A3 ROC-Kurven

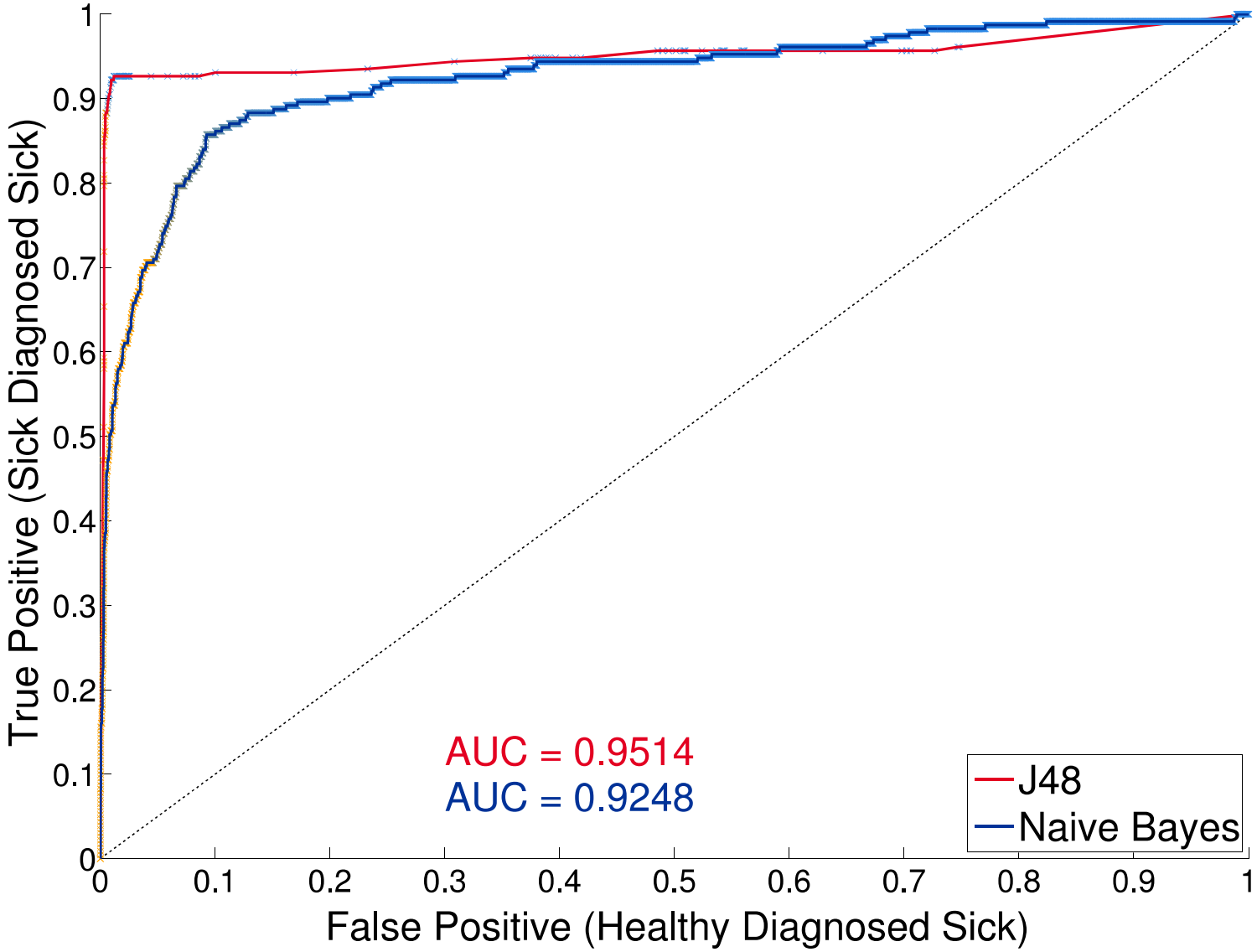
TECHNISCHE UNIVERSITÄT DARMSTADT

- gewählter Datensatz: ”sick.arff”
- der Datensatz bezieht sich auf Schilddrüsenkrankheiten (Thyroid disease) [1]
- klassifizieren von Personen in die Klassen ”krank” (sick) und ”gesund” (negative)
- Klassifizierer
  - J48 (Entscheidungsbaum)
  - Naive Bayes

[1] Thyroid disease records supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Syndney, Australia

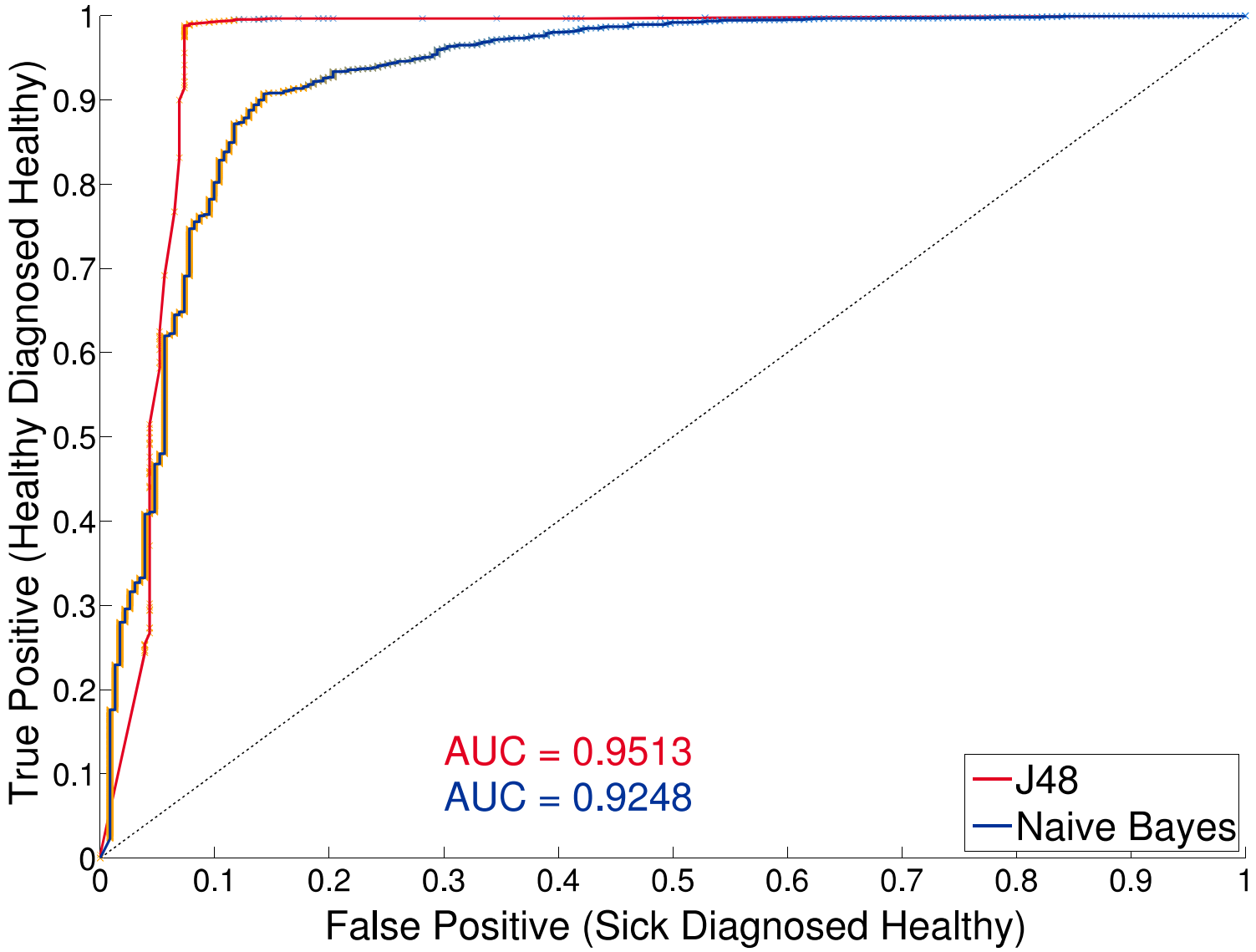
A3 ROC-Kurven
Klasse sick

TECHNISCHE UNIVERSITÄT DARMSTADT



A3 ROC-Kurven
Klasse negative

TECHNISCHE UNIVERSITÄT DARMSTADT



A3 ROC-Kurven Interpretation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Bei der Bewertung eines solchen Datensatzes muss man sich grundsätzlich die Frage stellen, ob es schlimmer ist eine gesunde Person als krank einzustufen, oder eine kranke Person als gesund
- ROC Kurven erlauben es den Threshold eines Klassifizierers zu wählen
  - für unseren Datensatz zum Beispiel so, dass verhindert wird dass kranke Personen unbehandelt bleiben

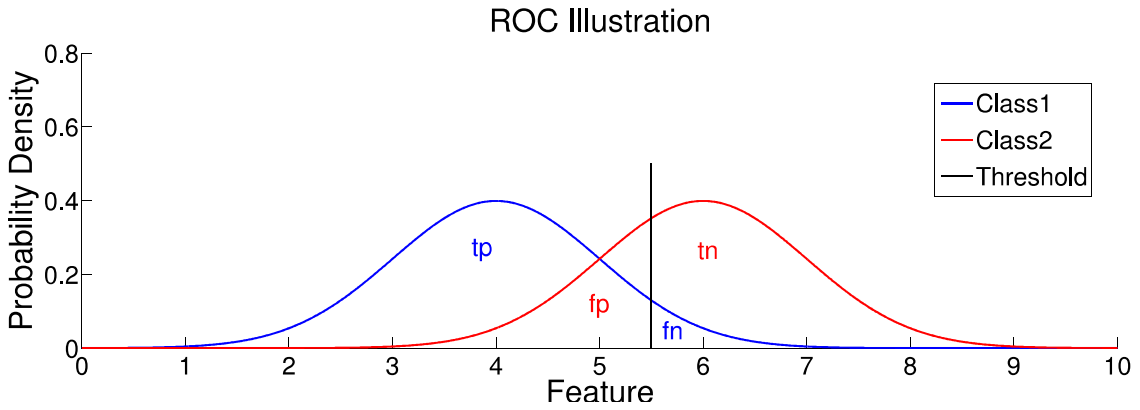


Abbildung : Die Form einer ROC Kurve wird bestimmt von der Überlappung. Je weniger Überlappung vorhanden ist desto besser kann der Klassifizierer arbeiten.

A3 ROC-Kurven
Interpretation - J48

TECHNISCHE UNIVERSITÄT DARMSTADT

- J48 klassifiziert diesen Datensatz sehr gut
  - nach einem sehr steilen Anstieg fast keine Verbesserung mehr bei den true positives (sick diagnosed sick)
  - eckiger Kurvenverlauf
  - keine größeren Höhlungen (concavities)
  - bereits bei 1% false positives (healthy diagnosed sick) werden 93% true positives (sick diagnosed sick) klassifiziert

A3 ROC-Kurven
Interpretation - Naive Bayes

TECHNISCHE UNIVERSITÄT DARMSTADT

- Naive Bayes klassifiziert den Datensatz auch sehr gut, aber etwas schlechter als J48
  - Kurve verläuft runder und etwas weniger steil
  - ebenfalls keine größeren Höhlungen (concavities)
  - bei 10% false positives (healthy diagnosed sick) werden jedoch auch bereits fast 85% true positives (sick diagnosed sick) klassifiziert
  - erst bei ungefähr 40% false positives (healthy diagnosed sick) ist eine ähnliche Anzahl an true positives (sick diagnosed sick) wie bei J48

A3 ROC-Kurven
Interpretation - AUC

TECHNISCHE UNIVERSITÄT DARMSTADT

- Die Fläche unter der ROC-Kurve (AUC) kann als Wahrscheinlichkeit gesehen werden
  - P(P(sick person = sick) $>$ P(healthy person = sick))
  - wenn ein zufälliges positives Sample und ein zufälliges negatives Sample genommen werden ist die Wahrscheinlichkeit, dass das positive Sample als solches klassifiziert wird höher als dass das negative als positiv klassifiziert wird
- die Fläche unter der ROC-Kurve ist bei J48 etwas höher als bei Naive Bayes
- J48 hat eine Fläche von 0.9514 und Naive Bayes eine von 0.9248
- insgesamt gesehen sind also beide Klassifizierer bei diesem Datensatz sehr gut