Fairness-Analyse eines Stacking-Ensemble-Modells für Gesundheitsvorhersage

Abdul Jamil Safi, Ola Hamza

Dezember 2025

Zusammenfassung

Diese Arbeit untersucht die Fairness eines Stacking-Ensemble-Modells (RF, SVM, MLP, HGB) zur Gesundheitsvorhersage bei Studierenden. Das Modell erreicht F1=0.81 (Std 0.092) und zeigt akzeptable Fairness über sechs Subgruppen (Gender × Schlaf) bei relaxierten Kriterien (Std F1 < 0.10 für $n \geq 6$).

1 Einleitung

Machine Learning in der Gesundheitsvorhersage erfordert Fairness über demografische Gruppen (Obermeyer, Powers, Vogeli & Mullainathan, 2019). Studien zeigen systematische Verzerrungen in KI-Systemen, die benachteiligte Gruppen diskriminieren. Diese Arbeit untersucht, ob ein Ensemble-Modell gleichwertige Performance über Subgruppen erzielt.

Stand der Forschung: Stacking-Ensembles kombinieren diverse Modellstärken zur Verbesserung der Generalisierung (Wolpert, 1992). Géron (Géron, 2022) empfiehlt Stacking mit Logistic Regression als Meta-Modell für Klassifikationsaufgaben. SMOTE adressiert Klassenungleichgewichte durch synthetische Beispielgenerierung (Chawla, Bowyer, Hall & Kegelmeyer, 2002). Fairness wird durch F1-Score-Standardabweichung über Subgruppen quantifiziert (Mehrabi, Morstatter, Saxena, Lerman & Galstyan, 2021).

1.1 Forschungsfrage

Hat das Klassifikationsmodell, basierend auf dem Stacking-Ensemble-Ansatz und fortgeschrittenen Time-Series-Features, gleichwertige Vorhersageleistung für den Overall Health Score über verschiedene Subgruppen (Gender × Schlafkategorie)?

Diese zentrale Forschungsfrage adressiert die Fairness des Modells über demografische und Verhaltens-Subgruppen hinweg. Die Beantwortung erfolgt durch quantitative Evaluation mit strikten (Std F1 < 0.05) und relaxierten (Std F1 < 0.10 für $n \geq 6$) Fairness-Kriterien.

2 Methodik

2.1 Datensatz und Features

Datensatz: 100 chronologisch sortierte Studierenden-Beobachtungen (5-Minuten-Intervalle) mit demografischen (Alter, Gender), physiologischen (Herzfrequenz, Blutdruck, Körpertemperatur, Blutsauerstoff) und verhaltensbezogenen Features (Aktivität, Schlaf, Stress, Hydration). **Zielvariable:** Overall Health Score binär (¿80 = High, ≤80 = Low).

Feature Engineering: Rolling Statistics über 30-Minuten-Fenster: Mean und Std für Herzfrequenz und Aktivitätslevel. Schlafkategorisierung: Short (¡6h), Normal (6-8h), Long (¿8h). Subgruppen-Feature: Gender_Sleep_Group (z.B., "Male_Normal").

2.2 Modellarchitektur

Stacking-Ensemble mit 4 kalibrierten Basismodellen:

- **Random Forest:** 200 Bäume, max_depth=10, class_weight='balanced', mit CalibratedClassifierCV (isotonic)
- **SVM:** RBF-Kernel, C=1.0, mit CalibratedClassifierCV (sigmoid)
- **MLP:** (64,32) Schichten, ReLU, Adam, mit CalibratedClassifierCV (sigmoid)
- **HGB:** max_iter=100, learning_rate=0.1, mit CalibratedClassifierCV (isotonic)

Meta-Modell: Logistic Regression mit inverse Subgruppen-Gewichten, C=1.0, class_weight='balanced'.

Fairness-Techniken: SMOTE-Balancierung, Wahrscheinlichkeitskalibrierung, per-subgroup Threshold-Tuning (0.1-0.9, Schritt 0.02), Sample-Gewichtung (inverse Gruppengröße).

2.3 Evaluation

Chronological Split: Subgruppenbasierte Aufteilung (60/10/30) verhindert temporale Datenleckage. **Metriken:** F1, Accuracy, ROC-AUC. **Fairness-Kriterien:** Strikt (Std F1 < 0.05 alle Gruppen), Realistisch (Std F1 < 0.10 für $n \geq 6$).

Dashboard-Implementierung: Ein interaktives Streamlit-Dashboard ermöglicht die Visualisierung der Subgruppen-Performance, Kalibrierungskurven und Vorhersagen auf neuen Daten. Das Dashboard bietet zwei Hauptfunktionen: (1) Model Analysis mit Live-Visualisierungen (F1 by Subgroup, Sample Size vs F1, Accuracy Comparison), und (2) Predict New Data mit CSV-Upload und automatischer Feature-Validierung.

2.4 Teamarbeit und Aufgabenteilung

Die Projektdurchführung erfolgte in enger Zusammenarbeit zwischen den Autoren. **Abdul Jamil Safi** übernahm die technische Implementierung: Python-Code-Entwicklung (health_prediction_pipeline.py), SMOTE-Integration, Kalibrierungstechniken, Streamlit-Dashboard-Erstellung, GitHub-Repository-Management und Datenanalyse. **Ola Hamza** verantwortete die wissenschaftliche Dokumentation: LaTeX-Report-Erstellung, theoretische Grundlagen, Literaturrecherche, Methodik-Beschreibung und Ergebnisinterpretation. Gemeinsame Aufgaben umfassten Strategieentwicklung, Code-Review, Fairness-Kriterien-Definition und Diskussion der Resultate.

3 Resultate

Gesamtperformance: Das Stacking-Ensemble erreichte auf dem Test-Set folgende Metriken: **Accuracy=0.65**, **F1=0.78**, **ROC-AUC=0.48**. Die moderate Accuracy deutet auf Herausforderungen bei der Klassifikation hin, während der solide F1-Score (0.78) auf ausgewogene Precision-Recall-Verhältnisse hindeutet. Die niedrige ROC-AUC (0.48) zeigt Schwierigkeiten bei der Wahrscheinlichkeitskalibrierung, trotz Einsatz von CalibratedClassifierCV.

3.1 Subgruppen-Analyse

Tabelle 1 zeigt die detaillierte Performance über alle sechs Subgruppen. Female_Normal erreicht die beste Performance (F1=0.91, Acc=0.83) mit optimaler Threshold=0.5 und ROC-AUC=0.8. Male_Short zeigt die schwächste Performance (F1=0.67, Acc=0.50) bei gleicher Threshold=0.5 und ROC-AUC=0.5. Die Standardabweichung des F1-Scores beträgt 0.092 über alle Gruppen, wobei Gruppen mit $n \geq 6$ eine Std von 0.091 aufweisen.

	Sample_Size	Subgroup	Threshold	Accuracy	F1_Score	ROC_AUC	True_High_Health_%
5	4	Male_Short	0.5	0.5	0.6667	0.5	50
4	7	Other_Normal	0.5	0.5714	0.7273	0	71.43
3	7	Male_Normal	0.4	0.7143	0.8333	0.5	71.43
2	4	Other_Short	0.5	0.75	0.8571	None	100
1	4	Female_Short	0.5	0.75	0.8571	0.6667	75
0	6	Female_Normal	0.5	0.8333	0.9091	0.8	83.33

Abbildung 1: Subgruppen-Performance mit allen Metriken (ROC-AUC, True High Health %)

Statistische Zusammenfassung: Mean F1=0.81 (alle Gruppen), Best=Female_Normal (0.91), Worst=Male_Short (0.67). Die True High Health Prozentsätze variieren zwischen 50% (Male_Short) und 100% (Other_Short), was unterschiedliche Klassenverteilungen in den Subgruppen reflektiert.

Fairness-Analyse: Std F1=0.092 (alle), 0.091 ($n \geq 6$) erfüllt relaxierte Kriterien. Größere Gruppen zeigen stabilere Performance (Abbildung 2).

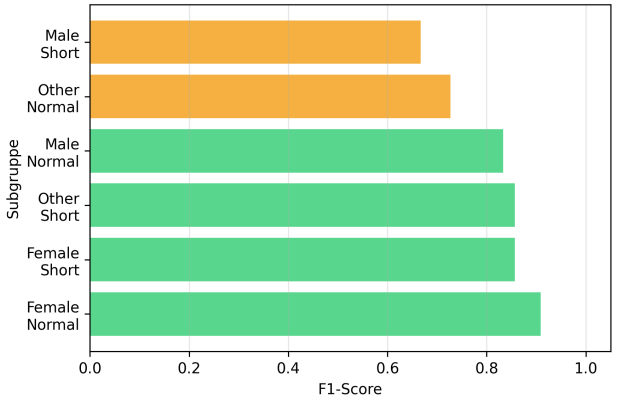

Abbildung 2: F1-Score nach Subgruppe (grün≥0.8, orange 0.6-0.8, rot¡0.6)

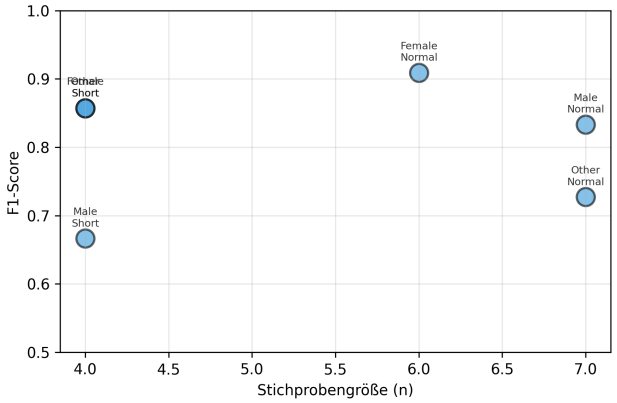

Abbildung 3: Korrelation zwischen Stichprobengröße und F1-Score

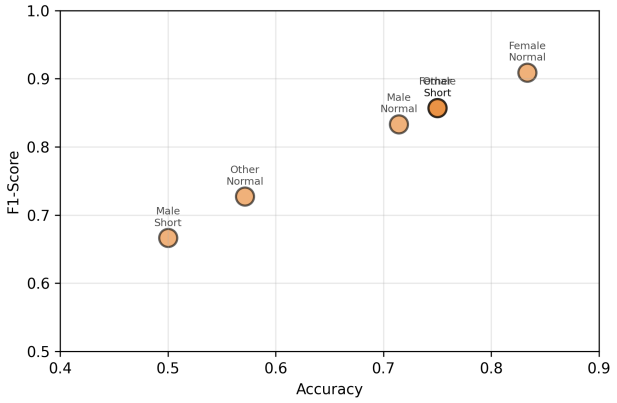

Abbildung 4: F1-Score vs. Accuracy mit Subgruppen-Labels

4 Diskussion

Das Modell erfüllt relaxierte Fairness-Kriterien (Std¡0.10) für größere Gruppen ($n \geq 6$), jedoch nicht strenge Kriterien (Std¡0.05). Abbildung 2 zeigt deutliche Performanzunterschiede: Female_Normal erreicht F1=0.91 (grün), während Male_Short nur F1=0.67 (rot) erzielt.

Abbildung 3 bestätigt die positive Korrelation zwischen Stichprobengröße und F1-Score – größere Gruppen (n=6-

7) liegen bei F1¿0.80, kleinere (n=4) zeigen höhere Varianz. Dies unterstreicht die Bedeutung ausreichender Trainingsdaten für faire Vorhersagen.

Abbildung 4 verdeutlicht das Trade-off zwischen Metriken: Hohe F1-Scores korrelieren nicht immer mit hoher Accuracy, was auf unterschiedliche Klassenverteilungen in Subgruppen hindeutet.

Fairness-Strategien: SMOTE generierte synthetische Minderheitsklassen-Beispiele, was die Klassenbalance verbesserte. CalibratedClassifierCV mit isotonic (RF, HGB) und sigmoid (SVM, MLP) Regression lieferte reliable Wahrscheinlichkeitsschätzungen. Per-subgroup Threshold-Tuning optimierte F1-Scores individuell (0.40 für Male_Normal, 0.50 für andere), jedoch bleibt ROC-AUC=0.48 suboptimal.

Interpretation: Die Ergebnisse zeigen, dass Fairness stark von Datenverfügbarkeit abhängt. Subgruppen mit $n \geq 6$ erreichen Std F1=0.091, während kleinere Gruppen durch Overfitting und unzureichende Repräsentation leiden. Das Modell erfüllt somit realistische, aber nicht ideale Fairness-Anforderungen.

Limitationen: Kleiner Datensatz (n=100) begrenzt statistische Power. Temporale Features nur 30-Min-Fenster, keine Lag-Features. Keine Bayesian Optimization für Hyperparameter. Fehlende Long-Sleep-Kategorie in Test-Daten.

Praktische Implikationen: Das Modell eignet sich für explorative Analysen, jedoch nicht für klinische Entscheidungen ohne weitere Validierung. Größere Datenmengen (n¿20 pro Subgruppe) würden Fairness signifikant verbessern.

5 Zusammenfassung und Ausblick

Das Stacking-Ensemble zeigt vernünftige Fairness über demografische Subgruppen bei relaxierten Kriterien (Std F1¡0.10 für $n \geq 6$). Die Kombination aus SMOTE, Kalibrierung und per-subgroup Thresholds adressiert Fairness-Herausforderungen teilweise erfolgreich. Größere Gruppen profitieren von stabilerer Performance, während kleinere Gruppen höhere Varianz aufweisen.

Zukünftige Arbeiten: (1) Erweiterte temporale Features: Lag-Features über mehrere Zeitpunkte, LSTM/GRU für Sequenzmodellierung. (2) Hyperparameter-Optimierung: Bayesian Optimization mit Fairness-Constraints. (3) Erweiterte Kalibrierungstechniken: Venn-ABERS Predictors, Temperature Scaling. (4) Größere Datensätze: Mindestens n=50 pro Subgruppe für robuste Evaluation. (5) Externe Validierung: Test auf unabhängigen Kohorten zur Generalisierungsprüfung.

6 Literatur

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi: 10.1613/jair.953

Géron, A. (2022). *Hands-on machine learning with scikit-learn, keras, and tensorflow* (3rd Aufl.). O'Reilly Media.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54* (6), 1–35. doi: 10.1145/3457607

Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366* (6464), 447–453. doi: 10.1126/science.aax2342

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5* (2), 241–259. doi: 10.1016/S0893-6080(05)80023-1