Forklarbar AI (XAI)

ELMED219: Momentliste X01–X08

ELMED219

Vår 2026

Oversikt

1. **Hvorfor forklarbarhet?**
   - X01: Forklare hvorfor forklarbarhet er viktig i medisinsk AI
2. **Typer forklarbarhet**
   - X02: Skille mellom global og lokal forklarbarhet
   - X03: Skille mellom ante-hoc og post-hoc forklarbarhet
3. **XAI-metoder**
   - X04: Beskrive SHAP (SHapley Additive exPlanations)
   - X05: Beskrive LIME (Local Interpretable Model-agnostic Explanations)
   - X06: Forklare Grad-CAM for CNN-visualisering
4. **Begrensninger og utfordringer**
   - X07: Diskutere begrensninger ved XAI-metoder
   - X08: Kjenne til attention-visualisering i LLM

X01: Forklare hvorfor forklarbarhet er viktig i medisinsk AI

Utfordringen med "black box" AI:

- Komplekse modeller (dype nettverk, LLM) gir prediksjoner uten å forklare *hvorfor*
- I medisin er dette problematisk av flere grunner

Hvorfor forklarbarhet er kritisk i medisin:

1. **Tillit:** Leger og pasienter må stole på anbefalingene
2. **Ansvar:** Ved feil – hvem har ansvaret?
3. **Regulering:** EU AI Act krever forklarbarhet for høyrisiko-AI
4. **Læring:** Forstå hva modellen har lært (og feil-lært)
5. **Feilsøking:** Identifisere bias og svakheter
6. **Klinisk integrasjon:** AI som beslutningsstøtte, ikke autonom

Medisinsk praksis

En lege som ikke kan forklare hvorfor de anbefaler en behandling vil miste pasientens tillit. Det samme gjelder AI.

X02: Skille mellom global og lokal forklarbarhet

Global forklarbarhet:

- Forklarer modellens **generelle oppførsel**
- "Hvilke features er viktigst totalt sett?"
- "Hvordan påvirker alder prediksjonen generelt?"

Metoder:

- Feature importance
- Partial dependence plots
- Globale SHAP-verdier

Lokal forklarbarhet:

- Forklarer én **spesifikk prediksjon**
- "Hvorfor fikk *denne* pasienten høy risikoscore?"
- "Hvilke piksler påvirket klassifiseringen?"

Metoder:

- LIME
- SHAP (individuelle verdier)
- Grad-CAM (for bilder)

Begge er viktige

Global: Forstå modellen som helhet, validere mot domenkunnskap
Lokal: Forklare beslutninger til pasienter og kollegaer

X03: Skille mellom ante-hoc og post-hoc forklarbarhet

Ante-hoc (innebygd):

- Forklarbarhet **bygget inn** i modellen
- Modellen er designet for å være tolkbar
- "Interpretable by design"

Eksempler:

- Lineær regresjon (koeffisienter)
- Beslutningstrær (regler)
- Attention-vekter (delvis)

Fordel: Ekte forklaring
Ulempe: Ofte mindre nøyaktig

Post-hoc (etterpå):

- Forklaringsmetode **lagt til** etter trening
- Modellen er fortsatt en "black box"
- Separate verktøy for tolkning

Eksempler:

- SHAP, LIME
- Grad-CAM
- Saliency maps

Fordel: Kan brukes på komplekse modeller
Ulempe: Approksimerer, kan være misvisende

Trade-off

Ofte et valg mellom tolkbarhet (ante-hoc) og ytelse (kompleks modell + post-hoc)

X04: Beskrive SHAP (SHapley Additive exPlanations)

SHAP: Basert på spillteori

- Shapley-verdier: Fair fordeling av "gevinst" mellom spillere
- Her: Hvor mye bidrar hver feature til prediksjonen?

Hvordan SHAP fungerer:

1. For hver feature: Beregn gjennomsnittlig bidrag til prediksjon
2. Tar hensyn til alle mulige kombinasjoner av features
3. Gir en verdi per feature for hver prediksjon

Styrker:

- Teoretisk solid (Shapley-aksiomer)
- Konsistent og lokalt nøyaktig
- Fungerer for alle modelltyper

Visualiseringer:

- **Waterfall plot:** Én prediksjon
- **Summary plot:** Global oversikt
- **Force plot:** Interaktiv

Python

```
import shap; explainer = shap.Explainer(model); shap_values = explainer(X)
```

X05: Beskrive LIME (Local Interpretable Model-agnostic Explanations)

LIME: Lokal tilnærming med enkel modell

Hovedidé:

1. Velg én prediksjon å forklare
2. Generer **perturberte** (forstyrrede) versjoner av input
3. Kjør den komplekse modellen på alle versjoner
4. Tren en **enkel, tolkbar modell** (f.eks. lineær) lokalt
5. Bruk den enkle modellen til å forklare

Styrker:

- Modell-agnostisk
- Intuitivt forståelig
- Fungerer for tekst, bilder, tabeller

Svakheter:

- Ustabil – ulike kjøringer kan gi ulike svar
- Valg av nabolag er vilkårlig
- Kan være misvisende

Eksempel: Tekstklassifisering

LIME viser hvilke ord som bidro mest til klassifiseringen "spam" vs. "ikke spam"

X06: Forklare Grad-CAM for CNN-visualisering

Grad-CAM: Gradient-weighted Class Activation Mapping

- Spesialisert for **konvolusjonelle nettverk (CNN)**
- Viser **hvilke bildområder** som påvirket klassifiseringen

Hvordan det fungerer:

1. Beregn gradienter av output m.h.p. feature maps i siste konvolusjonslag
2. Vekt feature maps med gjennomsnittlig gradient
3. Kombiner til et **heatmap** som overlays på originalbildet

Medisinsk anvendelse:

- Røntgen/CT: "Modellen fokuserte på dette området for lungefunn"
- Dermatologi: "Disse pikslene påvirket melanom-diagnosen"
- Validering: Sjekk at modellen ser på riktig område!

Viktig innsikt

Grad-CAM kan avsløre om modellen bruker **spuriøse korrelasjoner** (f.eks. ser på tekst i bildet, ikke patologi)

X07: Diskutere begrensninger ved XAI-metoder

XAI er ikke perfekt – viktige begrensninger:

1. **Forklaringer er approksimeringer**
   - Post-hoc metoder forklarer ikke den *ekte* modellen
   - Kan være misvisende eller inkonsistente
2. **Ustabilitet**
   - LIME kan gi ulike forklaringer for samme input
   - Små endringer i input kan gi store endringer i forklaring
3. **Fortolkningsbyrde**
   - Hvem skal tolke forklaringene? Krever ekspertise
   - Risiko for overtillit til forklaringer
4. **Beregningsmessig kost**
   - SHAP kan være svært treg for store modeller

Husk

XAI er et verktøy for innsikt, ikke en garanti for at modellen er "riktig" eller rettferdig.

X08: Kjenne til attention-visualisering i LLM

Attention-visualisering i LLM: Innebygd "forklaring"?

- Transformer-modeller bruker **self-attention**
- Attention-vekter viser hvilke tokens modellen "ser på"
- Kan visualiseres som heatmaps

Eksempel:

*"Pasienten har **diabetes** og tar **metformin**"*

Attention-visualisering kan vise at "metformin" har høy attention på "diabetes"

Muligheter:

- Innsikt i modellens fokus
- Ante-hoc (innebygd i modellen)
- Lett tilgjengelig

Begrensninger:

- Attention $\neq$ forklaring
- Mange lag, mange "heads"
- Kan være vanskelig å tolke
- Viser korrelasjon, ikke kausalitet

Forsiktighet

Attention-visualiseringer bør brukes som supplement, ikke som endelig forklaring.

Oppsummering: Forklarbar AI (XAI)

Nøkkelpunkter:

- **X01:** Forklarbarhet er kritisk i medisin (tillit, ansvar, regulering)
- **X02:** Global vs. lokal forklarbarhet
- **X03:** Ante-hoc (innebygd) vs. post-hoc (etterpå)
- **X04:** SHAP – spillteoretisk, konsistent
- **X05:** LIME – lokal tilnærming, modell-agnostisk
- **X06:** Grad-CAM – visualiser CNN-fokus i bilder
- **X07:** Begrensninger – approksimeringer, ustabilitet
- **X08:** Attention-visualisering i LLM

Praktisk i Lab 2 og Lab 3

- Lab 2: Bruk Grad-CAM for å tolke CNN-klassifisering
- Lab 3: Diskuter forklarbarhet og transparens i LLM