtig anzumerken, dass die Verteilung der Seltenheitswerte keine direkte Einschätzung der Effektivität der Methoden darstellt, jedoch wertvolle Einblicke bietet.

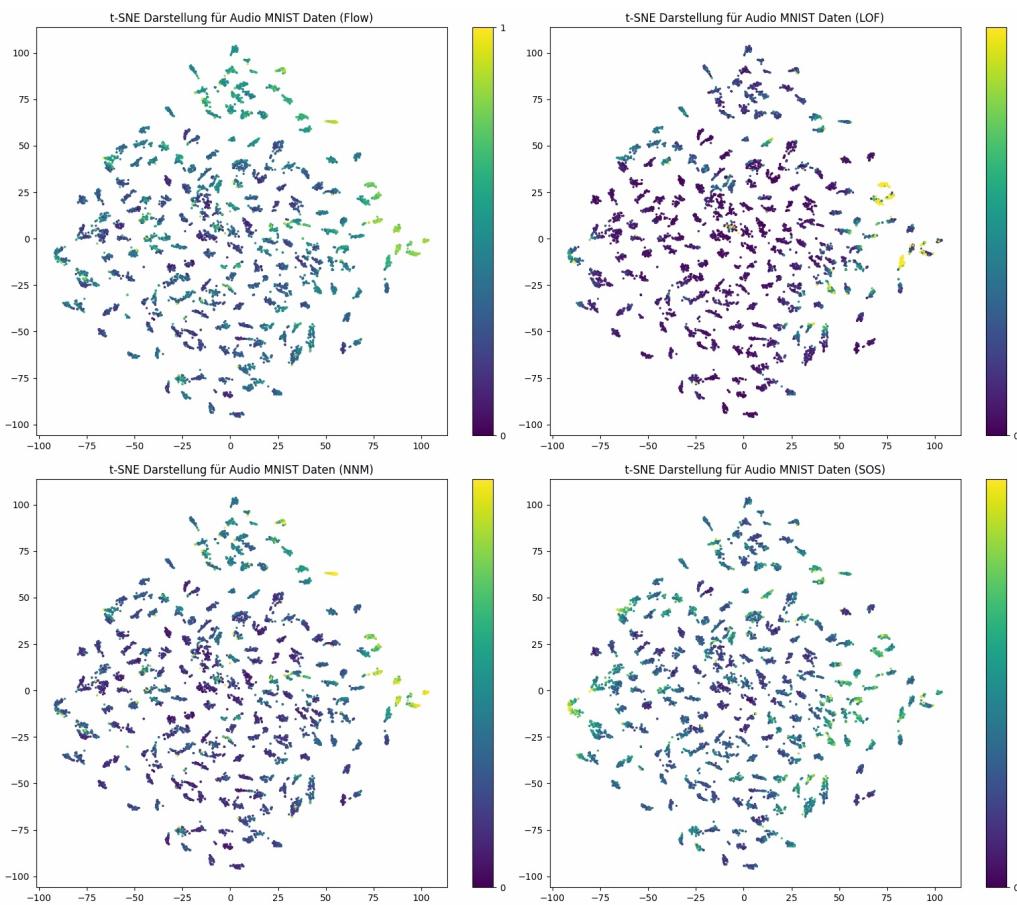


Abbildung 5.4: t-SNE Darstellungen der Seltenheitswerte von Audio MNIST Daten mit ausgewählten Methoden (heller dargestellte Punkte entsprechen selteneren Werten)

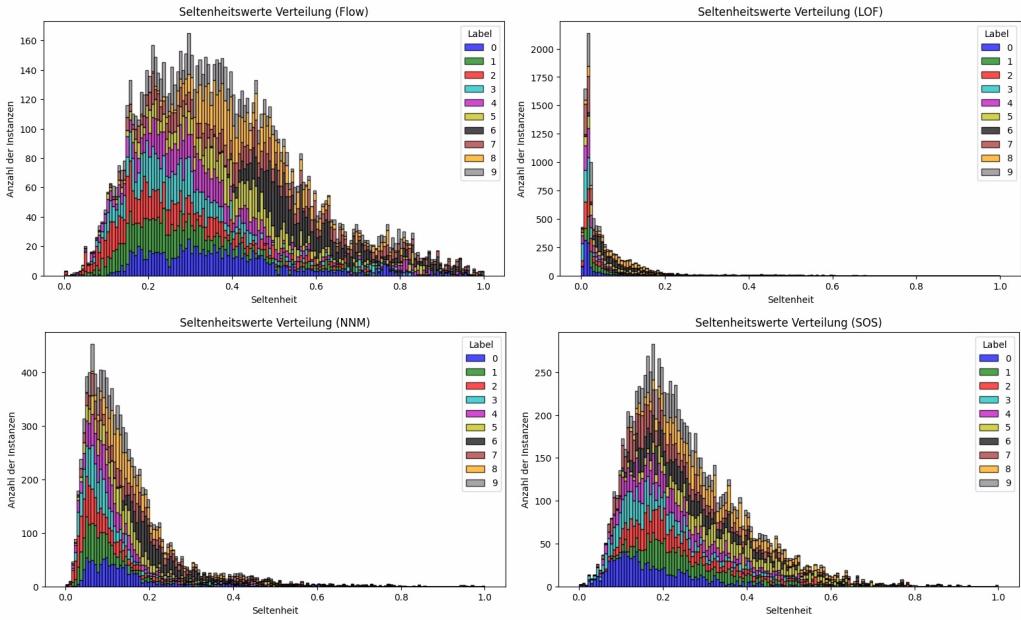


Abbildung 5.5: Histogramme der Seltenheitswerte von Audio MNIST Daten mit ausgewählten Methoden (gefärbt nach gesprochener Ziffer)

Die Audio-Daten lassen sich nicht so klar clustern wie die MNIST-Daten, was zu weniger offensichtlichen t-SNE Darstellungen führt (siehe Abb. 5.4). Es ist erkennbar, dass die Datenpunkte, die von allen Methoden als seltener eingestuft werden, tendenziell im rechten Bereich der t-SNE Darstellung gruppiert sind. Im Gegensatz zu den MNIST-Daten deutet die Visualisierung darauf hin, dass LOF im lokalen Bereich, also innerhalb der Cluster, nur wenige Datenpunkte als selten identifiziert. Im Unterschied dazu scheinen die anderen untersuchten Methoden eine größere Anzahl an Datenpunkten als selten zu klassifizieren. Die Betrachtung der Histogramme der Seltenheitsverteilung (siehe Abb. 5.5) zeigt, dass die Verteilung der Datenpunkte meistens gleichmäßig ist, mit der Ziffer „6“, die als auffallend seltener im Vergleich zu den übrigen Ziffern hervorsticht.

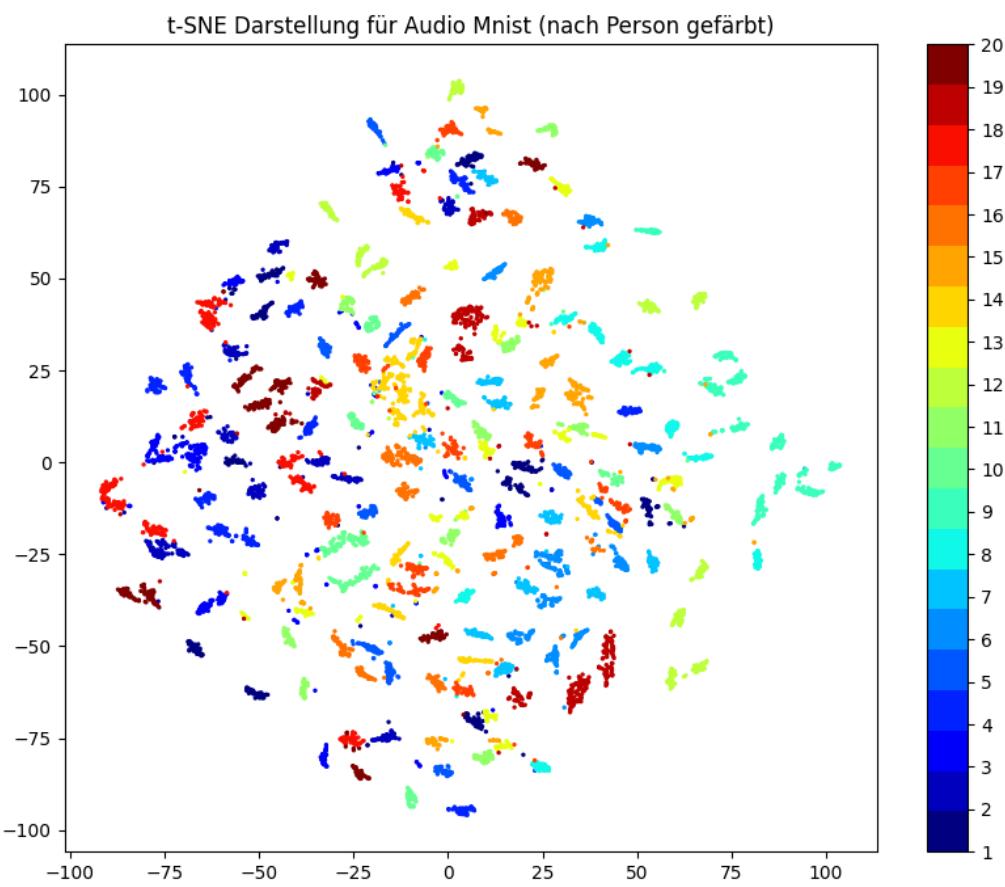


Abbildung 5.6: t-SNE Darstellung der Audio MNIST Teildaten gefärbt nach Person

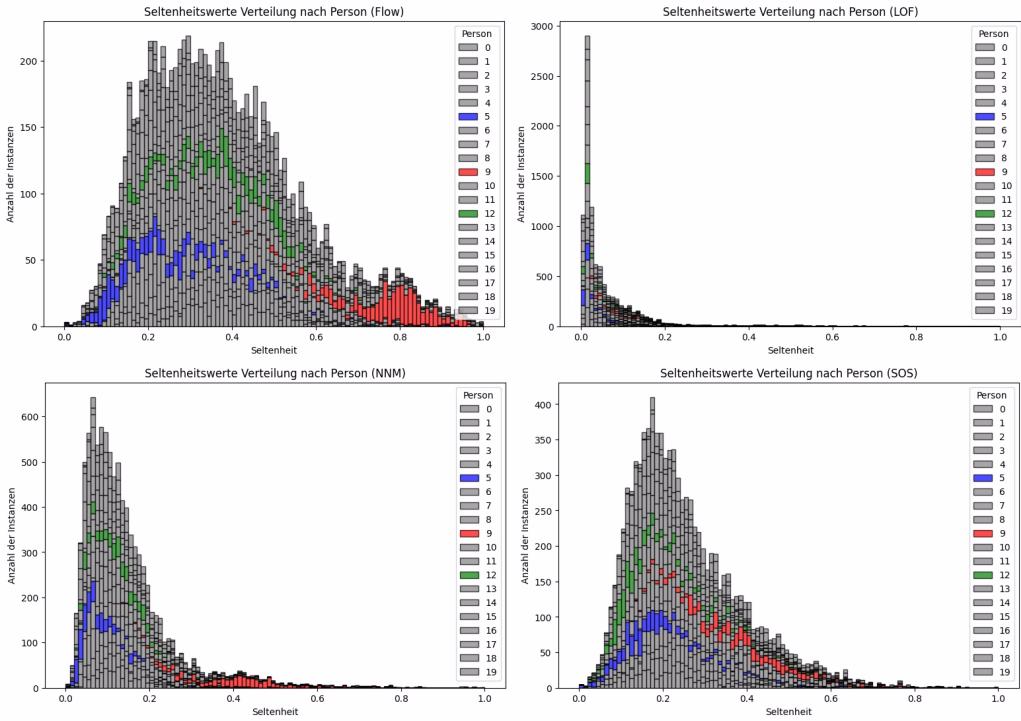


Abbildung 5.7: Histogramme der Seltenheitswerte von Audio MNIST Daten mit ausgewählten Methoden (gefärbt nach Sprecher\*innen)

Betrachtet man Abbildung 5.6, wird deutlich, dass die hohen Seltenheitswerte spezifischen Personen zugeordnet werden können, was darauf hindeutet, dass die Sprecher\*innenidentität einen signifikanten Einfluss auf die Seltenheitsbewertung hat. Insbesondere scheint die Identität eine größere Rolle zu spielen als die ausgesprochenen Ziffern selbst. Dies wird auch in Abbildung 5.7 ersichtlich, wo vor allem Daten von Person 9, die auch im rechten Bereich der t-SNE-Darstellung positioniert ist, von den meisten Methoden als seltener klassifiziert werden. Diese Beobachtungen legen nahe, dass bei der Analyse von Audio-Daten die individuellen Charakteristika der Sprecherinnen und Sprecher, wie ihre Stimmfarbe, Aussprache oder Akzentuierung, wesentliche Faktoren für die Seltenheitseinstufung sind.

Dies bietet eine umfassende Übersicht über die Methoden zur Quantifizierung von Seltenheit in den beiden Datensätzen. Weiterführend wird diese detaillierte Einsicht genutzt, um die zweite Hypothese von Kapitel 5.1 in

mehrere Experimente zu gliedern, die sowohl mit Daten durchgeführt werden, die als selten als auch solche, die als weniger selten eingestuft werden, um ein umfassendes Verständnis der Seltenheitsdynamiken in verschiedenen Datensatzkontexten zu erlangen.

Basierend auf den initialen Experimenten und der tiefgreifenden Analyse der Seltenheitswerte wurden spezifische Ziffern und Sprecher\*innen für die künstliche Reduzierung ausgewählt. Die Auswahlkriterien berücksichtigten die Ergebnisse der Seltenheitsbewertung und zielten darauf ab, ein umfassendes Verständnis der Seltenheitsdynamiken zu erlangen.

Für den MNIST-Datensatz wurden die Ziffern „0“, „1“ und „2“ für die experimentelle Fokussierung ausgewählt, die jeweils unterschiedliche Seltenheitswerte repräsentieren. Die gewählten Instanzenzahlen sind 700, 500, 250, 100, 50 und 10. Diese Staffelung ermöglicht eine gründliche Untersuchung des Einflusses der Datenmenge auf die Seltenheitsbewertung. Die Ziffer „0“ wurde als Vertreter des mittleren Seltenheitsbereichs gewählt. Sie ist interessant, da sie aufgrund ihrer geschlossenen Form und weniger variierenden Darstellung eine besondere Rolle in der Bewertung der Seltenheit spielt. Die Ziffer „1“ wurde aufgrund ihrer geringen Seltenheit ausgewählt. Sie ist in der Regel einfach zu identifizieren und zeigt weniger Variationen im Vergleich zu anderen Ziffern, was sie zu einem guten Kandidaten für die Analyse der Seltenheit bei häufig vorkommenden Datenpunkten macht. Die Ziffer „2“ wurde wegen ihrer hohen Seltenheit ausgewählt. Sie kann in verschiedenen ungewöhnlichen Formen auftreten, was sie zu einem idealen Kandidaten für die Untersuchung von Seltenheitsmerkmalen macht.

Im Audio MNIST-Datensatz wurden die Ziffern „0“, „6“ und „2“ für die vertiefende Analyse herangezogen. Auch hier wurde die gleiche Menge an Instanzen reduziert. Ähnlich wie im MNIST-Datensatz repräsentiert die Ziffer „0“ den mittleren Seltenheitsbereich, bietet aber im auditiven Kontext eine einzigartige Perspektive aufgrund der unterschiedlichen Aussprachevarianten. Die Ziffer „6“ wurde aufgrund ihrer hohen Seltenheit ausgewählt, da

sie in den vorherigen Analysen als besonders selten hervorgetreten ist, was sie zu einem interessanten Fall für weiterführende Untersuchungen macht. Die Ziffer „2“ repräsentiert die Datenpunkte mit niedriger Seltenheit und dient als Referenzpunkt, um die Auswirkungen von Seltenheit auf die Klassifikationsleistung zu verstehen.

Zusätzlich zur Fokussierung auf spezifische Ziffern wurde im Audio MNIST-Datensatz auch eine Reduzierung von Audiodateien bestimmter Sprecherinnen und Sprecher beschlossen. Die Anzahl der Audiodateien pro gesprochener Ziffer werden auf 40, 20, 10 und 5 reduziert. Hierfür wurden Person 9, Person 5 und Person 12 ausgewählt. Person 9, die hohe Seltenheitswerte aufwies, bietet die Möglichkeit, die Auswirkungen von Seltenheit auf die Datenrepräsentation zu untersuchen. Person 5 wurde aufgrund der geringen Seltenheitswerte ausgewählt, um die Kontraste in der Seltenheitsverteilung besser verstehen zu können. Person 12, die einen mittleren Seltenheitsbereich repräsentiert und zusätzlich als einzige Frau im Teildatensatz interessant ist, bietet eine einzigartige Perspektive, die insbesondere die Diversität der Stimmcharakteristiken in den Vordergrund stellt.

Diese gezielte Auswahl ermöglicht es, die Hypothesen über die Seltenheit in Datenpunkten weiter zu prüfen und zu verfeinern. Die durchgeführten Experimente sollen aufzeigen, wie sich die künstliche Veränderung der Datensatzgröße auf die Seltenheitswerte auswirkt und inwiefern die entwickelten Methoden zur Bewertung der Seltenheit diese Veränderungen erkennen und quantifizieren können.

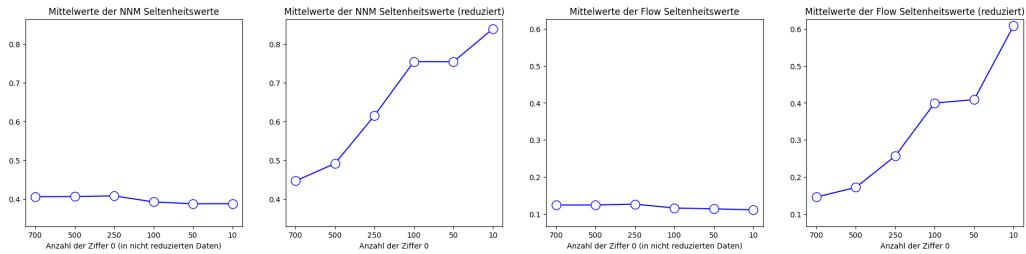


Abbildung 5.8: Reduzierung der Ziffer 0 im MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow)

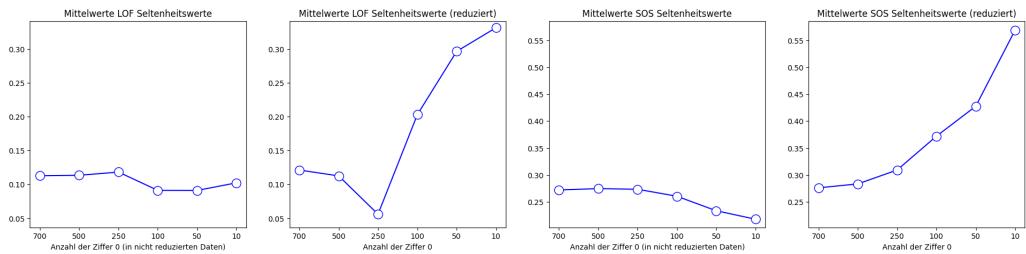


Abbildung 5.9: Reduzierung der Ziffer 0 im MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS)

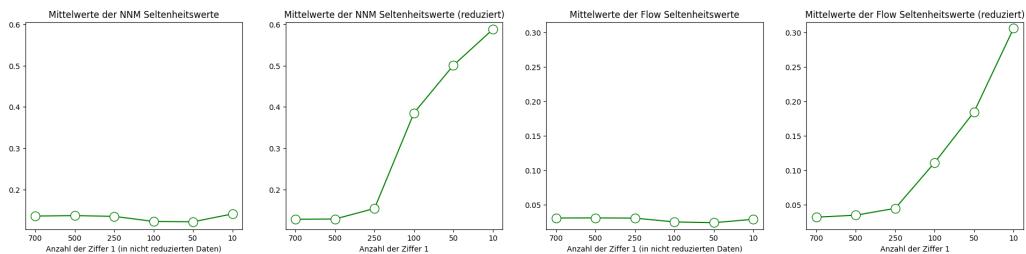


Abbildung 5.10: Reduzierung der Ziffer 1 im MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow)

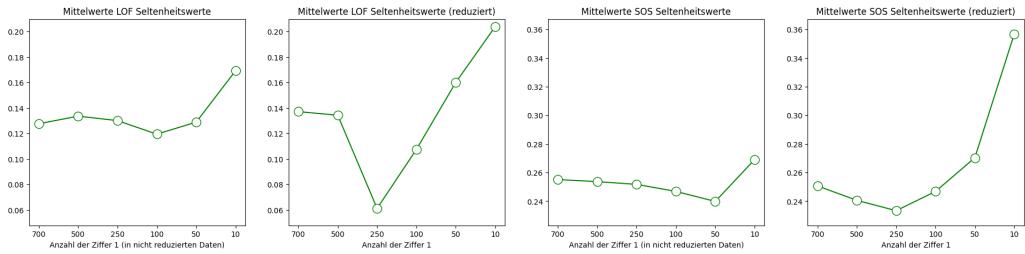


Abbildung 5.11: Reduzierung der Ziffer 1 im MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS)

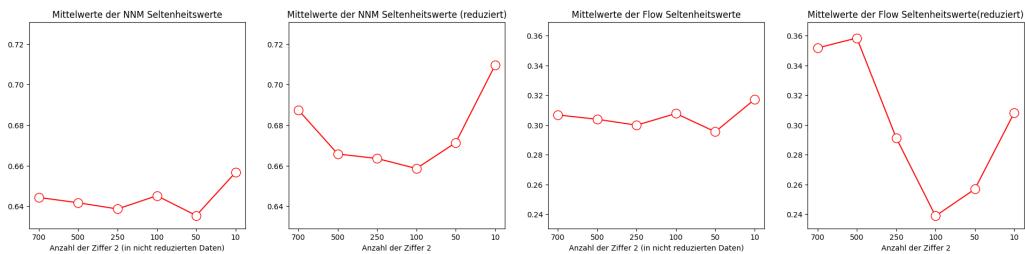


Abbildung 5.12: Reduzierung der Ziffer 2 im MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow)

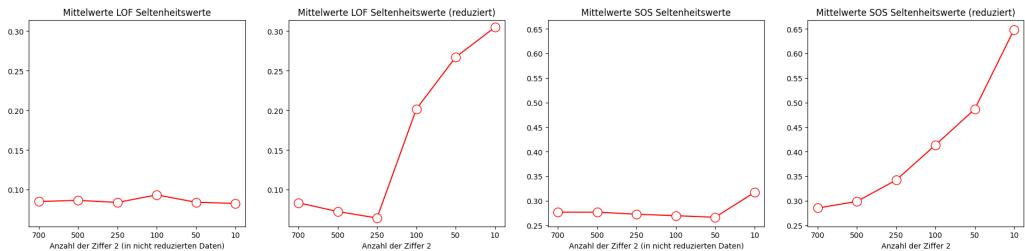


Abbildung 5.13: Reduzierung der Ziffer 2 im MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS)

Die Analyse der Effekte der Datenreduzierung auf die Seltenheitswerte wird durch die Abbildungen 5.8, 5.9, 5.10, 5.11, 5.12, und 5.13 detailliert illustriert. Diese Abbildungen visualisieren, wie sich die Seltenheitswerte verändern, wenn die Anzahl der Instanzen einer bestimmten Ziffer im MNIST-Datensatz systematisch verringert wird. Die beobachteten Trends bieten Einblicke in die Wirksamkeit der verwendeten Methoden unter variierenden Bedingungen der Datensatzgröße.

In den Abbildungen 5.8 und 5.9 ist erkennbar, dass die Seltenheitswerte für die Ziffer 0 steigen, sobald die Anzahl der Instanzen reduziert wird. Diese Ergebnisse stützen die Hypothese, dass eine Verringerung der Daten zu einem Anstieg der Seltenheitswerte führt. Interessanterweise zeigt sich bei der Methode LOF zunächst ein Abfall der Seltenheitswerte, was auf die spezifische Sensitivität der Methode gegenüber der Einstellung der Hyperparameter, insbesondere  $n_{neighbors}$ , zurückzuführen sein könnte.

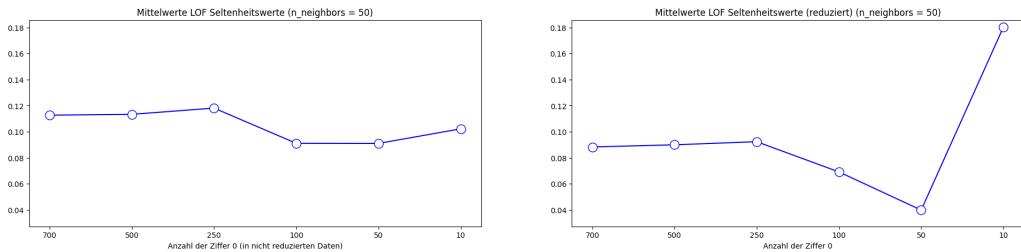


Abbildung 5.14: Reduzierung der Ziffer 0 im MNIST Teildatensatzes auf bestimmte Mengen (LOF mit  $n_{neighbors} = 50$ )

Eine Änderung des  $n_{neighbors}$  Parameters (siehe Abb. 5.14) in der Reduzierungsexperimenten offenbart, dass die zuvor beobachtete Abnahme der Seltenheitswerte maßgeblich auf die Wahl dieses spezifischen Parameters zurückzuführen ist. Durch die Modifikation des  $n_{neighbors}$ -Werts zeigt sich eine deutliche Veränderung im Verhalten der Seltenheitsbewertung, was die kritische Rolle der Parameterkonfiguration in der Effektivität der angewandten Methode unterstreicht.

Bei der Betrachtung der Abbildungen 5.10 und 5.11 für die Ziffer 1 wird ebenfalls ein genereller Trend des Anstiegs der Seltenheitswerte nach der Reduzierung deutlich. Obwohl die Methoden Flow und NNM konsistente Ergebnisse zeigen, ist bei den Anomalieerkennungsmethoden LOF und SOS ein initialer Abfall der Seltenheitswerte zu beobachten.

Die Darstellungen für die Ziffer 2, illustriert in den Abbildungen 5.12 und 5.13, zeigen ein anderes Muster. Hier scheinen die Methoden NNM und Flow

Schwierigkeiten zu haben, die bereits hohe Ausgangsseltenheit der Ziffer adäquat zu erfassen. Insbesondere bei Flow sinken die Seltenheitswerte nach der Reduzierung, was auf die Herausforderung hinweist, die bereits vorhandene Seltenheit weiter zu erhöhen. Im Gegensatz dazu zeigen die Anomalieerkennungsmethoden LOF und SOS bessere Anpassungen, wobei SOS eine kontinuierliche Zunahme der Seltenheitswerte aufzeigt. Dieses Phänomen könnte darauf zurückzuführen sein, dass die initiale Seltenheit der Ziffer 2 schon hoch ist und somit die Herausforderung besteht, diesen hohen Ausgangspunkt noch weiter zu steigern.

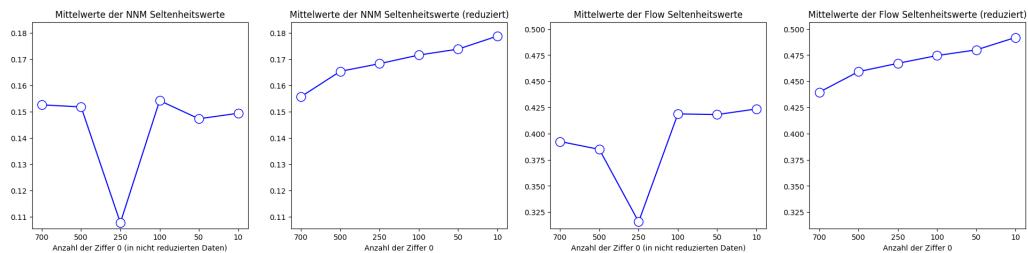


Abbildung 5.15: Reduzierung der Ziffer 0 im Audio MNIST Teildatensatz auf bestimmte Mengen (NNM und Flow)

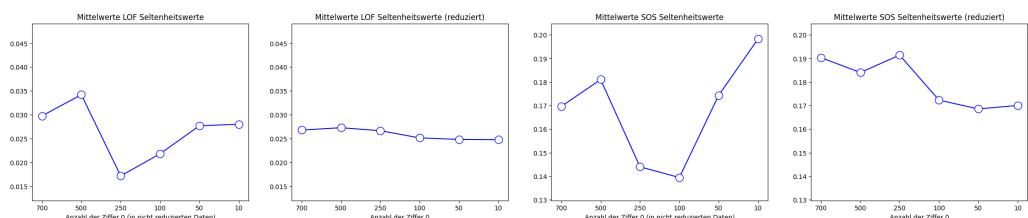


Abbildung 5.16: Reduzierung der Ziffer 0 im Audio MNIST Teildatensatz auf bestimmte Mengen (LOF und SOS)

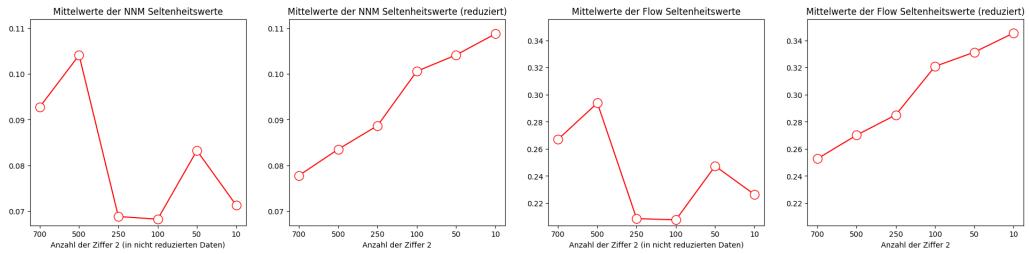


Abbildung 5.17: Reduzierung der Ziffer 2 im Audio MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow)

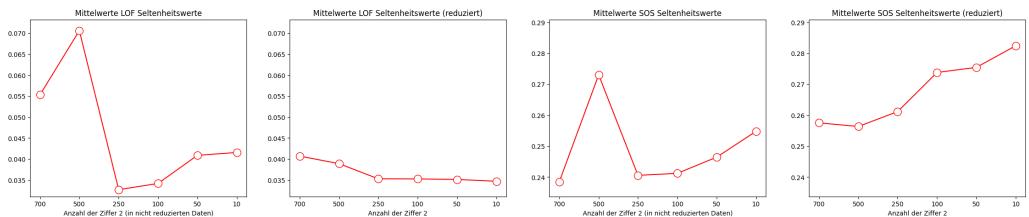


Abbildung 5.18: Reduzierung der Ziffer 2 im Audio MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS)

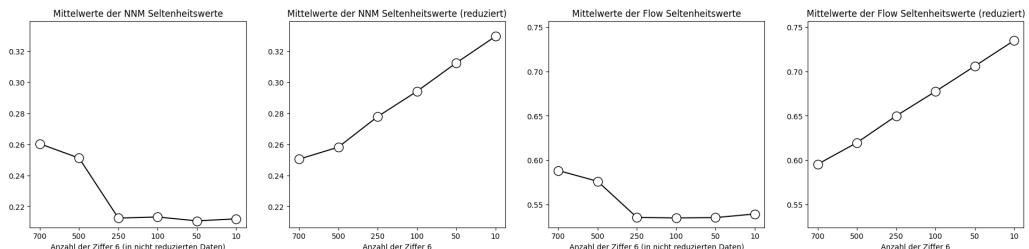


Abbildung 5.19: Reduzierung der Ziffer 6 im Audio MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow)

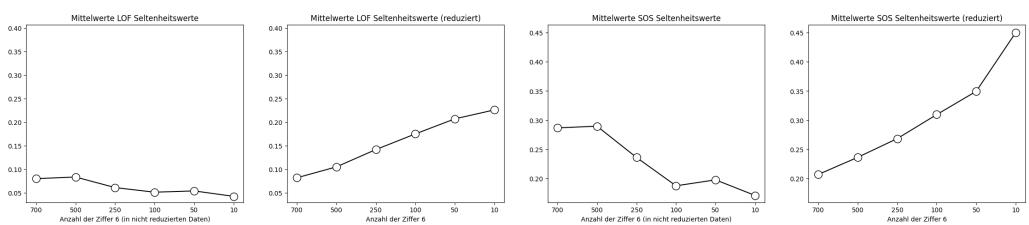


Abbildung 5.20: Reduzierung der Ziffer 6 im Audio MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS)

Die Untersuchung der Reduzierungseffekte innerhalb des Audio MNIST Daten, illustriert durch die Abbildungen 5.15 bis 5.20, liefert aufschlussreiche Erkenntnisse über die Anpassung der Seltenheitswerte in Reaktion auf die veränderte Datensatzgröße. Generell bestätigen die Ergebnisse die Hypothese, dass eine Reduzierung der Dateninstanzen zu einer Erhöhung der Seltenheitswerte führt, was die Effektivität der angewandten Methoden unterstreicht.

Die Abbildungen, die sich auf die Reduzierung der Ziffern im Audio MNIST-Datensatz beziehen, zeigen überwiegend, dass die angewandten Methoden in der Lage sind, die Seltenheitswerte nach der Datenreduktion zu erhöhen. Dies bestätigt die Annahme, dass die Verringerung der Verfügbarkeit bestimmter Ziffern ihre Seltenheit innerhalb des Gesamtdatensatzes erhöht, was eine wichtige Bestätigung der zugrundeliegenden Hypothesen darstellt.

Jedoch gibt es spezifische Fälle, in denen die Anomalieerkennungsmethode LOF nicht wie erwartet funktioniert. Insbesondere bei den Abbildungen 5.16 und 5.18 ist zu beobachten, dass die Seltenheitswerte nach der Reduktion der Daten tatsächlich sinken, statt wie vorhergesagt zu steigen. Dies ist auch hier mit der Auswahl der  $n_{neighbors}$  Parameters zu begründen. Auch die Methode SOS zeigt in Abbildung 5.16, die sich auf die Ziffer 0 bezieht, eine Reduktion der Seltenheitswerte, was ebenfalls unerwartet ist.

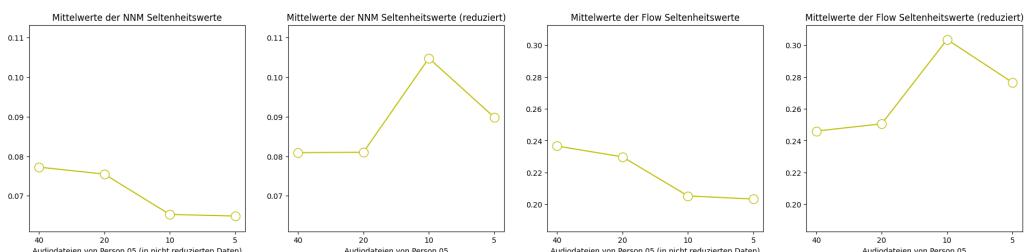


Abbildung 5.21: Reduzierung der Audiodateien von Person 5 im Audio MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow)

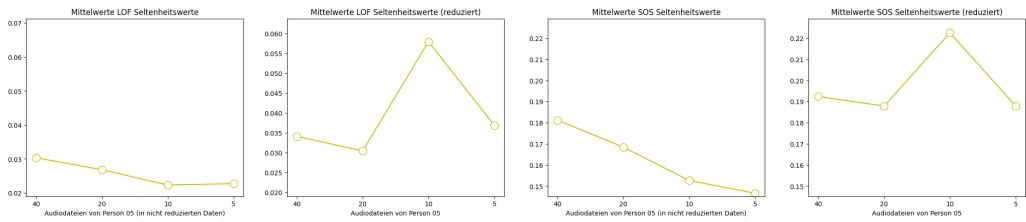


Abbildung 5.22: Reduzierung der Audiodateien von Person 5 im Audio MNIST Teildatensatz auf bestimmte Mengen (LOF und SOS)

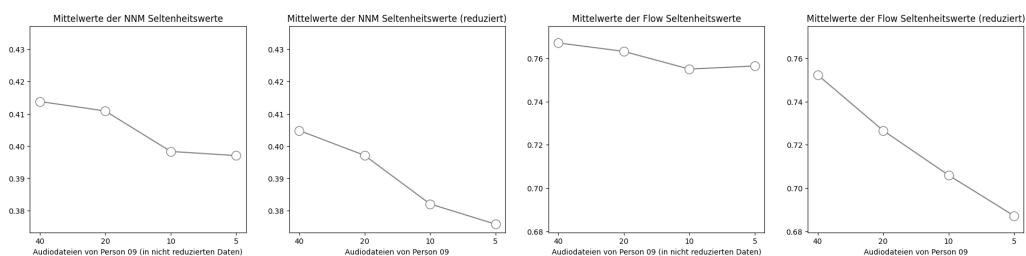


Abbildung 5.23: Reduzierung der Audiodateien von Person 9 im Audio MNIST Teildatensatz auf bestimmte Mengen (NNM und Flow)

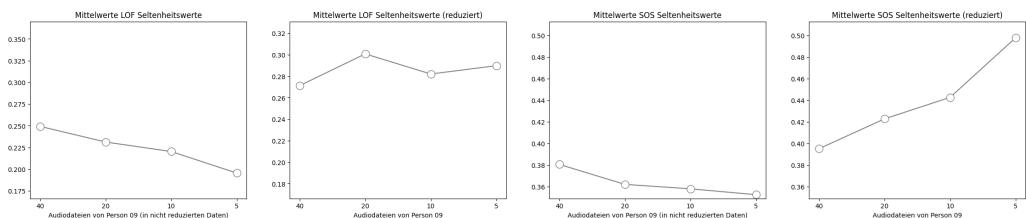


Abbildung 5.24: Reduzierung der Audiodateien von Person 9 im Audio MNIST Teildatensatz auf bestimmte Mengen (LOF und SOS)

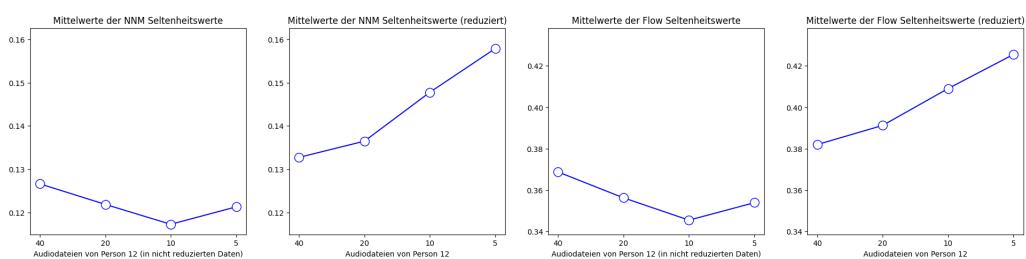


Abbildung 5.25: Reduzierung der Audiodateien von Person 12 im Audio MNIST Teildatensatz auf bestimmte Mengen (NNM und Flow)

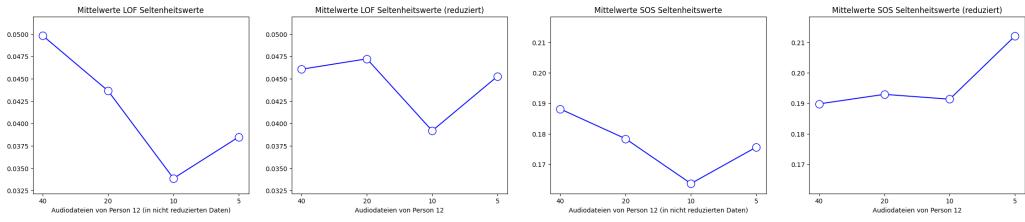


Abbildung 5.26: Reduzierung der Audiodateien von Person 12 im Audio MNIST Teildatensatz auf bestimmte Mengen (LOF und SOS)

Die Analyse der Reduzierungseffekte bezüglich der Sprecher\*innen im Audio MNIST-Datensatz, dargestellt in den Abbildungen 5.21 bis 5.26, bietet weitere Einblicke in die Anpassungsfähigkeit der Seltenheitsbewertungsmethoden. Die Ergebnisse für die Reduzierung von Person 5 und 12 bestätigen durchgehend die Hypothese, da eine Verringerung der Dateninstanzen zu einer Erhöhung der Seltenheitswerte führt. Diese Beobachtungen untermauern die Annahme, dass die Seltenheit eines Datenpunktes innerhalb des Gesamtdatensatzes zunimmt, wenn die Verfügbarkeit dieses spezifischen Datenpunktes reduziert wird.

Besonders bemerkenswert ist die Analyse der Reduzierung von Sprecher 9, dessen Daten bereits in der initialen Analyse (siehe Abb. 5.23) als relativ selten identifiziert wurden. In diesem Fall zeigen die Methoden NNM und Flow unerwartete Ergebnisse: anstatt die Seltenheitswerte zu erhöhen, verringern sich diese nach jeder Reduzierung der Dateninstanzen. Diese Beobachtung könnte darauf hindeuten, dass beide Methoden Schwierigkeiten haben, bereits als selten klassifizierte Daten noch seltener zu machen. Die initial hohe Seltenheit dieser Sprachdaten könnte eine Grenze darstellen, an der die Fähigkeit dieser Methoden, Seltenheit zu erkennen und zu quantifizieren, herausfordert wird.

Im Kontrast dazu scheinen die Methoden LOF und insbesondere SOS effektiv zu arbeiten, indem sie eine konsistente Steigerung der Seltenheitswerte bei jeder Reduktion der Daten von Sprecher 9 liefern (siehe Abb. 5.24). Dies zeigt, dass diese Methoden besser an die spezifischen Herausforderungen an-

gepasst sind, die sich aus der bereits hohen Ausgangsseltenheit ergeben. LOF und SOS zeigen sich resilient gegenüber den Schwierigkeiten, die sich aus der weiteren Verstärkung der Seltenheit bei bereits seltenen Daten ergeben, und bestätigen ihre Eignung für die Bewertung von Seltenheitswerten in solchen spezifischen Kontexten.

### 5.3.2 Madrid Tripadvisor Rezensionen

**Top 5:**  
Mittelwert: 0.5171237680930735 Index: 3991  
Headline: Questo posto necessita innanzitutto di una raccomandazione: dimenticate le parole antipasto, primo e secondo e fidatevi del mio consiglio. Perché non è un ristorante, non di per sé. Altrimenti rischiate di non finire il cibo. In questo abbiamo preso: un insalata con il bescola e il pulpito gallego di antipasto. Il polpo era qualcosa di divino. Uno di noi ha preso un menu del dia, con trippa e pollo, mentre gli altri hanno preso una porzione di riso funghi e tartufo, buonissima, e soprattutto due porzioni di carne arrosto. Questo merita una descrizione. La carne viene portata in quantità industriale cruda e ogni tavolo viene fornito di una sorta di formaggio. Una figata pazzesca. Poi abbiamo preso patate fritte, un vassoi, dolce, una bottiglia di vino, gassosa, birra e acqua. Un caffè, 138 euro in totale. Se avessimo saputo quanto erano enormi le dosi, avremmo preso meno roba, comunque da provare.

Mittelwert: 0.4144770441660847 Index: 616  
Headline: Linguini de gambón, pizza cuatro quesos and especially milhojas de dulce was incredible, thank u estafany

Mittelwert: 0.407730572860746 Index: 552  
Headline: No les puedo recomendar este restaurante más! Comenzando con la ensalada Doré (verdes, salmón, aguacate y vinagre balsámico), una paella de mariscos (para 2 personas) la tarta de limón, café y un shot de baileys. No se arrepentirán de venir. Nuestro mesero Miguel Alejandro también fue excelente. I can't recommend this restaurant enough! I had the Doré salad (greens, salmon and avocados with balsamic vinegar), seafood paella (for two), lemon tart, coffee and baileys. You wont regret coming here. The staff was also very welcoming. Our server Miguel Alejandro was very attentive and excellent at his job. Le service était excellent. Je vous recommande ce restaurant. J'ai commandé la salade Doré (du saumon, d'avocats et du vinaigre balsamique), paella de fruit de mer (pour deux), de la tarte de citron, du café et du baileys. Les serveurs étaient très attentifs, spécifiquement Miguel Alejandro, le notre.

Mittelwert: 0.3257744708341661 Index: 4492  
Headline: Splendid food and hospitality! Good spirit with friends after long day's at the EAU conference!!!!!!!!!!!!!!

Mittelwert: 0.3113533868577557 Index: 4893  
Headline: We went for lunch and were very pleased. We would definitely return if in Madrid again. Food: - Octopus was very tasty and tender. "Pulpo a la brasa" (\*\*\*\*\*) - Cochinitillo confitado a alta temperatura y lavado en su jugo" was excellent (\*\*\*\*\*) - Ribeye (grilled, from Escorial) was spectacular. I had it medium ("al punto mas"), rather than medium-rare because it's tougher than corn-fed U.S. beef. "Taco de lomo de buey a la brasa" (\*\*\*\*\*) - Hamburger Wagyu was average and did not earn the special designation "Wagyu" (\*\*\*) - Tapas: Veggies, croquettes and salmorejo (tomato soup) were excellent (\*\*\*\*\*) Wine: - 2015 Verdejo from Rueda was excellent: dry and mineral with a hint of fruit great with veggies and octopus (\*\*\*\*\*) - 2012 crianza Ribera del Duero Valdubón was full bodied and strong (perhaps too strong for my taste) enjoyable with beef (\*\*) -Service: - Professional, efficient and very friendly. - Ambiance: - Classic Spanish elegance. Even the king has a private dining room at Cafe de Oriente. -Location: Central next to the royal palace. Great idea to eat here after touring the palace and/or the adjacent church.

Abbildung 5.27: Die fünf Rezensionen aus dem Madrid-Datensatz mit den höchsten Seltenheitswerten, ermittelt durch die NNM-Methode

Last 2:  
Mittelwert: 0.018788712645340565 Index: 3740  
Headline: Make dinner reservations at **Casa Benigna**. Do it now. You will not be disappointed. I do not normally open a review so strongly, but this was the best meal of over ten days in Spain. It was not only a meal, but an experience perfect for a couple looking for a quiet unique night together. We will stay with them again first. I thought that maybe our Uber driver had made a mistake. The restaurant facade is so unassuming, and the door was locked. Using the knocker, the friendly staff opened the door to a unique and charming establishment. Each table has some unique decorations that make it special. Our table was set and waiting for us, not only with place settings but with the start of our first course, beans and olives. Olive oil, balsamic, and French butter soon followed. The next course was a small bowl of vegetable and cod soup. It was lovely, and included in the 6€ per person cover charge, as was the bread. Next we had our starter. We chose the herring. The marinated herring had a more subtle flavor than we get in the USA, and the variety of toppings made each piece a unique treat. Both paellas we ordered were excellent. The seafood and the Iberian pork paellas had the right mix of smoky and savory, and the rice was cooked to perfection. Each bite was an absolute pleasure. Our meal concluded with light sparkling wine and smooth hot chocolate, even without having ordered dessert. As if this amazing meal was not enough, I was invited back into the kitchen by Norberto, the owner. There I met the chefs who cooked our culinary experience. They were enthusiastic and clearly used to meeting the customers. Norberto and I chatted as if we were old friends. We talked about business, my home state of Massachusetts, and life in Spain. We had a wonderful night at **Casa Benigna**, and I will 100% return when I am next in Madrid. I may stop in the city just for another great meal with Norberto and his crew.

Mittelwert: 0.019339876894993623 Index: 2678  
Headline: Another great find on **Tripadvisor**. Being Spanish, having lived in Madrid and liking Basque-style food, I thought I had been to every decent Basque-style Asador worth the name in this city, and **Pelotari** was not on my list. As things turned out, I was clearly missing one important name. I was entertaining a close friend who also loves Basque food and who insisted on going elsewhere (and paying the night's bill for having it his way). As I was visiting, he was kind enough to accept going to **Pelotari**, a place neither of us knew. The only condition was that dinner was on me. I'm picky with food quality, and Basque food leaves little room for concessions in that field, as food preparations are basic. You cannot disguise the quality of meat being served if it is just charcoal grilled and with some sea salt over it. No fancy sauces to hide bare bones reality. We settled for some **appetizers**. **Jamón** was exquisite, and salted anchovies were good without being great. A savory and very convincing cod omelette completed the first round. Then came big **chuletillas** (steaks) to share. It was from here that this old Meat was ably charcoal grilled, tender, slightly spicy. I'm running out of adjectives to describe it. It was **BEST** I've had in a long time. The meal was washed down with a nice **Ribera del Duero** wine. I found prices on the wine list to be adequate, whereas the wine selection was downright dull. Few nice bottles, and the one we picked was not available, so we had to settle for a second option. I guess the crisis has impacted the quality and amount of bottles restaurateurs keep in inventory these days. We had some desserts, not worth mentioning, and a not so great (our waiter pretended otherwise) gin & tonic to complete a great meal. To their credit, the cocktails were on the house, and not because we complained or anything. Service was competent and polite. It was a slow August Saturday when we visited, so I cannot judge how things will fare in busier days. Check came down to around 200 EUR, and wine was a thing of that. Pricey, but money well spent. Dining room is on the classic side of things, and I did not like the lighting (no windows), but those were minor peccadilloes when thinking of how things fared altogether. We will definitely come back.

Abbildung 5.28: Die zwei Rezensionen aus dem Madrid-Datensatz mit den niedrigsten Seltenheitswerten, ermittelt durch die NNM-Methode

Bei der Anwendung von NNM auf den Madrid-Datensatz zeigt sich, dass Rezensionen in anderen Sprachen als Englisch häufig als selten eingestuft werden (siehe Abb. 5.27). Dies ist teilweise darauf zurückzuführen, dass für die Erstellung der Ähnlichkeitsmatrix ein Modell verwendet wurde, das auf Englisch trainiert ist. Zudem werden Texte, die eine hohe Anzahl von Sonderzeichen enthalten, ebenfalls als selten identifiziert. Im Gegensatz dazu verdeutlichen die beiden Rezensionen mit den niedrigsten Seltenheitswerten (siehe Abb. 5.28), dass ausführliche Texte mit einfachen, klaren Wortstrukturen als weniger selten betrachtet werden. Die Wörter die in diesen Rezensionen vorkommen, treten wahrscheinlich öfter im gesamten Datensatz auf, was dazu führt, dass die entsprechenden Rezensionen als üblich oder typisch für den Datensatz angesehen werden.

Obwohl in den präsentierten Beispielen ausschließlich die Ergebnisse der NNM-Methode gezeigt werden, dient dies lediglich als Beispiel für die Anwendung von Textdaten. Es ist zu beachten, dass sowohl die Flow-Methode als auch die Aureißer-Methoden sehr ähnliche oder sogar identische Ergebnisse in der Seltenheitsbewertung erzielen können.

### 5.3.3 Molekulare Daten

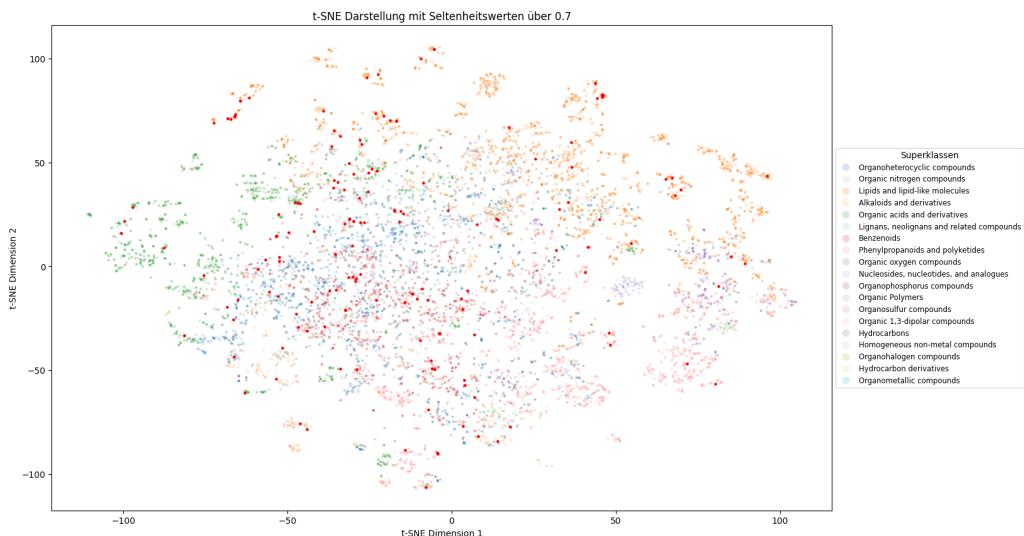


Abbildung 5.29: t-SNE-Visualisierung molekularer Daten: Punkte mit einem Seltenheitswert über 0,7 durch NNM bestimmt, sind rot hervorgehoben.

Bei der Anwendung von NNM auf molekulare Daten identifiziert die Methode bestimmte Punkte als besonders selten, was in der Visualisierung deutlich wird (siehe Abb. 5.29). Um die Ergebnisse übersichtlicher zu gestalten, werden die Daten in Superklassen kategorisiert. Diese Klassifizierung hilft dabei, zu veranschaulichen, in welchen spezifischen Gruppen sich die seltensten Moleküle konzentrieren. Auffällig ist, dass insbesondere in den Superklassen „Lipide und lipidähnliche Moleküle“ sowie „Alkaloide und Derivate“ eine hohe Dichte an seltenen Datenpunkten zu verzeichnen ist. Diese Einteilung dient hier vorrangig als Demonstrationsbeispiel für die Leistungsfähigkeit der Methode. Eine weitergehende Unterteilung in zusätzliche Gruppen könnte die Ergebnisse noch detaillierter darstellen, allerdings zielt die aktuelle Darstellung darauf ab, einen ersten Einblick in das Potenzial der Methode zu geben.

# 6 Diskussion und Ausblick

In diesem Kapitel werden die Ergebnisse der Studie im Hinblick auf die Validität der entwickelten Methoden zur Bewertung der Seltenheit in hochdimensionalen Datensätzen diskutiert. Durch die sorgfältige Analyse der durchgeführten Experimente mit dem MNIST- und Audio MNIST-Datensatz sowie den zusätzlichen Untersuchungen von Text- und Molekulardatensätzen konnten wertvolle Erkenntnisse über die Leistungsfähigkeit und Grenzen der neuen Verfahren gewonnen werden. Darüber hinaus wird ein Ausblick für zukünftige Arbeiten und Entwicklungen skizziert.

## 6.1 Diskussion der Ergebnisse

Die Analyse und Diskussion der Ergebnisse aus den durchgeführten Experimenten bieten aufschlussreiche Einblicke in die Leistungsfähigkeit der entwickelten Methoden zur Bewertung der Seltenheit von Datenpunkten. Die selbstentwickelten Methoden, NNM und Flow, haben sich überwiegend als effektiv erwiesen, indem sie in den meisten Fällen angemessene Seltenheitswerte zuweisen konnten. Diese Methoden zeigten eine robuste Performance über ein breites Spektrum von Datenkonfigurationen hinweg, mit Ausnahme von Situationen, in denen Datenpunkte reduziert wurden, die bereits als selten eingestuft waren. In diesen speziellen Fällen zeigten NNM und Flow Herausforderungen, die Seltenheitswerte nach der Reduzierung angemessen zu erhöhen, was auf eine mögliche Grenze ihrer Anpassungsfähigkeit in Szenarien mit bereits hohen Seltenheitswerten hindeutet.

Eine besondere Stärke zeigte die Methode SOS, die konsistent in der Lage war, die Seltenheit der bereits seltenen Datenpunkte nach weiteren Reduzierungen zu erhöhen. Dies unterstreicht die Zuverlässigkeit und Robustheit von SOS, auch in komplexen Szenarien mit sich dynamisch verändernden Datensätzen. Allerdings lieferte sie in vereinzelten Fällen, wie beispielsweise in Abbildung 5.16, unerwartete oder atypische Ergebnisse. Auf der anderen Seite erwies sich LOF als eine Methode, die nützliche Ergebnisse lieferte, allerdings mit einer ausgeprägten Sensitivität bezüglich der Einstellung ihrer Parameter, insbesondere  $n\_neighbors$ . Diese Sensitivität erfordert eine sorgfältige Kalibrierung, um die bestmöglichen Ergebnisse zu erzielen und die Methode effektiv einzusetzen.

Die in Kapitel 5.1 formulierten Hypothesen wurden durch die experimentellen Ergebnisse aus Kapitel 5.2 größtenteils bestätigt. Es zeigte sich, dass keine der untersuchten Methoden durchweg unzureichend oder ungeeignet war, allerdings offenbarten sich spezifische Stärken und Schwächen, die in bestimmten Kontexten relevant werden. Die selbstentwickelten Methoden NNM und Flow bieten einen praktikablen und leicht anwendbaren Ansatz, um die Seltenheit in hochdimensionalen Datensätzen zu bewerten, was ihre Integration in weiterführende analytische Prozesse erleichtert.

Zusammenfassend lässt sich sagen, dass die durchgeführten Untersuchungen erfolgreich waren und Methoden hervorgebracht haben, die effektiv die Seltenheit in Datensätzen quantifizieren. Diese Methoden erfüllen die in dieser Arbeit gestellten Herausforderungen und bieten solide Ansätze für die Bewertung der Seltenheit, was sie zu wertvollen Werkzeugen für die Datenanalyse in verschiedenen Anwendungsbereichen macht. Im Vergleich zu anderen Methoden zeichnen sich die entwickelten Verfahren besonders durch ihre Anwendungsflexibilität aus. Sie lassen sich auf verschiedene Arten von Daten wie Bild, Audio, Text und Molekulardaten anwenden. Diese Universalität ist besonders wertvoll in multidisziplinären Forschungsfeldern, wo Daten aus unterschiedlichen Quellen und in verschiedenen Formaten vorliegen.

## 6.2 Ausblick

Die durchgeführten Untersuchungen und die daraus resultierenden Erkenntnisse legen ein solides Fundament für zukünftige Forschungsrichtungen und die Weiterentwicklung der Methoden zur Bewertung der Seltenheit von Datenpunkten in hochdimensionalen Datensätzen. Die ermittelten Seltenheitswerte bieten wertvolle Informationen, die für fortgeschrittene Anwendungen genutzt werden können, wie beispielsweise in der Erforschung von Naturstoffen. Hier könnten die identifizierten seltenen Daten dazu beitragen, Modelle gezielt mit diesen wertvollen und seltenen Informationen zu trainieren, was besonders in Bereichen von Bedeutung ist, wo seltene Ereignisse entscheidende Einblicke oder Durchbrüche ermöglichen können.

Eine zentrale Richtung für zukünftige Arbeiten ist die erweiterte Validierung und Optimierung der entwickelten Methoden. Durch den Einsatz dieser Methoden in einem breiteren Spektrum von Anwendungsfällen und Datensätzen könnten ihre Zuverlässigkeit und Genauigkeit weiter gestärkt werden. Dabei könnte insbesondere die Anpassung und Feinabstimmung der Algorithmen im Fokus stehen, um die Methoden noch präziser und effektiver zu gestalten.

Die Integration von neuen Technologien, insbesondere aus den Bereichen des maschinellen Lernens und der künstlichen Intelligenz, verspricht bedeutende Fortschritte. Die Anwendung von Deep Learning und anderen fortschrittlichen Algorithmen könnte neue Möglichkeiten eröffnen, insbesondere in der Analyse komplexer und unstrukturierter Daten, und so die Fähigkeit zur Erkennung und Bewertung von Seltenheit erheblich verbessern.

Die Übertragung der entwickelten Methoden auf verschiedene Fachgebiete stellt eine spannende Perspektive dar. In Disziplinen wie der Biomedizin, der Pharmakologie oder der Umweltwissenschaft könnten diese Methoden entscheidend dazu beitragen, seltene Ereignisse oder Muster zu identifizieren, die für spezifische wissenschaftliche oder praktische Fragestellungen von Be-

deutung sind. Die Zusammenarbeit mit Fachexpert\*innen aus unterschiedlichen Bereichen könnte dazu führen, dass die Methoden maßgeschneidert für spezielle Anforderungen und Fragestellungen angepasst werden.

Zukünftige Forschungen könnten sich darauf konzentrieren, die Effektivität der vorgestellten Methoden in der Naturstoffforschung weiter zu validieren und zu optimieren. Hier könnten die identifizierten seltenen Daten dazu beitragen, Modelle gezielt mit diesen wertvollen und seltenen Informationen zu trainieren, was besonders in Bereichen von Bedeutung ist, wo seltene Ereignisse entscheidende Einblicke oder Durchbrüche ermöglichen können. Dabei könnte der Schwerpunkt auf der Anpassung der Algorithmen liegen, um die spezifischen Herausforderungen dieser Domäne noch besser adressieren zu können. Durch die Integration fortschrittlicher Technologien aus den Bereichen des maschinellen Lernens und der künstlichen Intelligenz könnten die Methoden verfeinert werden, um die Identifizierung seltener Molekülstrukturen in Naturstoffdatenbanken noch präziser und effizienter zu gestalten.

Darüber hinaus ist die Entwicklung von benutzungsfreundlichen Tools und Softwarepaketen, die die implementierten Methoden zugänglich machen, ein wichtiger Schritt, um die praktische Anwendbarkeit der Forschungsergebnisse zu erweitern. Dies würde es einem breiteren Spektrum von Nutzenden ermöglichen, von den entwickelten Methoden zu profitieren, ohne sich intensiv mit den zugrundeliegenden Algorithmen auseinandersetzen zu müssen.

Insgesamt eröffnen die Ergebnisse dieser Arbeit vielfältige Wege für zukünftige Forschungen und Entwicklungen. Sie bilden eine solide Grundlage, von der aus die Methoden zur Seltenheitsbewertung weiterentwickelt und für ein breites Spektrum von Anwendungen nutzbar gemacht werden können, was das Potenzial hat, sowohl das wissenschaftliche Verständnis als auch praktische Anwendungen erheblich zu bereichern.

# Abkürzungsverzeichnis

**DIF** Deep Isolation Forest

**ECOD** Empirical-Cumulative-distribution-based Outlier Detection

**InChI** International Chemical Identifier

**KNN** K-Nearest Neighbors

**LOF** Local Outlier Factor

**MNIST** Modified National Institute of Standards and Technology

**NNM** Nächste-Nachbarn-Methode

**PYOD** Python Outlier Detection

**SOS** Stochastic Outlier Selection

**SMILES** Simplified Molecular Input Line Entry System

**t-SNE** t-Distributed Stochastic Neighbor Embedding

# Liste der Algorithmen

1	Nächste-Nachbarn-Methode . . . . .	30
2	Flow Methode . . . . .	33

# Abbildungsverzeichnis

2.1	Seltenheitswerte von MNIST Daten (siehe Kapitel 3.2) anhand von ECOD . . . . .	8
5.1	Links: t-SNE Darstellung der MNIST Teildaten gefärbt nach Ziffer. Rechts: t-SNE Darstellung der Audio MNIST Teildaten gefärbt nach Ziffer . . . . .	39
5.2	t-SNE Darstellungen der Seltenheitswerte von MNIST Daten mit ausgewählten Methoden (heller dargestellte Punkte entsprechen selteneren Werten) . . . . .	41
5.3	Histogramme der Seltenheitswerte von MNIST Daten mit ausgewählten Methoden . . . . .	42
5.4	t-SNE Darstellungen der Seltenheitswerte von Audio MNIST Daten mit ausgewählten Methoden (heller dargestellte Punkte entsprechen selteneren Werten) . . . . .	43
5.5	Histogramme der Seltenheitswerte von Audio MNIST Daten mit ausgewählten Methoden (gefärbt nach gesprochener Ziffer) .	44
5.6	t-SNE Darstellung der Audio MNIST Teildaten gefärbt nach Person . . . . .	45
5.7	Histogramme der Seltenheitswerte von Audio MNIST Daten mit ausgewählten Methoden (gefärbt nach Sprecher*innen) .	46
5.8	Reduzierung der Ziffer 0 im MNIST Teildatensatz auf bestimmte Mengen (NNM und Flow) . . . . .	49
5.9	Reduzierung der Ziffer 0 im MNIST Teildatensatz auf bestimmte Mengen (LOF und SOS) . . . . .	49

5.10	Reduzierung der Ziffer 1 im MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow) . . . . .	49
5.11	Reduzierung der Ziffer 1 im MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS) . . . . .	50
5.12	Reduzierung der Ziffer 2 im MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow) . . . . .	50
5.13	Reduzierung der Ziffer 2 im MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS) . . . . .	50
5.14	Reduzierung der Ziffer 0 im MNIST Teildatensatzes auf bestimmte Mengen (LOF mit $n_{neighbors} = 50$ ) . . . . .	51
5.15	Reduzierung der Ziffer 0 im Audio MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow) . . . . .	52
5.16	Reduzierung der Ziffer 0 im Audio MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS) . . . . .	52
5.17	Reduzierung der Ziffer 2 im Audio MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow) . . . . .	53
5.18	Reduzierung der Ziffer 2 im Audio MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS) . . . . .	53
5.19	Reduzierung der Ziffer 6 im Audio MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow) . . . . .	53
5.20	Reduzierung der Ziffer 6 im Audio MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS) . . . . .	53
5.21	Reduzierung der Audiodateien von Person 5 im Audio MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow) . . . . .	54
5.22	Reduzierung der Audiodateien von Person 5 im Audio MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS) . . . . .	55
5.23	Reduzierung der Audiodateien von Person 9 im Audio MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow) . . . . .	55
5.24	Reduzierung der Audiodateien von Person 9 im Audio MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS) . . . . .	55
5.25	Reduzierung der Audiodateien von Person 12 im Audio MNIST Teildatensatzes auf bestimmte Mengen (NNM und Flow) . . . . .	55

5.26 Reduzierung der Audiodateien von Person 12 im Audio MNIST Teildatensatzes auf bestimmte Mengen (LOF und SOS) . . . . .	56
5.27 Die fünf Rezensionen aus dem Madrid-Datensatz mit den höchsten Seltenheitswerten, ermittelt durch die NNM-Methode . . . . .	57
5.28 Die zwei Rezensionen aus dem Madrid-Datensatz mit den niedrigsten Seltenheitswerten, ermittelt durch die NNM-Methode . . . . .	58
5.29 t-SNE-Visualisierung molekularer Daten: Punkte mit einem Seltenheitswert über 0,7 durch NNM bestimmt, sind rot hervorgehoben. . . . .	59

# Literaturverzeichnis

[Abadi u. a. 2016] ABADI, Martín ; BARHAM, Paul ; CHEN, Jianmin ; CHEN, Zhifeng ; DAVIS, Andy ; DEAN, Jeffrey ; DEVIN, Matthieu ; GHEMAWAT, Sanjay ; IRVING, Geoffrey ; ISARD, Michael u. a.: Tensorflow: A system for large-scale machine learning. In: *12th { USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, S. 265–283

[Angiulli 2018] ANGIULLI, Fabrizio: On the behavior of intrinsically high-dimensional spaces: Distances, direct and reverse nearest neighbors, and hubness. In: *Journal of Machine Learning Research* 18 (2018), 04, S. 1–60

[Banerjee u. a. 2023] BANERJEE, Jineta ; TARONI, Jaclyn N. ; ALLAWAY, Robert J. ; PRASAD, Deepashree V. ; GUINNEY, Justin ; GREENE, Casey: Machine learning in rare disease. In: *Nature Methods* 20 (2023), Nr. 6, S. 803–814

[Becker u. a. 2018] BECKER, Sören ; ACKERMANN, Marcel ; LAPUSCHKIN, Sebastian ; MÜLLER, Klaus-Robert ; SAMEK, Wojciech: Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. (2018), 07

[Becker u. a. 2023] BECKER, Sören ; VIELHABEN, Johanna ; ACKERMANN, Marcel ; MÜLLER, Klaus-Robert ; LAPUSCHKIN, Sebastian ; SAMEK, Wojciech: AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark. In: *Journal of the Franklin Institute* (2023). <http://dx.doi.org/https://doi.org/10.1016/j.jfranklin.2023.11.038>. – DOI <https://doi.org/10.1016/j.jfranklin.2023.11.038>. – ISSN 0016–0032

- [Bellman u. Kalaba 1959] BELLMAN, Richard ; KALABA, Robert: On adaptive control processes. In: *IRE Transactions on Automatic Control* 4 (1959), Nr. 2, S. 1–9
- [Breunig u. a. 2000] BREUNIG, Markus M. ; KRIEGEL, Hans-Peter ; NG, Raymond T. ; SANDER, Jörg: LOF: identifying density-based local outliers. (2000), 93–104. <http://dx.doi.org/10.1145/342009.335388>. – DOI 10.1145/342009.335388. ISBN 1581132174
- [Burkel u. a. 2012] BURKEL, E. ; IGE, Oladeji O. ; UMORU, Lasisi E. ; ARIBO, Sunday: Natural Products: A Minefield of Biomaterials. In: *ISRN Materials Science* 2012 (2012), 983062. <http://dx.doi.org/10.5402/2012/983062>. – DOI 10.5402/2012/983062. ISBN 2356–7872
- [Cha 2007] CHA, Sung-Hyuk: Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions. In: *Int. J. Math. Model. Meth. Appl. Sci.* 1 (2007), 01
- [Chandola u. a. 2009] CHANDOLA, Varun ; BANERJEE, Arindam ; KUMAR, Vipin: Anomaly detection: A survey. In: *ACM Comput. Surv.* 41 (2009), jul, Nr. 3. <http://dx.doi.org/10.1145/1541880.1541882>. – DOI 10.1145/1541880.1541882. – ISSN 0360–0300
- [Chauhan u. a. 2018] CHAUHAN, Rahul ; GHANSHALA, Kamal K. ; JOSHI, R.C: Convolutional Neural Network (CNN) for Image Detection and Recognition. In: *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 2018, S. 278–282
- [Chen u. a. 2009] CHEN, Shihyen ; MA, Bin ; ZHANG, Kaizhong: On the similarity metric and the distance metric. 410 (2009), may, Nr. 24–25, 2365–2376. <http://dx.doi.org/10.1016/j.tcs.2009.02.023>. – DOI 10.1016/j.tcs.2009.02.023. – ISSN 0304–3975
- [Chollet u. a. 2015] CHOLLET, François u. a.: *Keras*. <https://keras.io>, 2015. – Letzter Zugriff: 23.01.2024

- [Cukier u. Mayer-Schoenberger 2013] CUKIER, Kenneth ; MAYER-SCHOENBERGER, Viktor: The rise of big data: How it's changing the way we think about the world. In: *Foreign Aff.* 92 (2013), S. 28
- [Dang u. a. 2015] DANG, Taurus T. ; NGAN, Henry Y. ; LIU, Wei: Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. (2015), S. 507–510. <http://dx.doi.org/10.1109/ICDSP.2015.7251924>. – DOI 10.1109/ICDSP.2015.7251924
- [Djoumbou Feunang u. a. 2016] DJOUMBOU FEUNANG, Yannick u. a.: Classy-Fire: automated chemical classification with a comprehensive, computable taxonomy. In: *Journal of Cheminformatics* 8 (2016), Nr. 1, 61. <http://dx.doi.org/10.1186/s13321-016-0174-y>. – DOI 10.1186/s13321-016-0174-y. ISBN 1758-2946
- [Donoho u. a. 2000] DONOHO, David L. u. a.: High-dimensional data analysis: The curses and blessings of dimensionality. In: *AMS math challenges lecture* 1 (2000), Nr. 2000, S. 7–8, 18
- [Draghici 2012] DRAGHICI, S.: *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. 2. Chapman and Hall/CRC, 2012. – 565–579 S. <http://dx.doi.org/10.1201/b11566>. <http://dx.doi.org/10.1201/b11566>
- [Ellison u. Agrawal 2005] ELLISON, Aaron ; AGRAWAL, Anurag: The Statistics of Rarity 1. In: *Ecology* 86 (2005), 05, S. 1079–1080. <http://dx.doi.org/10.1890/04-1456>. – DOI 10.1890/04-1456
- [Feng u. a. 2023] FENG, Cindy ; LI, Longhai ; XU, Chang: Advancements in predicting and modeling rare event outcomes for enhanced decision-making. In: *BMC Medical Research Methodology* 23 (2023), Nr. 1, S. 243
- [Gaston 1994] In: GASTON, Kevin J.: *What is rarity?* Dordrecht : Springer Netherlands, 1994, S. 1–21
- [Han u. a. 2022a] HAN, Jiyeon ; CHOI, Hwanil ; CHOI, Yunjey ; KIM, Junho ; HA, Jung-Woo ; CHOI, Jaesik: Rarity score: A new metric to evaluate the

uncommonness of synthesized images. In: *arXiv preprint arXiv:2206.08549* (2022)

[Han u. a. 2022b] HAN, Songqiao ; HU, Xiyang ; HUANG, Hailiang ; JIANG, Mingqi ; ZHAO, Yue: ADBench: Anomaly Detection Benchmark. (2022)

[Harris u. a. 2020] HARRIS, Charles R. u. a.: Array programming with NumPy. In: *Nature* 585 (2020), September, Nr. 7825, 357–362. <http://dx.doi.org/10.1038/s41586-020-2649-2>. – DOI 10.1038/s41586-020-2649-2

[Honnibal u. a. 2020] HONNIBAL, Matthew ; MONTANI, Ines ; VAN LANDEGHEM, Sofie ; BOYD, Adriane: spaCy: Industrial-strength Natural Language Processing in Python. (2020). <http://dx.doi.org/10.5281/zenodo.1212303>. – DOI 10.5281/zenodo.1212303

[Hunter 2007] HUNTER, John D.: Matplotlib: A 2D graphics environment. In: *Computing in science & engineering* 9 (2007), Nr. 3, S. 90–95

[Janssens 2013] JANSSENS, J.H.M.: Outlier selection and one-class classification. (2013). ISBN 9789082027310. – Series: TiCC Ph.D. Series Volume: 27

[King u. Zeng 2002] KING, Gary ; ZENG, Langche: Logistic Regression in Rare Events Data. In: *Political Analysis* 9 (2002), 09. <http://dx.doi.org/10.1093/oxfordjournals.pan.a004868>. – DOI 10.1093/oxfordjournals.pan.a004868

[Landrum u. a. 2024] LANDRUM, Greg u. a.: *rdkit/rdkit: Release\_2023.09.5*. <http://dx.doi.org/10.5281/zenodo.591637>. Version: Februar 2024

[Lecun u. a. 1998] LECUN, Y. ; BOTTOU, L. ; BENGIO, Y. ; HAFFNER, P.: Gradient-based learning applied to document recognition. In: *Proceedings of the IEEE* 86 (1998), Nr. 11, S. 2278–2324. <http://dx.doi.org/10.1109/5.726791>. – DOI 10.1109/5.726791

- [LeCun u. a. 2015] LECUN, Yann ; BENGIO, Y. ; HINTON, Geoffrey: Deep Learning. In: *Nature* 521 (2015), 05, S. 436–44. <http://dx.doi.org/10.1038/nature14539>. – DOI 10.1038/nature14539
- [Li u. a. 2023] LI, Zheng ; ZHAO, Yue ; HU, Xiyang ; BOTTA, Nicola ; IO-NESCU, Cezar ; CHEN, George H.: ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. In: *IEEE Transactions on Knowledge and Data Engineering* 35 (2023), Dezember, Nr. 12, 12181–12193. <http://dx.doi.org/10.1109/tkde.2022.3159580>. – DOI 10.1109/tkde.2022.3159580. – ISSN 2326–3865
- [Liu u. a. 2008] LIU, Fei T. ; TING, Kai M. ; ZHOU, Zhi-Hua: Isolation Forest. (2008), S. 413–422. <http://dx.doi.org/10.1109/ICDM.2008.17>. – DOI 10.1109/ICDM.2008.17
- [Liu u. a. 2017] LIU, Shusen ; MALJOVEC, Dan ; WANG, Bei ; BREMER, Peer-Timo ; PASCUCCI, Valerio: Visualizing High-Dimensional Data: Advances in the Past Decade. In: *IEEE Transactions on Visualization and Computer Graphics* 23 (2017), Nr. 3, S. 1249–1268. <http://dx.doi.org/10.1109/TVCG.2016.2640960>. – DOI 10.1109/TVCG.2016.2640960
- [Lommers u. a. 2023] LOMMERS, Kristof ; STORCHEUS, Dmitry ; ELSAADANY, Abdelmoez ; KANCHERLA, Adi ; BAIOUNY, Mohamed: Pixel Rarity Score: Rarity Learned Directly From Pixel Data. In: *Available at SSRN* (2023). <http://dx.doi.org/10.2139/ssrn.4350207>. – DOI 10.2139/ssrn.4350207
- [López-Riobóo Botana u. a. 2022] LÓPEZ-RIOBÓO BOTANA, Inés L. ; ALONSO-BETANZOS, Amparo ; BOLÓN-CANEDO, Verónica ; GUIJARRO-BERDIÑAS, Béatrice: A TripAdvisor Dataset for Dyadic Context Analysis. (2022). <http://dx.doi.org/10.5281/zenodo.6583422>. – DOI 10.5281/zenodo.6583422
- [Märtens u. a. 2022] MÄRTENS, Kaspar ; BORTOLOMEAZZI, Michele ; MONTORSI, Lucia ; SPENCER, Jo ; CICCARELLI, Francesca ; YAU, Christopher:

Rarity: Discovering rare cell populations from single-cell imaging data. In: *bioRxiv* (2022)

[McFee u. a. 2023] MCCEE u. a.: *librosa/librosa: 0.10.1.* <http://dx.doi.org/10.5281/zenodo.8252662>. Version: August 2023

[Muhr u. a. 2023] MUHR, David ; AFFENZELLER, Michael ; KÜNG, Josef: A Probabilistic Transformation of Distance-Based Outliers. In: *Machine Learning and Knowledge Extraction* 5 (2023), Nr. 3, 782–802. <http://dx.doi.org/10.3390/make5030042>. – DOI 10.3390/make5030042. – ISSN 2504–4990

[O’Boyle 2012] O’BOYLE, Noel M.: Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. In: *Journal of Cheminformatics* 4 (2012), Nr. 1, 22. <http://dx.doi.org/10.1186/1758-2946-4-22>. – DOI 10.1186/1758-2946-4-22. ISBN 1758–2946

[OpenAI 2024] OPENAI: *ChatGPT: Ein innovatives Sprachmodell.* <https://www.openai.com/>, 2024. – Zugriff am: 2024-03-12

[Oracle 2002] ORACLE: *Data Warehousing Guide.* [https://docs.oracle.com/cd/B10501\\_01/server.920/a96520/dimensio.htm#11840](https://docs.oracle.com/cd/B10501_01/server.920/a96520/dimensio.htm#11840), 2002. – Zugriff am 31. März 2024

[Pedregosa u. a. 2011] PEDREGOSA, F. u. a.: Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830

[Rafii u. Pardo 2012] RAFII, Zafar ; PARDO, Bryan: Music/Voice Separation Using the Similarity Matrix. (2012), S. 583–588

[Ramaswamy u. a. 2000] RAMASWAMY, Sridhar ; RASTOGI, Rajeev ; SHIM, Kyuseok: Efficient algorithms for mining outliers from large data sets. (2000), 427–438. <http://dx.doi.org/10.1145/342009.335437>. – DOI 10.1145/342009.335437. ISBN 1581132174

- [Ranjan u. a. 2023] RANJAN, Mritunjay ; BAROT, Krishna ; KHAIRNAR, Vaisnavi ; RAWAL, Vaishnavi ; PIMPALGAONKAR, Anujaa ; SAXENA, Shilpi ; SATTAR, Arif: Python: Empowering Data Science Applications and Research. In: *Journal of Operating Systems Development Trends* 10 (2023), 08, S. 27–33. <http://dx.doi.org/10.37591/joosdt.v10i1.576>. – DOI 10.37591/joosdt.v10i1.576
- [Rusdiana u. a. 2021] RUSDIANA, Uus ; ERNAWATI, Iin ; FALIH, Noor ; ARISTA, Artika: Comparison of Distance Metrics on Fuzzy C-Means Algorithm Through Customer Segmentation. (2021), S. 307–311. <http://dx.doi.org/10.1109/ICIMCIS53775.2021.9699206>. – DOI 10.1109/ICIMCIS53775.2021.9699206
- [Sathe u. Aggarwal 2016] SATHE, Saket ; AGGARWAL, Charu C.: Subspace Outlier Detection in Linear Time with Randomized Hashing. (2016), S. 459–468. <http://dx.doi.org/10.1109/ICDM.2016.0057>. – DOI 10.1109/ICDM.2016.0057
- [Song u. a. 2022] SONG, Xiaojia ; XIE, Tao ; FISCHER, Stephen: Accelerating kNN search in high dimensional datasets on FPGA by reducing external memory access. In: *Future Generation Computer Systems* 137 (2022), 189–200. <http://dx.doi.org/https://doi.org/10.1016/j.future.2022.07.009>. – DOI https://doi.org/10.1016/j.future.2022.07.009. – ISSN 0167–739X
- [Van Der Maaten u. a. 2009] VAN DER MAATEN, Laurens ; POSTMA, Eric O. ; HERIK, H J. d. u. a.: Dimensionality reduction: A comparative review. In: *Journal of Machine Learning Research* 10 (2009), Nr. 66-71, S. 13
- [Van Rossum 2020] VAN ROSSUM, Guido: *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020. – 425 S.
- [Virtanen u. a. 2020] VIRTANEN, Pauli ; GOMMERS u. a.: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. In: *Nature Methods* 17 (2020), S. 261–272. <http://dx.doi.org/10.1038/s41592-019-0686-2>. – DOI 10.1038/s41592-019-0686-2

- [Waggoner 2021] WAGGONER, Philip D.: Modern Dimension Reduction. In: *CoRR* abs/2103.06885 (2021). <https://arxiv.org/abs/2103.06885>
- [Wang u. a. 2016] WANG, Mingxun u. a.: Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. In: *Nature Biotechnology* 34 (2016), 05, S. 828–837. <http://dx.doi.org/10.1038/nbt.3597>. – DOI 10.1038/nbt.3597
- [Wei u. a. 2023] WEI, Bin u. a.: Global analysis of the biosynthetic chemical space of marine prokaryotes. In: *Microbiome* 11 (2023), Nr. 1, 144. <http://dx.doi.org/10.1186/s40168-023-01573-3>. – DOI 10.1186/s40168-023-01573-3. ISBN 2049–2618
- [Weiss 2004] WEISS, Gary: Mining with rarity. In: *ACM SIGKDD Explorations Newsletter* 6 (2004), 06. <http://dx.doi.org/10.1145/1007730.1007734>. – DOI 10.1145/1007730.1007734
- [Xu u. a. 2023] XU, Hongzuo ; PANG, Guansong ; WANG, Yijie ; WANG, Yongjun: Deep Isolation Forest for Anomaly Detection. In: *IEEE Transactions on Knowledge and Data Engineering* 35 (2023), Dezember, Nr. 12, 12591–12604. <http://dx.doi.org/10.1109/tkde.2023.3270293>. – DOI 10.1109/tkde.2023.3270293. – ISSN 2326–3865
- [Zhao u. a. 2019] ZHAO, Yue ; NASRULLAH, Zain ; LI, Zheng: PyOD: A Python Toolbox for Scalable Outlier Detection. In: *Journal of Machine Learning Research* 20 (2019), Nr. 96, 1-7. <http://jmlr.org/papers/v20/19-011.html>