Å gjøre kunstig intelligens til mer enn en svart boks

Transparente maskinlæringssystemer for medisinske anvendelser

Andrea Theomine Marheim Storås

Simula Metropolitan Center for Digital Engineering AS (SimulaMet)
E-post *andreatheomine@gmail.com*

TITTEL
Beyond the Black Box: Transparent Machine Learning Systems for Medical Applications
ISBN: 978-82-8364-588-0
https://hdl.handle.net/11250/3155666

VEILEDERE
Pål Halvorsen, SimulaMet; Michael A. Riegler, SimulaMet; Tor P. Utheim, Oslo universitetssykehus, og Inga Strümke, NTNU.

STED OG TIDSPUNKT FOR DISPUTAS
Oslomet – storbyuniversitetet, vår 2024

HOVEDBUDSKAP

Maskinlæring har stort potensial innen helse, men mangel på forståelse av hvordan teknologien fungerer kan begrense bruken.

Vi undersøker og utvikler metoder for transparente maskinlæringssystemer.

Tett samarbeid med helsepersonell er viktig for at teknologien blir anerkjent og brukt i klinikk.

BAKGRUNN OG HENSIKT

Kunstig intelligens (KI) og maskinlæring (ML) utgjør en stadig større del av hverdagen vår. ChatGPT oppnådde raskt høy popularitet i årsskiftet 2022/2023 (1) og bidro til å tilgjengeliggjøre ML-teknologi for allmennheten. KI kan defineres som informasjonsteknologi som gir datamaskiner evne til å løse oppgaver på en intelligent måte. ML er en type KI der datamaskiner mottar treningsdata for å lære å løse oppgavene uten at de eksplisitt fortelles hvordan. Innen helse har ML-modeller vist imponerende resultater for å løse mange oppgaver, for eksempel sykdomsdiagnostisering (2), prediksjon av proteinstrukturer (3) og persontilpasset behandling (4). ML har derfor et stort potensial for å bidra til å effektivisere helsevesenet ved å finne sammenhenger i og tolke store datamengder, samt assistere helsepersonell.

Figur 1 illustrerer komponentene som inngår i et medisinsk ML-system. Et sentralt element er ML-modellen som analyserer dataene og kommer med prediksjoner (komponent 3). Selv om ML-modeller løser enkelte oppgaver bedre enn mennesker, er det ofte uvisst nøyaktig hvordan modellene tar beslutninger. De blir derfor omtalt som svarte bokser. At ML-modeller er svarte bokser, bidrar til begrenset bruk innen høyrisikoområder som medisin. For å åpne de svarte boksene kan metoder som forklarer ML-modellene og deres beslutninger benyttes. Likevel er det mindre klart hvor godt forklaringsmetodene fungerer innen medisin. Datagrunnlaget modellen er trent på (komponent 1) er en annen viktig komponent i ML-systemet. Dersom datasettet modellen utvikles på avviker mye fra dataene modellen analyserer i praksis, kan modellen gjøre det dårligere enn først antatt. Derfor bør treningsdatasettet være tilgjengelig for inspeksjon. Preprosessering av data (komponent 2) inkluderer bruk av generative ML-modeller til å produsere syntetiske data som et alternativ eller supplement til originale helsedata. I tillegg til å tilgjengeliggjøre dataene ML-modellen er evaluert på, anbefales det å rapportere en samling av evalueringsmetrikker for et nyansert bilde av modellens styrker og svakheter (komponent 5).

Til slutt er brukernes preferanser essensielle for en vellykket implementering av ML-systemer (komponent 6). Dersom brukere som helsepersonell og pasienter ikke forstår hvordan ML-systemet fungerer eller hvordan det kan implementeres i arbeidsflyten, kan dette redusere tilliten og villigheten til å bruke systemet.

Forskningsspørsmålet doktorgradsarbeidet undersøker er: «Hva skal til for å gjøre ML-systemer nyttige for medisinske anvendelser?» Vi argumenterer for at transparens i hele systemet er nøkkelen. For å oppnå transparens bør alle komponentene være tilstrekkelig beskrevet og forklart til brukerne av systemet.

MATERIALE OG METODER

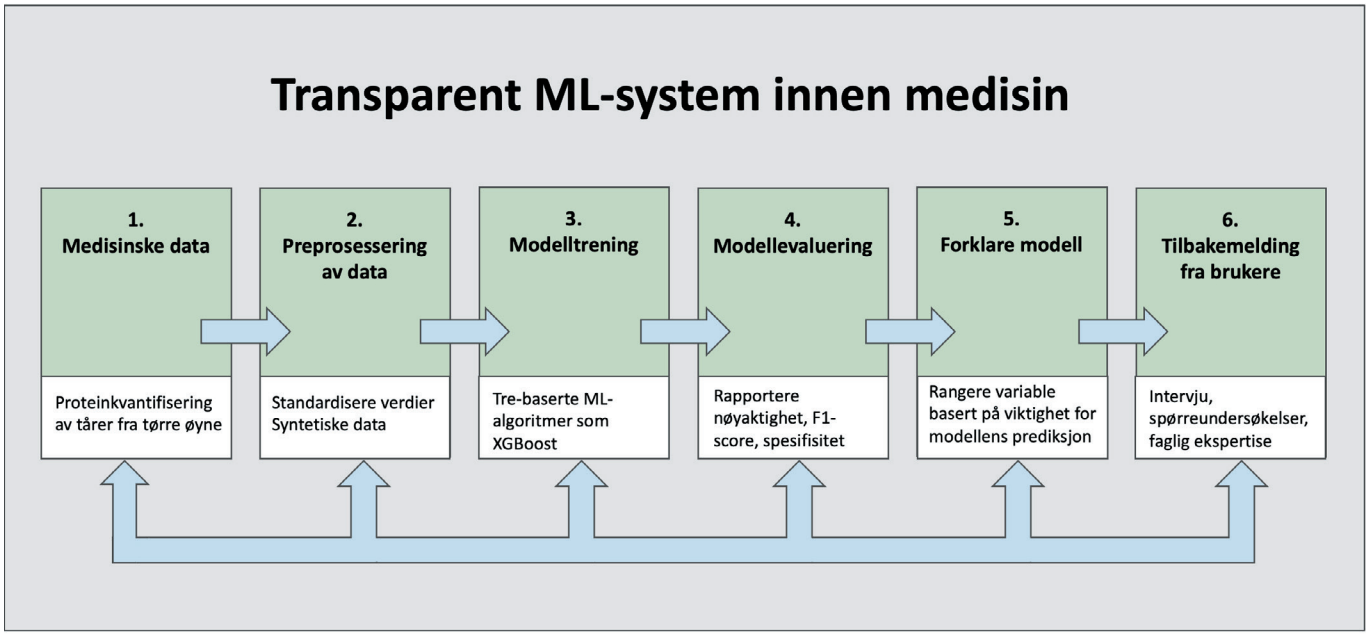
Litteratursøk ble utført for å få en oversikt over ML og hvordan ML-modeller blir forklart innen helse. For å trene og evaluere ML-modellene ble helsedata fra følgende syv medisinske områder benyttet: tørre øyne, kunstig befruktning, gastroenterologi, hode- og halskreft, nyretransplantasjon, kardiologi og diabetes retinopati. Modellene ble forklart med eksisterende forklaringsmetoder. Valg og sammenlikning av forklaringsmetoder var basert på type data og problemstilling, hyppig bruk i ML-feltet generelt og resultater oppnådd underveis i doktorgradsarbeidet. Vi utviklet også metoder for økt transparens i selve dataanalysen. Tilbakemeldinger fra medisinske eksperter stod sentralt for å tolke resultatene og vurdere klinisk nytte av ML-systemene.

RESULTATER

Tabell 1 viser resultatene av doktorgradsarbeidet gruppert etter deres bidrag til transparens i de respektive komponentene i et medisinsk ML-system fra figur 1. To medisinske datasett ble publisert (5, 6). Disse kan bidra til økt transparens av fremtidige ML-modeller.

Doktorgradsarbeidet bekreftet at ML-modeller kan analysere medisinske data nøyaktig på tvers av medisinske anvendelser. Innen tørre øyne og kunstig befruktning er det fortsatt rom for bedre modeller. Publisering av høykvalitets-datasett for disse medisinske områdene, slik som (6), bidrar forhåpentligvis til dette.

Et viktig funn er at ikke alle forklaringsmetoder møter behovene til medisinske eksperter, selv om metodene ofte benyttes til å forklare medisinske ML-modeller. Videre etterspurte ekspertene informasjon utover forklaringene av modellprediksjonene, slik som distribusjonen av treningsdataene og modellens usikkerhet (7). Resultatene bekrefter visjonen om et transpa-



Figur 1. *Skjematisk presentasjon av et transparent medisinsk ML-system. Alle komponenter 1 til 6 bør være tilgjengelige, godt beskrevet og tilstrekkelig forklart til brukerne. Hvite bokser gir eksempler på innhold i de ulike komponentene. Blå piler nederst i figuren indikerer at de ulike komponentene påvirker hverandre og utveksler informasjon, for eksempel at tilbakemeldinger fra brukerne kan påvirke valg av hvilke ML-algoritmer og forklaringsmetoder som brukes og at medisinske data benyttes til å evaluere modellen.*

Tabell 1. *Oversikt over resultater fra doktorgradsarbeidet.*

Komponent(er)	Medisinsk domene	Anvendelse
Medisinske data	Kunstig befruktning, gastroenterologi	Tilgjengeliggjøre datasett bestående av henholdsvis annoterte videoer av sperm-prøver og bilder fra mage-tarm med tilhørende spørsmål og svar
Preprosessering av data	Gastroenterologi, diverse tabulære helsedata	Generative ML-modeller for syntetiske bilder fra mage-tarm og syntetiske helsedata i tabellformat
Modelltrening og -evaluering	Tørre øyne, kunstig befruktning, gastroenterologi, hode- og nakkekreft, nyretransplan-tasjon, kardiologi, diabetes retinopati	Predikere alvorlighetsgrad og symptomer på tørre øyne, tolke PET- og CT-bilder og predikere overlevelse hos pasienter med hode- og nakkekreft, predikere systemisk takrolimus-eksponering i nyretransplanterte pasienter, tolke EKG, klassifisere diabetes retinopati. Øke transparens ved segmentering av polypper i bilder fra mage-tarm og ved gruppering av ekstraherte bilder fra videoer av *in vitro* kunstig befruktning.
Forklare modell	Tørre øyne, kunstig befruktning, gastroenterologi, hode- og nakkekreft, nyretransplan-tasjon, kardiologi, diabetes retinopati	Identifisere viktige faktorer relatert til tørre øyne og relevante symptomer, visualisere grupperinger av bilder fra videoer av *in vitro* kunstig befruktning, segmentere polypper i bilder fra mage-tarm, identifisere viktige faktorer for overlevelse av hode- og nakkekreft og for systemisk eksponering av takrolimus i nyretransplanterte pasienter, markere viktige områder i EKG gitt modellens beslutning, kvantifisere relevansen av ulike medisinske bildefunn for klassifisering av diabetes retinopati
Tilbakemelding fra brukere	Tørre øyne, nyretransplan-tasjon, gastroenterologi, kardiologi	Vurdere faktorer identifisert som potensielt viktige for tørre øyne og for systemisk eksponering av takrolimus i nyretransplanterte pasienter, undersøke foretrukne forklaringsmetoder for ML-modeller som detekterer polypper i bilder fra mage-tarm, vurdere klinisk nytte av forklaringsmetoder for ML-modeller som tolker EKG
Litteratursøk*	Tørre øyne, medisin generelt	Danne overblikk over nåværende status for ML og forklaringsmetoder innen medisin

**Ikke en komponent i seg selv, men danner grunnlag for arbeidet med komponentene.*

Forkortelser: CT, computertomografi; EKG, elektrokardiogram; ML, maskinlæring; PET, positronemisjonstomografi

rent medisinsk ML-system som går utover kun å forklare ML-modellen.

I en oppfølgingsstudie ble et bredt utvalg forklaringsmetoder rangert kvalitativt av medisinske eksperter og kvantitativt gjennom en automatisert metode. Forklaringene ble presentert som *heatmaps* (varmekart), som markerer de viktigste områdene i observasjonen som analyseres for modellens prediksjon. Resultatene viser at de medisinske ekspertenes foretrukne forklaringsmetoder avvek fra den kvantitative metoden. Dette understreker viktigheten av helsefaglig involvering for valg av forklaringsmetode. Studien vurderes for publikasjon i et vitenskapelig tidsskrift.

Vi erfarte at forklaringsmetoder kan bidra til ny medisinsk innsikt. I en studie der ML-modeller predikerte alvorlighetsgrad av tørre øyne basert på tårevæskens proteinsammensetning, brukte vi forklaringsmetoder som rangerte proteinene basert på viktighet for modellenes beslutninger. Flere av de viktigste proteinene er allerede ansett som biomarkører for tørre øyne, mens andre ikke er undersøkt med hensyn til tørre øyne ennå (8, 9). Videre studier av disse proteinene kan forhåpentligvis øke kunnskapen om og bidra til mer effektive verktøy for diagnostisering og behandling av denne sammensatte lidelsen.

DISKUSJON

Selv om eksisterende forklaringsmetoder for ML-modeller gir nyttig informasjon til teknologer, gjelder ikke dette nødvendigvis helsepersonell som skal bruke et ML-system. En årsak kan være at metodene er utviklet av teknologer for å forklare ML-modeller generelt. Automatiserte evalueringer av forklaringsmetodene vil ikke alltid samstemme med brukernes vurderinger. Fremtidig forskning bør undersøke hvordan eksisterende forklaringsmetoder kan skreddersys medisinske anvendelser og utvikle automatiserte evalueringsprosedyrer som reflekterer brukernes preferanser av modellforklaringer.

Tverrfaglig samarbeid mellom teknologer og helsepersonell er essensielt for å sikre transparente medisinske ML-systemer som er tilstrekkelig forklart til brukerne. Opplæring av helsepersonell om ML kan forenkle kommunikasjonen mellom systemutviklere og helsepersonell og øke aksepten for implementering av ML-systemer i klinikk.

KONKLUSJON

For at medisinske ML-systemer skal være nyttige, må de underliggende ML-modellene løse oppgavene sine presist og effektivt. Man bør i samarbeid med medisinske eksperter forklare modellprediksjonene på en måte som er nyttig for brukerne av systemet. Data brukt til trening og evaluering av modellen bør være åpent tilgjengelig. Istedenfor å fokusere på én enkelt komponent, skal hele ML-systemet være tilstrekkelig beskrevet og forklart. Helsepersonell som har en god forståelse av medisinske ML-systemer, er bedre rustet til å dra full nytte av dem på en forsvarlig måte som kommer pasientene til gode.

REFERANSER

1. Hu K. ChatGPT sets record for fastest-growing user base – analyst note. Reuters 2. februar 2022. www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (Lest 18. august 2024).
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019; 25: 44–56.
3. Abramson J, Adler J, Dunger J et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024; 630: 493–500.
4. Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019; 8: 2328–31.
5. Hicks SA, Storås A, Halvorsen, P et al. Overview of ImageCLEFmedical 2023 – Medical Visual Question Answering for Gastrointestinal Tract. CLEF 2023 Working Notes. CEUR Workshop Proceedings 2023; 3497: 1316–27.
6. Thambawita V, Hicks SA, Storås AM et al. VISEM-Tracking, a human spermatozoa tracking dataset. Sci Data 2023; 10: 1–8.
7. Storås AM, Andersen OE, Lockhart S et al. Usefulness of Heat Map Explanations for Deep-Learning-Based Electrocardiogram Analysis. Diagnostics 2023; 13: 2345.
8. Storås AM, Magnø M, Fineide FA et al. Identifying important proteins in meibomian gland dysfunction with explainable artificial intelligence. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) 2023: 204–9.
9. Storås AM, Fineide F, Magnø, M et al. Using machine learning model explanations to identify proteins related to severity of meibomian gland dysfunction. Sci Rep 2023; 13: 22946.

■