



Towards Reliable AI / ML Testing by Systematic Assessment of Test Data Quality

Janek Groß, Lisa Jöckel, Michael Kläs, Pascal Gerber

Bewertung von Testdaten Qualität

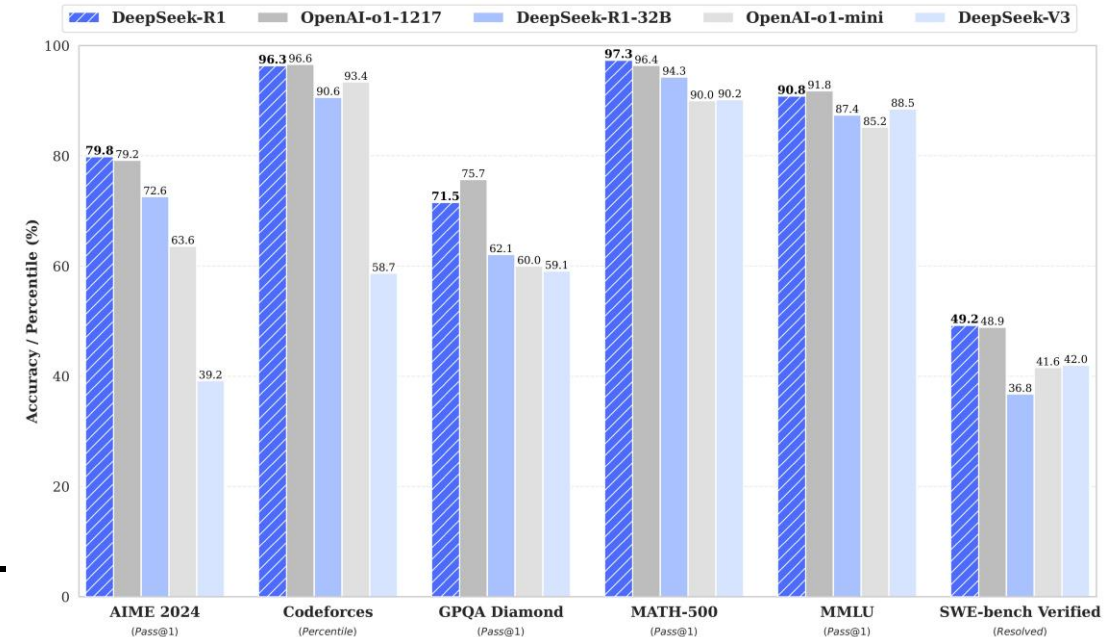
Motivation

Wie werden KI Modelle Verglichen?

- Anwendung auf Daten
- Berechnung von Metriken

Probleme:

- Nur deskriptiv, keine Information über Konfidenz.
- KI Testen gleicht eher statistischem Test.
- Voraussetzungen statistischer Tests oft nicht erfüllt.

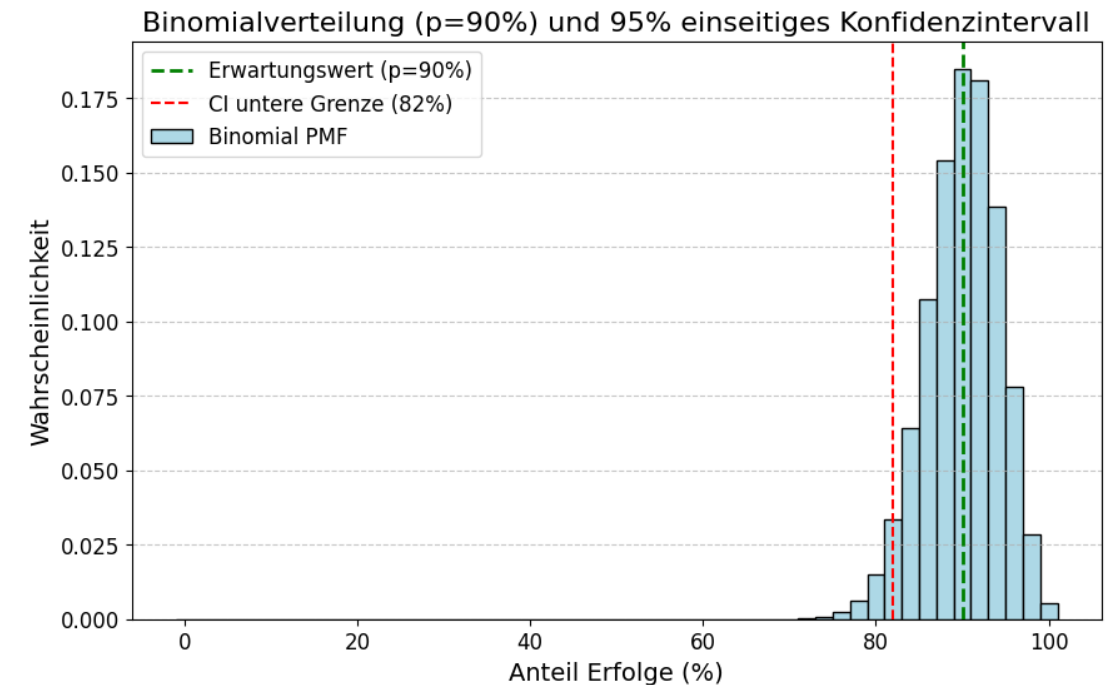


<https://github.com/deepseek-ai/DeepSeek-R1/blob/main/figures/benchmark.jpg>

Bedingung Statistischer Tests

Beispiel: Binomialverteilung

- Güte von KI Modell Ausgaben i.d.R. auf binäres richtig/falsch reduzierbar.
- Statistische Voraussetzung Konfidenzintervall der Binomialverteilung:
 - **Identische** und stochastisch **unabhängige** Verteilung der Testdaten
 - **Vorregistrierte Hypothesen**
 - **Externe Validität**



Vorgeschlagene Testdatenqualitätsaspekte

- **Representativeness:**

Die Modelleingaben folgen der gleichen Verteilung der auch die Daten die während der Anwendung beobachtet werden folgen. (**Identische Verteilung**, **Externe Validität**)

- **Label Validity:**

Die Daten sind korrekt gelabelt. (**Identische Verteilung**, **Externe Validität**)

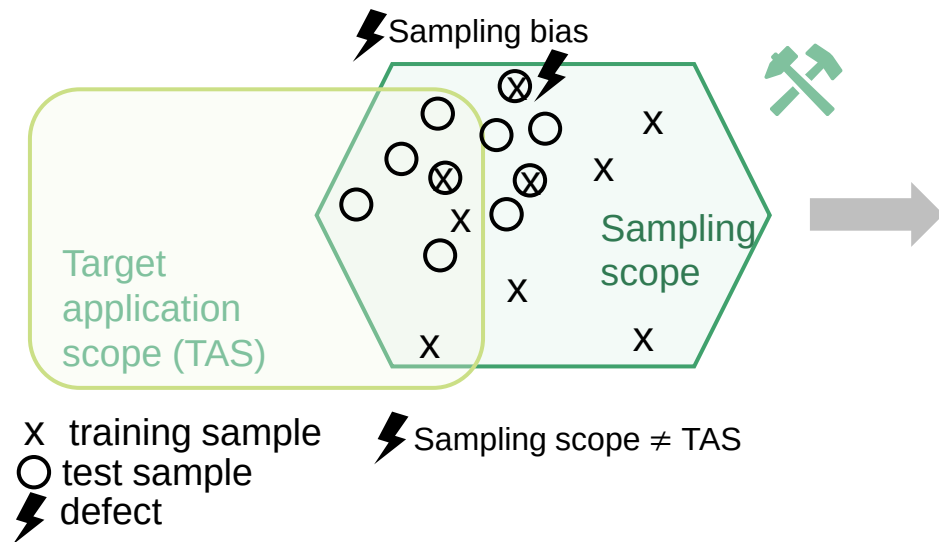
- **Self-containment:**

Die Datenpunkte erlauben keine Rückschlüsse über andere Datenpunkte aus dem Trainings- oder Testdatensatz (**Unabhängigkeit**, **Vorregistrierte Hypothesen**)

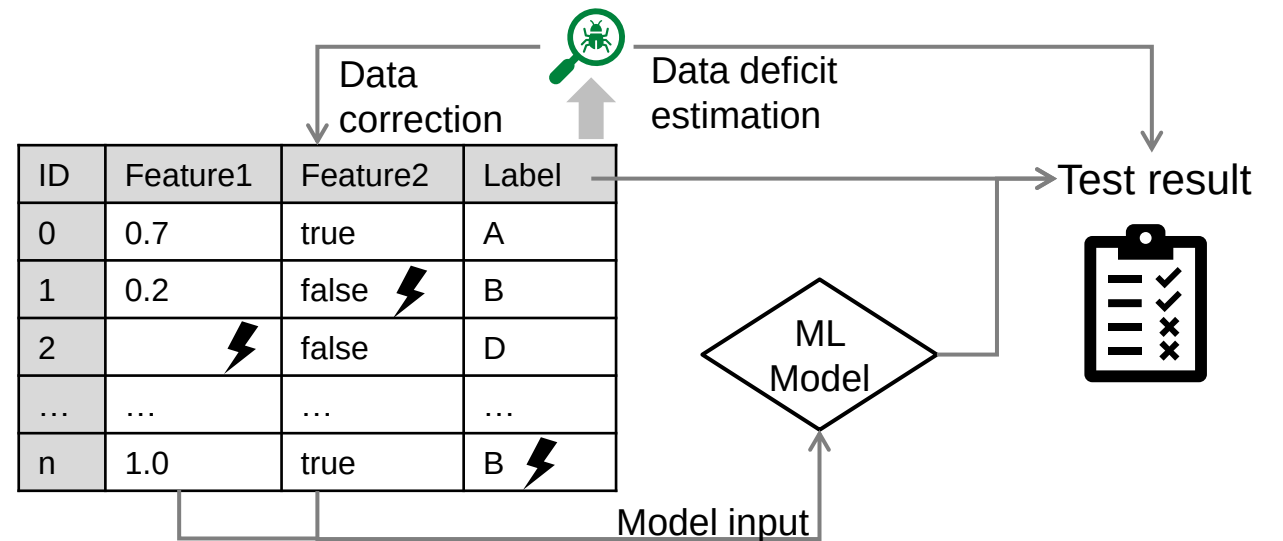
Beispielmethoden für Testdatenqualität

Überblick

Constructive data quality for data acquisition & preparation



Data quality analysis for data testing



Beispielmethoden für Testdatenqualität

Überblick

Self-containment

- Unseen test data
 -  Separation of training and test data
 -  Collection of new test data
 -  Prevention of information leaks
- Consideration of inherent dependencies
 -  Timeseries-aware data split

Representativeness

- Sampling strategies
 -  Random, stratified, or cluster sampling
-  Tests for homogeneity
-  Consideration of realistic quality deficits
-  Detection of outliers, extreme or missing data points

Label validity

- Systematic labeling
 -  Software tools
- Multiple labelers/iterations
 -  Reliability analysis
- Detection of wrong labels
 -  Confident learning¹

Beispiel Anwendung: Klassifikation von Atemwegserkrankungen

Datensatz

- Tabellarische Daten
- 38,537 Patienten
- 18 verschiedene Atemwegserkrankungen
- Public Health Department New Mexico
- 2019-2021

^ Symptoms ≡	# Age ≡	^ Sex ≡	^ Disease ≡	^ Treatment ≡	^ Nature ≡
coughing	5	female	Asthma	Omalizumab	high
tight feeling in the chest	4	female	Asthma	Mepolizumab	high
wheezing	6	male	Asthma	Mepolizumab	high
shortness of breath	7	male	Asthma	Mepolizumab	high

Beispiel Anwendung

Methode: Confident Learning¹

- Ziel:
Aufdecken von Labelfehlern, Abschätzung Anteil verbleibender Labelfehler
- Methode:
 - Entscheidungsbaum Classifier trainiert auf 70% der Daten
 - Klassenwahrscheinlichkeiten geben Aufschluss über mögliche Label Verwechselungen.
- Ergebnisse:
 - 79 von 98 künstlich erzeugten Label Fehlern konnten entdeckt werden
 - Auf den original Daten wurden keine Fehler entdeckt.

	Predicted label errors	Predicted correct labels	Σ
Artificially introduced label errors	79 True Positives	19 False Negatives	98
Originallabels	1 False Positive	9719 True Negatives	9720
Σ	80	9738	9818

Metric:	Value:
Accuracy	0.9980
Precision	0.9875
Recall	0.8061
F1 Score	0.8876
Specificity	0.9999

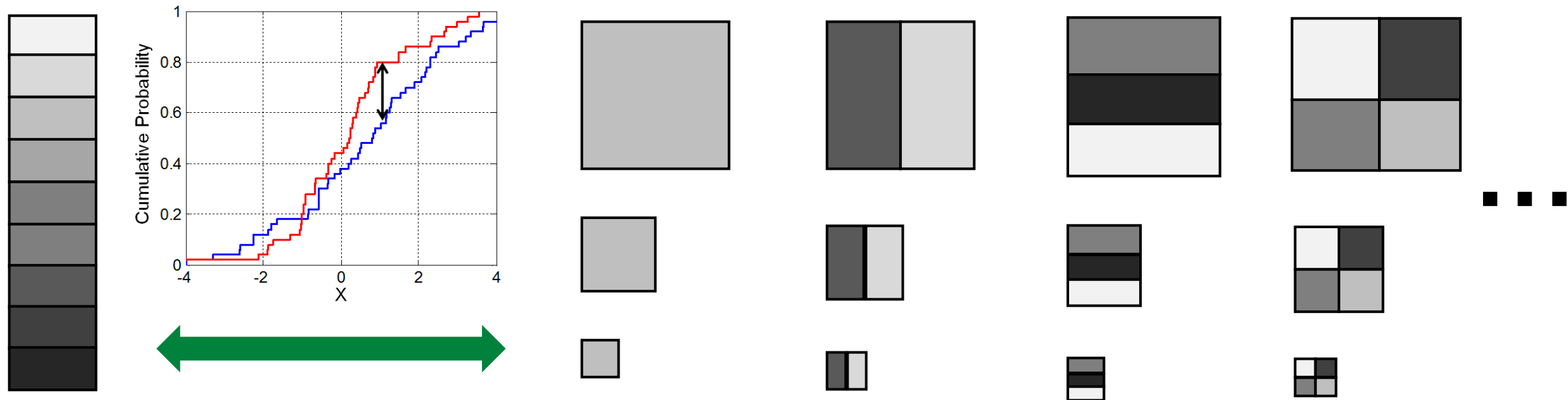
¹Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70, 1373-1411.

Beispiel Anwendung

Methode: Representativitätsanalyse der häufigsten Atemwegserkrankungen

- Ziel: Optimalen Tradeoff zwischen Menge und Vielfalt der Testdaten finden.

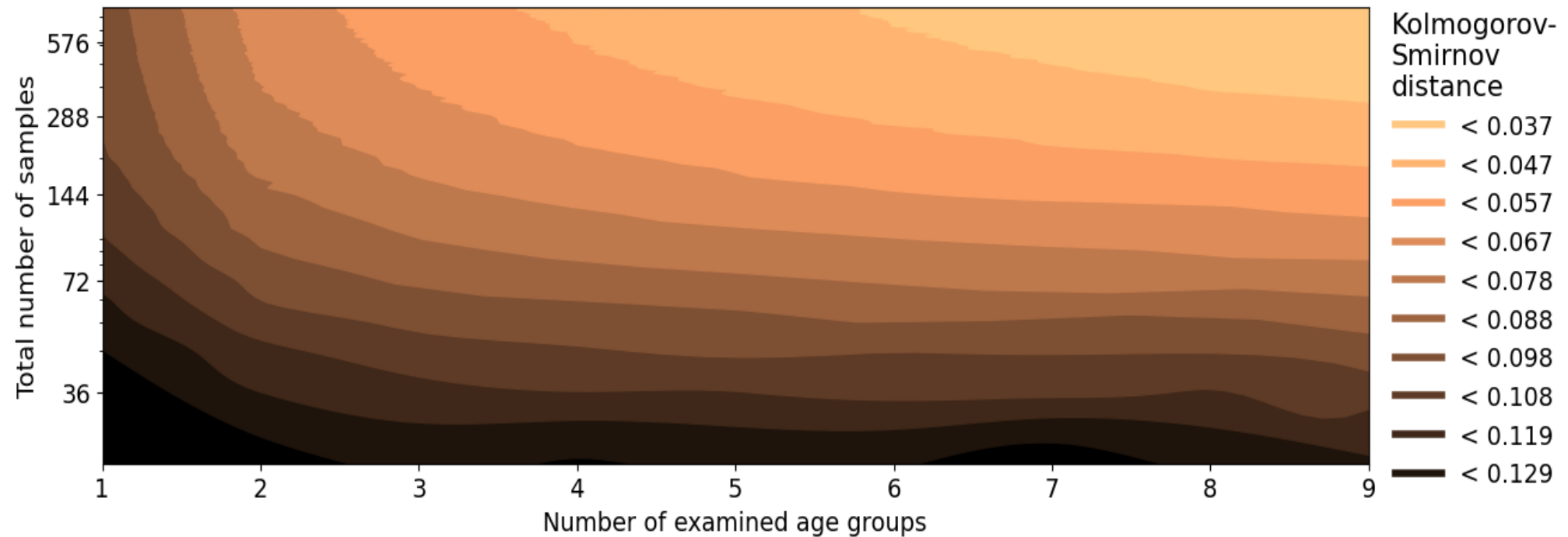
0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
-----	-------	-------	-------	-------	-------	-------	-------	-------



Beispiel Anwendung

Ergebnis: Repräsentativitätsanalyse der häufigsten Atemwegserkrankungen

- Ergebnis: Leitfaden für die Datensammlung





Vielen Dank für
Ihre Aufmerksamkeit.

Kontakt

Janek Groß
i3Mainz

E janek.gross@hs-mainz.de



Lisa Jöckel
Fraunhofer IESE

E lisa.joeckel@iese.fraunhofer.de



Dr. Michael Kläs
Fraunhofer IESE

E michael.klaes@iese.fraunhofer.de



Pascal Gerber
Fraunhofer IESE

E pascal.gerber@iese.fraunhofer.de

