

# Einstieg in die Untersuchung der Erklärbarkeit von Objektklassifikation auf ML-Basis am Beispiel eines Barcode-Detektors

---

Semih Kasap

28. Mai 2020

Institut der Pathologie, Charité Berlin  
Hochschule für Technik und Wirtschaft Berlin

# Überblick

- Vertrauen
- Ansätze
- Werkzeuge
- Anwendung auf den Barcode-Detektor

# Vertrauen

---

# Warum das Ganze?

**Sorge:** Wenn der Anwender einem KI-Modell nicht vertraut, wird dieser es nicht nutzen.

**Vertrauen**, emotionale Sicherheit, einem anderen Menschen und dem eigenen Dasein offen gegenübertreten und sich hingeben zu können(...).

Brockhaus, Vertrauen.

<http://brockhaus.de/ecs/enzy/article/vertrauen> (aufgerufen am 2020-04-28)

im KI-Kontext:

1. Einer **Vorhersage** vertrauen (z. B., ob der Anwender einer individuellen Vorhersage vertraut, um auf dieser Grundlage Aktionen zu tigen)
2. Einem **Modell** vertrauen (z. B., ob der Anwender sein Verhalten nachvollziehbar auf Grundlage seines Vertrauens in ein KI-Modells tigt, nachdem das Modell fr den Produktiveinsatz freigegeben wurde)

- Beide Definitionen stehen im direkten Verhältnis zum Verständnis eines Menschen in das Verhalten des Modells
- **Problem:** Wie kann dafür gesorgt werden, dass der Mensch das Modell nicht mehr als Black-Box sieht?

- 1. Ermitteln des Vertrauens auf eine individuelle Vorhersage ist fundamental, wenn das Modell zur Entscheidungsfindung genutzt wird
- In kritischen Anwendungsbereichen (Medizin oder Terrorismusbekämpfung) dürfen Vorhersagen nicht blind hingenommen werden, da Konsequenzen katastrophal werden können.

- **2. Evaluierung des Modells als Ganzes** bevor es „auf freier Wildbahn“ eingesetzt werden kann
- Der Anwender sollte sich auf das Modell verlassen können bzw. das Modell muss gute Ergebnisse bei realen Daten in Bezug auf den Metrics-of-Interest erzielen

## Evaluierung eines Modells heute?

---

- oft werden Genauigkeitsmetriken mithilfe eines verfügbaren Testdatensatzes ermittelt
- die reale Welt ist oft anders
- die gewohnten Evaluierungspraktiken können meist nicht das Ziel des Modells untermauern

## Was bedeutet das für uns?

---

- das Untersuchen individueller Vorhersagen und deren Erklärungen sind **zusätzlich** nötig für eine vertrauensvolle Lösung
- der Anwender muss unterstützt werden bei der Frage, welche Instanzen näher betrachtet werden müssen - vor allem bei großen Datensätzen

## Ansätze

---

# Fragestellungen

---

- interaktive vs. **statische** Erklärungen
- Verstehen der Daten oder des **Modells**?
- Erklärungen für das übergreifende Verhalten (global) oder **individuelle Beispiele (lokal)**?
- direkt interpretierbares Modell oder **nachträgliche Analyse**?
- Erklärungen basierend auf Beispielen oder **Features**?
- *Hinweis: Die relevanten Entscheidungen für den nachfolgenden Verlauf dieser Arbeit sind fett gedruckt.*

## Welche Ansätze passen?

---

- Contrastive Explanations Method (CEM) oder „CEM with Monotonic Attribute Functions“ (CEM-MAF)
- Local Interpretable Model-Agnostic Explanations[1] (LIME)
- SHapley Additive exPlanations (SHAP)

## Local Interpretable Model-Agnostic (LIME)

---

- konzentriert sich auf das Problem, auf eine individuelle Vorhersage zu vertrauen (Trusting-a-Prediction)
- ein Algorithmus, das Vorhersagen aller Klassifizierer oder Regressor, indem es diese Vorhersagen mit einem interpretierbaren Modell lokal annähert
- die Auswahl mehrerer individueller Vorhersagen werden als Lösung für das „Trusting-a-Model“-Problem herangezogen

# Werkzeuge

---

- AI Explainability 360 Toolkit: Eine Ansammlung von mehreren Erklärungsalgorithmen in einer Library gekapselt
- Python 3.6 und Anaconda

Anwendung auf den  
Barcode-Detektor (siehe  
Forschungsprojekt A+B)

---

# Problem

---

- tagtäglich kommen mehrere Objektträger am Institut für Pathologie der Charité Berlin an
- diese werden vorbereitet und gescannt
- vor jedem hochauflösenden Scan werden Bilder mit einer Handels-üblichen Digitalkamera aufgenommen
- **Fragestellung: Wie kann die Zuordnung der Objektträger auf eine Patientenakte automatisiert vorgenommen werden?**

# Lösung

---

- zwei Ansätze werden verfolgt: YOLOv3 und Faster R-CNN
- beide Ansätze werden mit jeweils vier verschiedenen Trainingsdatensätzen trainiert, validiert und getestet, um anschließend beide konkurrierenden Ansätze unter Verfolgung der Fragestellung auszuwerten/zu vergleichen

# Lösung

---

- da der YOLOv3-Ansatz mit Microsoft-Technologien bewerkstelligt ist (C# und Darknet als Library), das AIX360-Toolkit jedoch in Python implementiert ist, wird lediglich der Faster R-CNN-Ansatz erklärt
- es werden zuerst 15 Bilder aus den vier Datensätzen für die Erklärungen ausgewählt, die ein möglichst vertrauenswürdiges Bild herstellen können

# Kriterien für die Auswahl der Testbilder

- Ausgangsklasse: 1D-Barcodes/Strichcodes, 2D-Barcodes/Data-Matrix-Codes
- Qualität des Drucks des Barcodes: klar sichtbar, verschwommen, Fehler am gedruckten Barcode durch falsch kalibrierte oder schlechte Toner, teilweise mit Post-Its verdeckter Barcode
- Qualität des Bildes: klar oder verschwommen (z. B. durch falschen Fokus an der Digitalkamera oder am Scanner)
- Vorhandensein von Handschrift: frei von Handschrift oder vom Laborpersonal signierte Objektträger
- Färbung: diverse Färbungen des zu scannenden Objektes selbst (z. B. HE-Färbung)

- im nachfolgenden werden die 15 Bilder mitsamt ihren Erklärungen dargestellt
- pro individuelle Vorhersage werden vier Bilder angezeigt:
  - 1. Bild: Umrandung der aussagekräftigsten fünf positiven Features
  - 2. Bild: dasselbe, Rest wird ausgeblendet
  - 3. Bild: Anzeige der aussagekräftigsten zehn positiven (grün) und ggf. negativen (orange) Features
  - 4. Bild: Anzeige der 1000 aussagekräftigsten positiven (grün) und ggf. negativen (orange) Features mit einer Mindestgewichtung von 0,1

# Bilder



**Abbildung 1:** Data-Matrix-Code, klar gedruckter Barcode, gute Bildqualität, signiert

# Bilder



Abbildung 2: 1D Barcode, klar gedruckter Barcode, gute Bildqualität, unsigniert

# Bilder

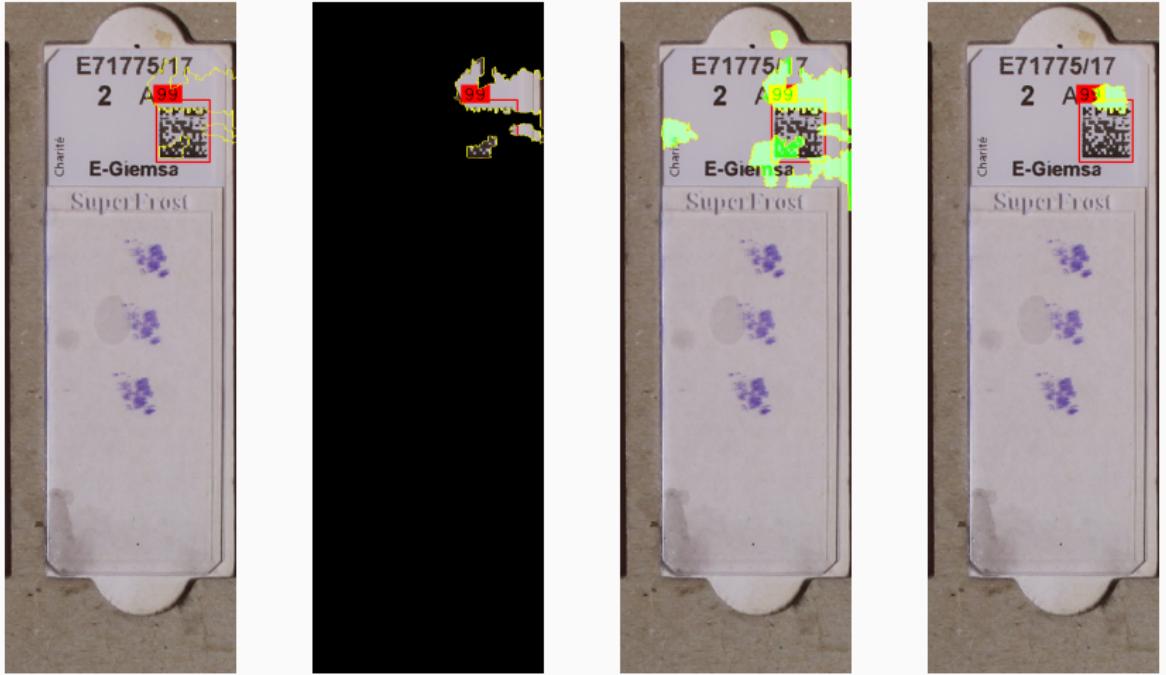


Abbildung 3: Data-Matrix-Code, schlechter Barcode-Druck, gute Bildqualität, unsigniert

# Bilder



**Abbildung 4:** Data-Matrix-Code, verschwommener Barcode, gute Bildqualität, unsigniert

# Bilder

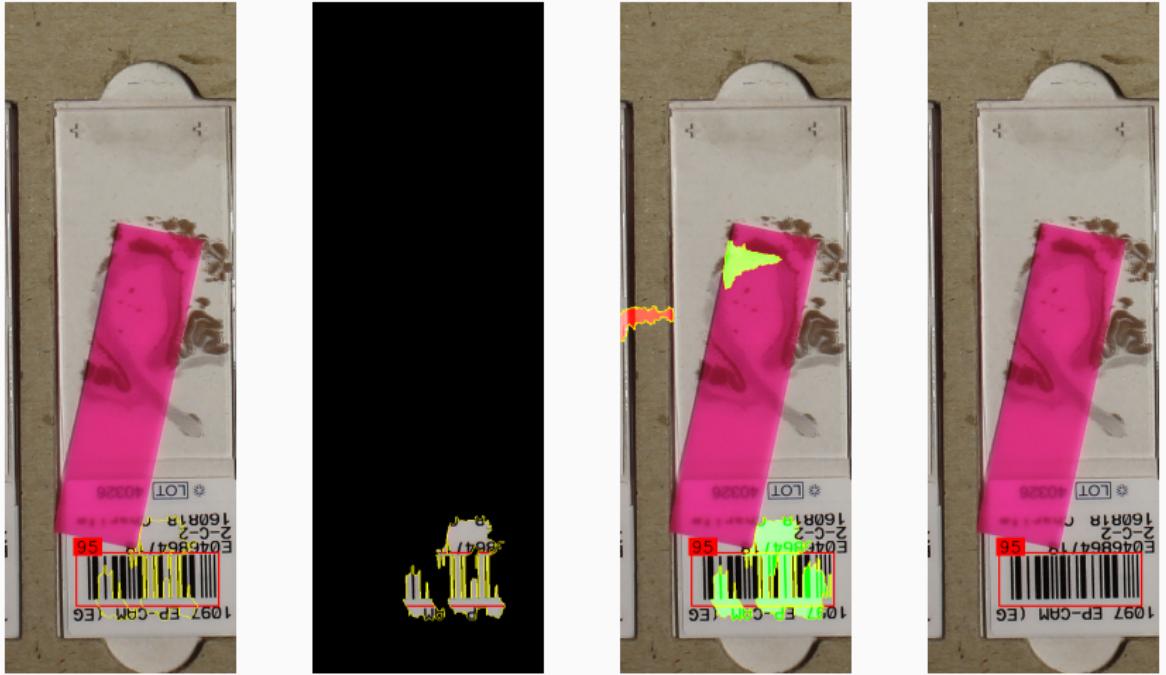


Abbildung 5: 1D Barcode, teilweise verdeckter Barcode, gute Bildqualität, signiert

# Bilder



Abbildung 6: 1D Barcode, klar gedruckter Barcode, gute Bildqualität, signiert

# Bilder



Abbildung 7: 1D Barcode, klar gedruckter Barcode, schlechte Bildqualität, unsigniert

# Bilder

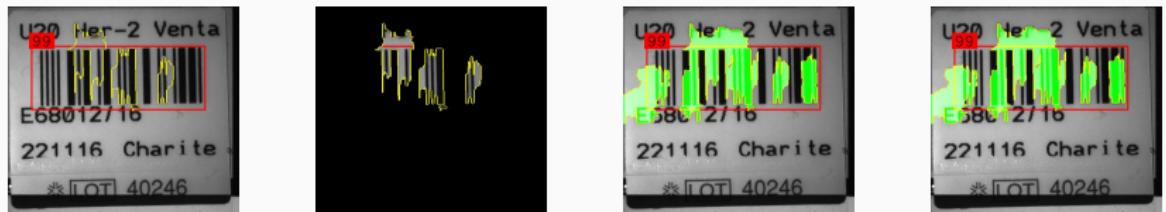


Abbildung 8: 1D Barcode, klar gedruckter Barcode, gute Bildqualität, unsigniert

# Bilder

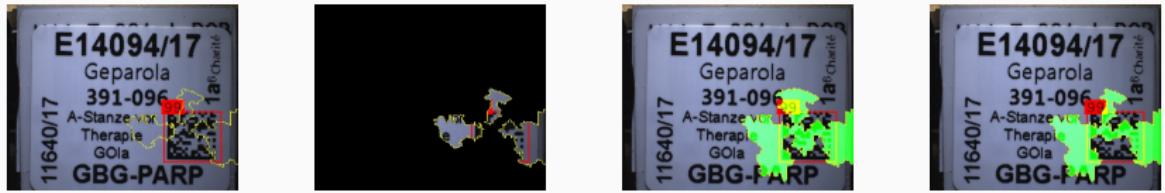


Abbildung 9: Data-Matrix-Code, klar gedruckter Barcode, gute Bildqualität, unsigniert

# Bilder

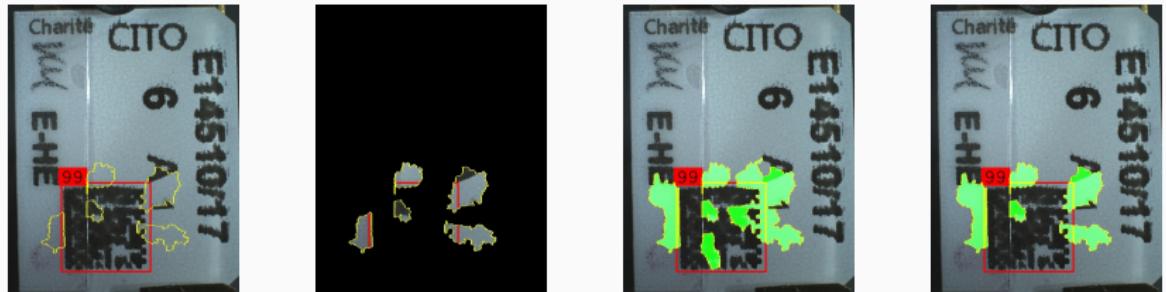


Abbildung 10: Data-Matrix-Code, verschwommener Barcode, gute Bildqualität, signiert

# Bilder

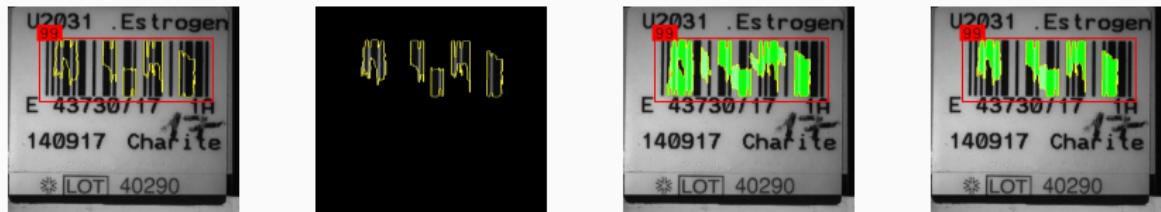


Abbildung 11: 1D Barcode, klar gedruckter Barcode, gute Bildqualität, signiert

# Bilder



**Abbildung 12:** Data-Matrix-Code, klar gedruckter Barcode, schlechte Bildqualität, unsigniert

# Bilder



Abbildung 13: 1D Barcode, klar gedruckter Barcode, schlechte Bildqualität, unsigniert

# Bilder

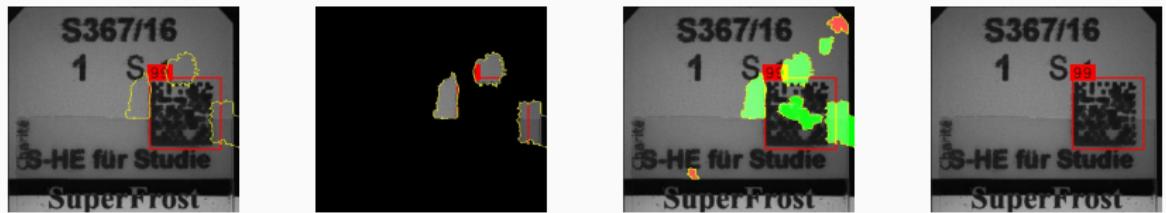


Abbildung 14: Data-Matrix-Code, verschwommener und teilweise überdeckter Barcode, gute Bildqualität, unsigniert

# Bilder

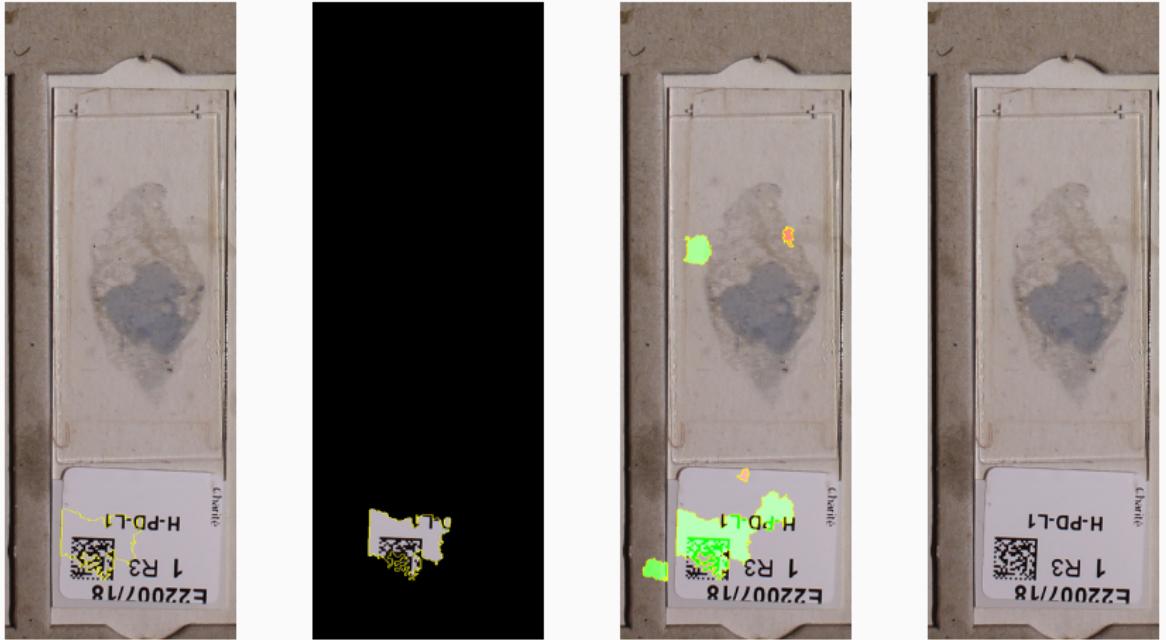


Abbildung 15: Data-Matrix-Code (nicht erkannt), klar gedruckter Barcode, gute Bildqualität, unsigniert

Fragen?

## Literatur i

-  Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin.  
"why should I trust you?": Explaining the predictions of  
any classifier.  
CoRR, abs/1602.04938, 2016.