UiT Norges arktiske universitet

Helsevitenskapelig fakultet

Kunstig intelligens som verktøy i bruddiagnostikk

En litteraturstudie

Henrik B. Dukefoss

Masteroppgave i profesjonsstudiet medisin, Med-3950 august 2020

Forord

Denne masteroppgaven ble valgt på bakgrunn av en sterk interesse for krysningen mellom medisin og teknologi. Teknologi er en stor del av den medisinske hverdagen og er blitt uunngåelig for å levere kvalitet i helsehjelp, spesielt i spesialisthelsetjenesten. Jeg forundrer meg ofte over hvordan helsesektoren ville sett ut uten de enorme fordelene teknologien bærer med seg. Samtidig blir framtidsoptimismen stor når man ser hva kombinasjonen av helsearbeidere og teknologiske hjelpemidler er i stand til å utrette. Blant annet står for meg PET-senteret bygd i Tromsø i løpet min studietid som et symbol på fagets stadige utvikling.

En stor takk til min bror, Erik B. Dukefoss, datateknolog ved NTNU med spesialisering i kunstig intelligens, for hjelp med de teknologiske aspektene ved oppgaven.

Tromsø, 30.08.20

Henrik B. Dukefoss

Innholdsfortegnelse

Sammendrag

Det teoretiske grunnlaget for kunstig intelligens (KI) er ingen ny idé, men etter blant annet en voldsom utvikling i prosessorkraft har fagfeltet sett en ny vår. Det mangler ikke på hverken lovord om hva teknologien er i stand til eller på løfter om hvilke potensielle problemer KI vil løse. Én definisjon av kunstig intelligens mange kan enes om er: «En datamaskin som er i stand til å løse oppgaver uten å få instruksjoner fra et menneske på hvordan den skal gjøre det, har kunstig intelligens». Kjernen i KI kan sies å være evnen til å lære. I motsetning til å skrive programmer til et dataprogram med forklaring på hva det skal gjøre basert på strenge logiske regler, utvikler KI selv algoritmene for problemløsning basert dataene den har til rådighet. For eksempel kan KI bruke en database av klassifiserte røntgen thorax, altså røntgenbilder merket med «fasiten», til å lage en algoritme med diagnostiske ferdigheter. Faget er sterkt inspirert av biologiske hjerner. På samme måte som nevroner i biologiske hjerner er koplet i nettverk, bruker KI digitale nevroner koplet i nettverk. Det er knyttet mye optimisme til anvendelse av KI i medisinske fagfelt. I denne oppgaven studeres testegenskapene til KI i bruddiagnostikk.

Denne litteraturstudien bygger på fem studier som har lagd egne kunstige nevrale nettverk for bruddeteksjon. Testegenskapene har også blitt sammenlignet med menneskelige ferdigheter. Litteraturgjennomgangen viste svært gode testegenskaper for KI-baserte verktøy i deteksjon av brudd. Både sensitiviteten og spesifisiteten til verktøyene var høyere enn ikke-spesialiserte leger og enkelte av nettverket var like gode som ortopedispesialister.

Nøkkelord/ forkortelser

Begrepene som brukes i kunstig intelligens blir fort unyanserte ettersom faget ikke alltid har klare begrepsdefinisjoner. For å unngå nyanseforskjeller i en eventuell egen forklaring i forhold til annen norsk litteratur, har jeg valgt å bruke Per Kristian Bjørkeng sine begrepsforklaringer.

Kunstig intelligens (KI): «En type dataprogrammer med evne til å nå komplekse mål. Disse trekker som regel lærdom fra miljø.»(1)

Artificial intelligens (AI): Engelsk; samme som KI.

Nevrale nett: «Et nettverk av digitale nevroner. Det består av et lag av inputnevroner (for data som skal vurderes) og et lag av outputnevroner (der svarene kommer ut). Mellom input og output finnes ett eller flere såkalte skjulte lag av nevroner. Dersom det er mer enn ett lag med nevroner mellom input og output, kaller vi det et **dypt nevralt nett, dyp læring eller deep learning**. Denne typen algoritme har hatt så stor suksess de siste årene at begrepet nesten er blitt synonymt med AI. Men kunstig intelligens omfatter i vitenskapelig forstand også mange andre typer algoritmer, i tillegg til mer tradisjonelle former for kunstig intelligens som er programmert av mennesker.»(1)

Algoritme: «En serie med instruksjoner i en datamaskin. Algoritmen tar imot inndata, bearbeider dem, og gir fra seg et resultat i form av utdata. En matoppskrift er et praktisk eksempel på algoritme. Algoritmene som brukes i AI, tar typisk imot diffuse inndata, bruker forskjellige matematiske metoder til å få mer orden på dem, og gir oss resultatene i form av mer ordnede utdata. En slik algoritme kan trenes til å gjenkjenne mønstre i data den aldri har sett før. Et mønster kan gjenkjennes selv om det ikke er identisk med det algoritmen har trent på».(1)

Maskinlæring: «En viktig gren av den vitenskapelige disiplinen kunstig intelligens. Befatter seg med utvikling av algoritmer som gjør datamaskiner i stand til å lære fra empiriske data eller interaksjon med miljøet, og utvikle atferd basert på slike data. Maskinlæring er den grenen av kunstig intelligens som har gjort størst fremskritt de siste årene, og brukes nå ofte synonymt med AI eller kunstig intelligens. Denne boken handler om kunstig intelligens, men det er i all hovedsak underkategorien maskinlæring den fokuserer på.»(1)

Digitalt nevron/perseptron: «En bitte liten algoritme bygget opp på en måte som minner om nevronene i hjernen vår. Den kan ta imot signaler fra andre nevroner gjennom digitale forbindelser kalt vekter. Inputvektene kombineres, og nevronsignalet sendes videre til andre nevroner.»(1)

Svak/smal og sterk/bred AI: S»vak eller smal AI betegner kunstig intelligens som er høyt spesialisert. Et typisk eksempel er et dypt nevralt nett som kan kategorisere bilder. Dette er dagens mest utbredte AI. Sterk eller bred AI tilhører fremtiden. Sterk AI trenger ikke bare være trent på én oppgave, men kan løse mange forskjellige komplekse oppgaver. Når én sterk AI-modell en gang i fremtiden kanskje kan løse alle oppgaver like godt som mennesker, sier vi at vi har utviklet Artificial General Intelligence (AGI) eller generell kunstig intelligens.»(1)

1.0 Innledning

Kunstig intelligens blir sett på som en nymotens oppfinnelse og er forbundet med en del mystikk. I realitet ble det matematiske grunnlaget for kunstig intelligens og maskinlæring utviklet på 50- og 60-tallet.(2) Selve begrepet kunstig intelligens hinter til at det er en komponent av intelligens involvert, noe som medfører at intelligens må defineres. Det finnes imidlertid ingen universell enighet om definisjonen av intelligens. Uansett kan det argumenteres for at kunstig intelligens ikke i seg selv er intelligent i det hele tatt, men snarere en statistisk metode nært beslektet metodene som brukes for å besvare vitenskapelige spørsmål allerede.

I innledningen skal jeg først beskrive det teoretiske grunnlaget for kunstig intelligens. Deretter presenteres problemstillingen og formålet med oppgaven, samt begrunnelse av valg av problemstilling og hvordan oppgaven avgrenses.

1.1 Historie

Historien til kunstig intelligens (KI) kan argumenteres for at begynner med den engelske matematikeren Alan Turing, kjent for å knekke tyskernes krypteringsmaskin, Enigma, under 2. verdenskrig. Turing teoretiserte en maskin med evne til å lære.(3) I 1956 arrangerte matematikkprofessor John McCarthy en sommerkonferanse ved Dartmouth College hvor ti ledende forskere samlet seg for å diskutere og undersøke rekkevidden og kompleksiteten av oppgaver en datamaskin kunne utføre. I søknaden for finansiering av konferansen lagde McCarthy begrepet kunsting intelligens.(3) Til tross for at konferansen ledet til, på den tiden, forbausende resultater, begrenset omfanget av programmene seg til logisk oppbygging av algoritmer: «hvis x er tilfelle, så skal du gjøre y». Den rådende ideen var at tilstrekkelig kompliserte, menneskeskapte algoritmer til slutt ville forutsi alle utfall av en gitt situasjon og respondere på riktig vis.(3) Et eksempel på det ville være å prøve og definere alle karakteristikaene til f.eks. et røntgen thorax og ta i betraktning alle anatomiske varianter av en brystkasse og alle måter sykdom kan presentere seg. Til tross for at denne metoden har vært essensiell innen informatikk-revolusjonen vi har sett de siste tiårene, har den ikke vært særlig

nyttig i for eksempel oppgaver som bildegjenkjenning, mønstergjenkjenning, språkforståelse og oppgaver som krever «intuisjon». På sommerkonferansen på Darthmouth ble også ideen om et nevralt nettverk med evne til å lære fra erfaring presentert. Likevel tok det flere år før idéen skulle bli realisert. Datidens prosessorkraft var også en klar begrensning for realisering av et operativt nevralt nettverk.

1.2 Teoretisk grunnlag for kunstig intelligens

Helt sentralt for KI er evnen til å lære. Under følger en definisjon av maskinlæring som er mye brukt:
"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."(4)

En enklere definisjon som mange kan enes om er: «...en datamaskin som er i stand til å løse oppgaver uten å få instruksjoner fra et menneske på hvordan den skal gjøre det, har kunstig intelligens».(5)

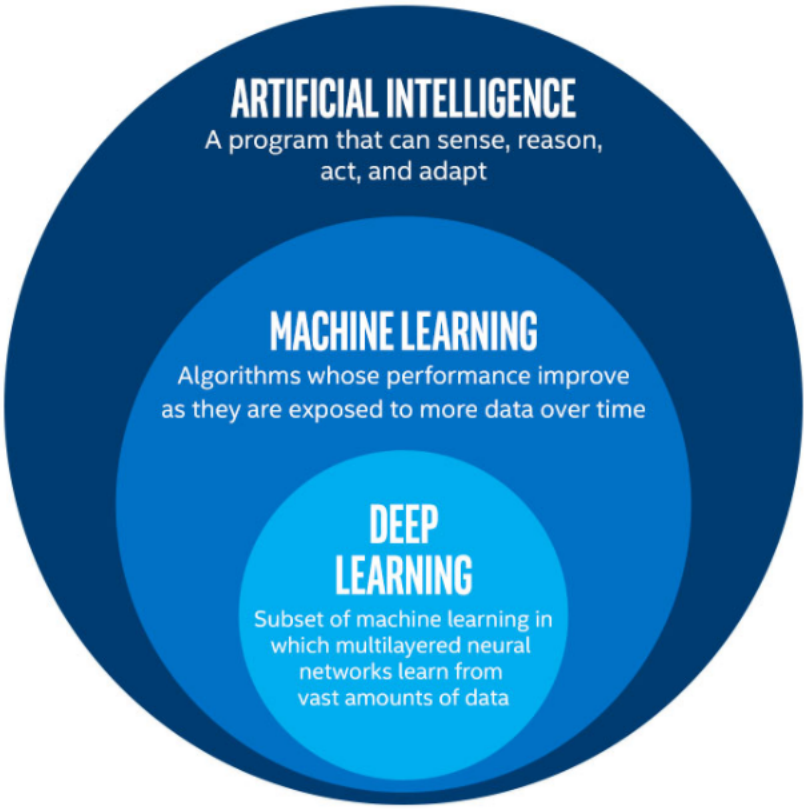
«Som fagfelt er kunstig intelligens en sammensmelting av datateknikk, logikk, matematikk, psykologi og nevrovitenskap.»(5)

Årsaken til at KI er avhengig av datateknikk, logikk og matematikk er naturlige følger av hvordan programmer og datamaskiner fungerer. At psykologi og nevrovitenskap spiller en rolle i fagfeltet er mindre åpenbart. På 60 -og 70-tallet ble «fagfeltet» videreført av nettopp mennesker med kompetanse i psykologi og nevrovitenskap. Årsaken til det er at KI er sterkt inspirert av biologiske hjerner.(5) På samme måte som biologiske hjerner består av nevroner koplet i nettverk, består KI av digitale nevroner koplet i nettverk. Essensen i KI er altså en maskin som evner å lære og er basert på etterligning av biologiske hjerners fysiologi.

De seneste tiårenes computerutvikling har vært muliggjort av menneskelagde algoritmer og koder. Det vil si at mennesker har programmert computere til å gjøre oppgavene de skal løse. Kunstig intelligens avviker fra dette, da programmereren kun setter opp rammeverket og heller lar dataene bestemme hvordan programmet skal løse problemet. I tillegg har kunstig

intelligens evnen til å justere sine egne algoritmer. Det gjør at teknologien tilsynelatende framstår intelligent.(5) Kunstig intelligens og maskinlæring er sterkt inspirert av biologiske hjerner. På samme måte som biologiske hjerner er et nettverk av nevroner, baserer KI seg på digitale nevroner knyttet sammen i et nettverk. Mens biologiske hjerner bruker sanser til å innhente data, mater man rådata inn i nettverket av digitale nevroner slik at programvaren kan lære av dataene.

1.3 Subkategorier av KI



(6)

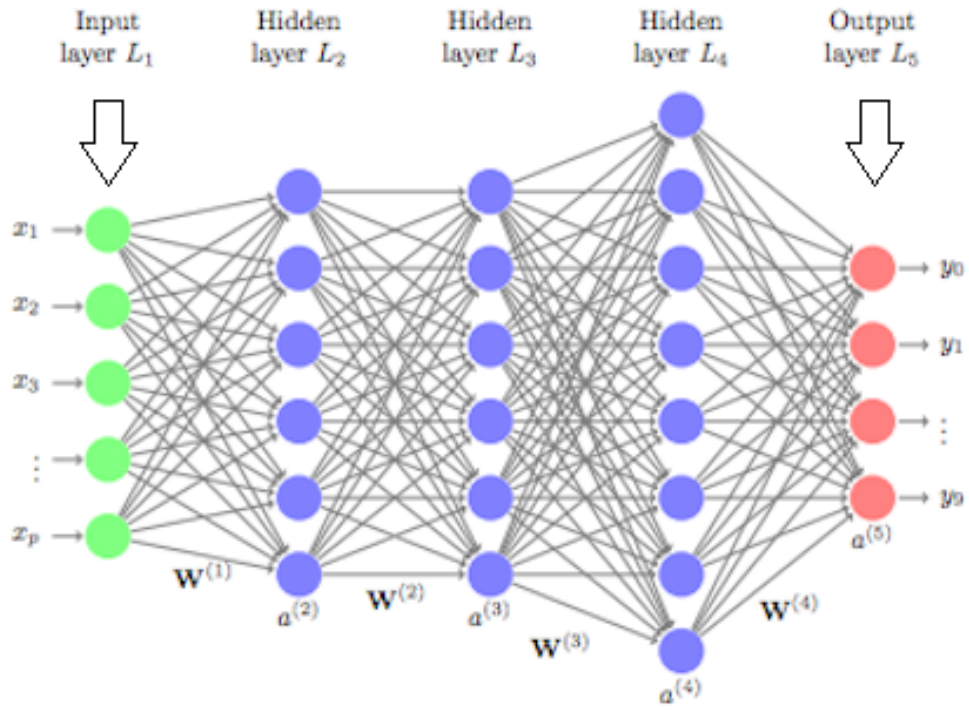
Som illustrasjonen over viser er KI en bred konseptbetegnelse. En viktig subkategori av KI er maskinlæring.(7) Maskinlæring innebærer systemer som kan lære. Maskinlæring deles ofte inn i tre hovedkategorier: Veiledet læring, ikke-veiledet læring og forsterkende læring. Før en forsterkende læringsalgoritme blir eksponert for data kan det ingenting. En typisk analogi vil være et nyfødt barn uten noen kjennskap til verden. På samme måte som barnet lærer av sanseinntrykk og erfaring, må programmet eksponeres for data før det kan brukes til noe som helst. På samme måte som et barn prøver og feiler gjør den forsterkende læringsalgoritmen

det samme. Den prøver ut ulike fremgangsmåter og beholder den fremgangsmåten som gjør det best ut ifra en predefinert målsetning.(8)

Veiledet læring kan være et spesielt nyttig verktøy i helsesektoren. Dataen den veiledende læringsalgoritmen kan eksponeres for kan være nesten hva som helst, så lenge den er klassifisert. La oss si at programmet blir matet med røntgenbilder, og bildene er klassifisert ut ifra om pasientene har pneumoni eller ikke.

Databasen av røntgen thorax bilder som brukes er merket med «ikke-pneumoni» eller «pneumoni». Deretter utsettes programmet for trening. I starten må programmet gjette uten informasjonsgrunnlag. Hvis gjetningen viser seg å være feil, justeres programmets indre modell før det forsøkes igjen på et nytt bilde. Systemet eller programmet kalles for nevrale nettverk. Den minste funksjonelle «komponenten» i nettverket kalles et perseptron.(7) Et perseptron kan sies å være et digitalt nevron og er en liten algoritme som kan motta og sende tallverdier. Verdiene som mottas kommer enten fra andre perseptroner eller fra omgivelsene. Omgivelsene kan for eksempel være pikslene fra et røntgen thorax. For å eksemplifisere kan vi fortsette å bruke det hypotetiske datasettet med røntgenbilder av pasienter med enten «pneumoni» eller «ikke-pneumoni». Pikslene som representerer penetransen av røntgenstråling i lungevevet vil være viktig faktor i for å vurdere «pneumoni»/«ikke-pneumoni». Første steg i algoritmen er å finne egenskaper i bildet gjennom konvulsjon hvor et filter anvendes på bildet og rekalkuleres med matriseoperasjoner for å finne et 'egenskapskart'.(9) Deretter reduseres dimensjonene i bildet for å ta høyde for romlige variasjoner som gjør at blant annet bilderotasjon ikke har betydning. Til slutt kobles det inn et nevralt nettverk som både gjør at prediksjonene blir bedre jo flere bilder den ser, samt foretar selve prediksjonene. Perseptronet som mottar verdiene enten fra bildet eller fra andre perseptroner foretar kalkulasjoner og sender informasjonen til andre perseptroner. Verdiene som sendes fra perseptron til perseptron er koplet i et lag og nettverk. Til slutt vil informasjonen akkumuleres i et siste perseptron som vil gi sitt svar «pneumoni» eller «ikke pneumoni», eventuelt tillagt sannsynligheten for resultatet. Dette resultatet sammenliknes så med klassifiseringen av bildet gjort av en lege. Programmet har da enten beregnet riktig eller feil. I begge tilfellene vil det kjøres en matematisk optimaliseringsalgoritme som justerer perseptronets vektleggelse. Det er i denne fasen programmet kan justeres til å finne ut hvilke faktorer som er avgjørende for å diagnostisere pasienten.

1.4 Nevrale nettverk



(10)

Som illustrasjonen over viser består nettverket av flere lag. Alle perseptronene i et lag er koplet til alle perseptronene i neste lag. Inngangslagets (Input layer) registrer data som skal analyseres og tolkes av lagene som kommer etter. Hvert perseptron i lagene «argumenter» for et gitt resultat. Argumentet blir så sendt videre til neste preseptron. Et perseptron mottar argumenter for eller imot et gitt resultat fra alle perseptronene i det forrige laget. Produktet av denne argumentasjonen avgjør hvilken informasjon perseptronet sender videre. På mange måter er metodikken lik den mennesker bruker i sin beslutningstagning.(11) Eksempelvis vil en kliniker vurdere sannsynligheten for pneumoni basert på en rekke faktorer og tillegge hver faktor en grad av betydning. At pasienten har hodepine tillegges for eksempel lavere grad av betydning enn symptomene hoste og dyspné for å avgjøre sannsynligheten for pneumoni. Med pneumoni som eksempel kan nettverket for eksempel få input om:

- Symptominformasjon som feber, hoste, dyspné, allmenntilstand, brystsmerter.
- Undersøkelsesinformasjon som mental forvirring, puls, respirasjonsfrekvens, blodtrykk og temperatur
- Laboratoriske tester som CRP og patogener.
- Komorbiditeter som lungesykdommer, nedsatt immunforsvar osv.
- Radiologiske undersøkelser som røntgen og CT.

Når nettverket trenes på data fra tusenvis av pasienter med tilhørende fasiten på om pasientene hadde eller ikke hadde pneumoni, vektes betydningen av faktorene. Fravær av radiologiske tegn på pneumoni vil man anta at nettverket vekter som et sterkt motargument imot at pasienten har pneumoni. Derimot kan være at produktet av all den andre informasjonen sammen veier tungt nok til at programmet diagnostiserer pasienten med pneumoni. Ettersom programmet selv vekter betydningen av faktorene basert på dataen det er trent på, vil programmet selv avgjøre hva som har betydning for sannsynligheten for pneumoni i et gitt tilfelle.

For å trene et nevralt nettverk til tilfredsstillende nivå er man avhengig av mye data. Ofte er det tilgangen på data som er hinderet til å lage gode nok programmer. Typisk vil 80% av den tilgjengelige dataen i et datasett bli brukt til å trene systemet, altså man opplyser programmet om resultatet var riktig eller galt. Mens 20% vil bli brukt til å teste det.(12) Da bruker man ikke dataen til å gjøre endringer i nettverket for å forbedre prediksjonene, men heller finner hvor mange prosent av dataen programmet klassifiserte riktig. Kvaliteten på datasettet er også fundamentalt for å oppnå et godt resultat. Et program vil ikke kunne bli bedre enn datasettet.

1.5 Medisinsk anvendelse av KI

KI har en bred anvendelse i medisinske fag. KI er spesielt egnet til oppgaver som krever mønstergjenkjenning. Fag som radiologi og patologi er derfor spesielt egnet til anvendelse av kunstig intelligens. Selv om KI fremdeles kan sies å være i sin barndom, har det allerede begynt å prestere på nivå som kan tenkes forenlig med klinisk utnyttelse. Videre følger eksempler for medisinsk anvendelse av KI. Studien «The potential for artificial intelligence in healthcare» har oppsummert mange potensielle bruksområder:(8)

- Tidlig bevisstgjøring av høyrisiko-sykdommer som sepsis og hjertesvikt.
- Diagnostisering og behandling av kreft basert på genetisk profil.
- Språkprosessering for strukturering av journaler.
- Kirurgiske roboter.
- Forutse pasienter med risiko for spesifikke sykdommer.
- Forutse sannsynlighet for reinnleggelser.
- Individuelt spesialtilpassede behandlingsplaner.
- Diagnostisering med radiologiske verktøy.
- Diagnostisering av patologiske preparater.

- Administrativ effektivisering.

Listen av potensielle bruksområdet til KI i helsevesenet er mye lengere. Flere AI-baserte verktøy har allerede funnet veien inn i klinisk praksis. Som verktøy i deteksjon av tidlig pneumoni-deteksjon og potensiell triagering av Covid-19 pasienter har forskere ved UC San Diego health brukt AI i klinisk forskning.(13) Forskerne brukte 22 000 bilder til å trene et nevralt nettverk i pneumoni diagnostisering. Programmet er ikke spesifikt for Covid-19, men tenkes å bli brukt til tidlig diagnostisering av Covid-19, spesielt for å avgjøre om forløpet blir behandlingskrevende eller ikke. «In one case, a patient in the Emergency Department who did not have any symptoms of COVID-19 underwent a chest X-ray for other reasons. Yet the AI readout of the X-ray indicated signs of early pneumonia, which was later confirmed by a radiologist. As a result, the patient was tested for COVID-19 and found to be positive for the illness.» KI har også blitt brukt som supplement til radiologer for å øke testegenskapene ved mammografier.(14) Selskapet cmAssist bruker KI i computer aided detection (CAD). Sensitiviteten for deteksjon av malignitet økte med i gjennomsnitt 27% ved bruk av teknologien. Andelen falske positive økte med under 1 %.

Til tross for høye forventninger og stadig nye eksempler på KI som presterer på høyt nivå, er det flere barrierer for klinisk bruk av teknologien. Pasienters villighet til å la algoritmer diagnostisere og lage behandlingsplaner er en av barrierene. Noen peker på at pasienter mener sine helsebehov er for unike til at algoritmer tilstrekkelig klarer å løse dem.(15)

1.6 Problemstilling

Problemstillingen er: Hva er sensitiviteten og spesifisiteten til verktøy basert på kunstig intelligens i bruddiagnostikk? Hvor gode er testegenskapene sammenlignet med leger, radiologer og ortopeder?

Bakgrunnen for problemstillingen bunner ut i en nysgjerrighet om hvor god teknologien har blitt. Ettersom fagfeltet kunstig intelligens har gjennomgått en omveltning bare det siste tiåret, har jeg forsøkt å unngå tematikk som innebærer potensielle resultater og bruksområder. I stedet avgrenses problemstillingen til verktøy og testegenskaper som er blitt testet og studert. For å avgrense oppgaven ytterligere har jeg valgt å fokusere på bruddiagnostikk. Det blir fort fristende å diskutere andre sider ved KI, som for eksempel de etiske implikasjonene

implementasjon medfører. I et forsøk på å avgrense oppgaven har jeg bevisst valgt å unngå temaer som ikke er relevant nok for testegenskapene til de KI-baserte verktøyene.

Det er også fristende og hoppe fra et medisinsk fagfelt til et annet, og undersøke testegenskapene til verktøyene på tvers av disipliner. For å unngå det og for å konkretisere oppgaven, valgte jeg derfor radiologi som undersøkelse, med ortopedi som fagfelt og frakturer som diagnose. Det har imidlertid vært vanskelig å avgrense oppgaven videre til en kroppsdel.

1.7 Formål

Formålet med denne oppgaven er å undersøke sensitiviteten og spesifisiteten til verktøy basert på kunstig intelligens i bruddiagnostikk, samt å sammenligne testegenskapene med menneskelige ferdigheter.

1.8 Språkbruk og synonymer

Ordforrådet som brukes til å beskrive KI er både komplekst tidvis uoversiktlig. Samtidig lages det stadig nye begreper. Mange av begrepene overlapper med biologiske slektninger av konseptene. I tillegg florerer det av tilsynelatende synonymer, med formål å beskrive nyanseforskjeller i tilnærmet samme konsept. I denne oppgaven har det vært vanskelig å være konsistent i språkbruken. «Nevrale nettverk», «KI-baserte verktøy», «programmer» og «algoritmer» er blant ordene som brukes i løpet av oppgaven til å beskrive det endelige verktøyet som basert på KI brukes til å løse medisinske oppgaver. Jeg beklager hyppig synonymbruk for å beskrive samme konsept.

2.0 Materiale og metode

Jeg har gjort følgende søk i PubMed: «Artificial intelligence orthopedics» og «Artificial intelligence fractures». «Artificial intelligence orthopedics» ga 387 treff, mens «Artificial intelligence fractures" ga 86 treff. En del av treffene overlappet for søkeordene jeg valgte. For å finne relevante artikler har jeg forsøkt å lete etter studier der forfatterne lagde egne nevrale nettverk, samt oppga sensitivitet og spesifisitet for programmene. I tillegg forsøkte jeg å finne studier som sammenlignet testegenskapene med legers testegenskaper. For å finne relevante artikler begrenset jeg søket til artikler publisert fra og med 2017. Grunnet den hyppige utviklingen i fagfeltet var det min vurdering at artikler og KI-baserte verktøy eldre enn det ikke ville være teknologisk tilstrekkelige lenger. Jeg har bare benyttet engelskspråklige studier. Til slutt valgte jeg ut studier jeg anså relevante nok til å bli inkludert. Ettersom oppgavens problemstilling og formål er smal, valgte jeg ikke å bruke mer enn fem studier som grunnlag for denne oppgaven. Dette ble også begrenset av mengden forskning gjort på feltet. Kvaliteten og innholdet på de fem studiene vurderte jeg også til å være gode og omfattende nok til løse problemstillingen til denne oppgaven.

Som verktøy for referanser har jeg brukt EndNote X9.

3.0 Resultater:

Studie 1:

I studien "Artificial intelligence for analyzing orthopedic trauma radiographs - Deep learning algorithms—are they on par with humans for diagnosing fractures?" har forfatterene brukt 256 458 røntgenbilder av hender, håndledd og ankler for å undersøke om et dypt nevralt nettverk er stand til å identifisere frakturer i ortopediske røntgenbilder.(16) Senere sammenligner studien programmets diagnostiske ferdigheter med to ortopediske overleger. I frakturdiagnostikken ga resultatet en nøyaktighet på 83 % det nevrale nettverket. De to ortopediske overlegene hadde en nøyaktighet på 82%.

Studie 2:

I studien "Deep neural network improves fracture detection by clinicians" utviklet forfatterene et dypt nevralt nettverk for å lokalisere brudd i røntgenbildet.(17) Datagrunnlaget for nettverket var 135 409 røntgenbilder vurdert av 18 ortopediske kirurger. Formålet med studiet var å sammenligne akuttmedisinske legers evne til å detektere frakturer røntgenbilder av håndledd alene og med assistanse fra det nevrale nettverket. Den gjennomsnittlige legens sensitivitet var 80.8% (95% CI, 76.7–84.1%) uten assistanse og 91.5% (95% CI, 89.3–92.9%) med assistanse. Legens gjennomsnittlige spesifisitet var 87.5% (95 CI, 85.3–89.5%) uten assistanse og 93.9% (95% CI, 92.9–94.9%) med assistanse. Den gjennomsnittlige legen erfarte en relativ reduksjon av feiltolking på 47% (95% CI, 37.4–53.9%).

Studie 3:

I studien "Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs" brukte forfatterne 25 505 røntgenbilder av bekken til å lage et deep convulsion neural network for diagnostisering av frakturer i bekken.(18) Formålet med studien var å undersøke nøyaktigheten, sensitivetet, raten av falske negative og AUC. For identifisering av hoftefrakturer oppnådde algoritmen en nøyaktighet på 91%, en sensitivitet på 98%, en falsk negativ rate på 2% og en AUC på 98%.

Studie 4:

I studien "Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments" lagde et CNNs for deteksjon av distale radius frakturer og sammenlignet dets egenskaper med radiologer og ortopeder.(19) Datasettet ble konstruert ved at to ortopediske overleger gjennomgikk 2 359 røntgenbilder av håndledd. Det nevrale nettverket hadde en nøyaktighet på 93%, en sensitivitet på 90% og en spesifisitet på 96%. De ortopediske legene hadde en nøyaktighet på 94%, en sensitivitet på 93% og en spesifisitet på 95%. Radiologene hadde en nøyaktighet på 84%, en sensitivitet på 81% og en spesifisitet på 87%.

Studie 5:

I studien "Automated detection and classification of the proximal humerus fracture by using deep learning algorithm" evaluerer forfatterne egenskapene til en «dyp læring» algoritme til å detektere og klassifisere proksimale humerus frakturer.(20) Datasettet ble bestod av 1 891 røntgenbilder av proximale humerus vurdert av speasialister. Egenskapene til programmet ble senere sammenlignet med 28 allmennleger, 11 ortopeder og 19 ortopeder med spesialisering i skulder. Det nevrale nettverket hadde en sensitivetet på 99%, en spesifisitet på 97%, og en Youden index på 97%. Allmennlegene hadde en sensitivitet på 82%, spesifisitet på 94% og Youden index på 77%. Ortopedene uten spesialisering i skulder hadde en sensitivitet på 93%, spesifisitet på 97% og Youden index på 77%. Ortopedene med spesialisering i skulder hadde en sensitivitet på 96%, en spesifisetet på 98% og en Youden index på 94%.

4.0 Diskusjon

Studiene som er valg har alle til felles at de ser på bruddiagnostikk med radiologiske verktøy. Alle bruker også røntgen som modalitet. Enkelte av studiene sammenligner resultatene fra KI-verktøyene med leger, radiologer og ortopeder. Det gir et bedre bilde av hvor gode testegenskapene faktisk er for den gitte undersøkelsen. Interessant nok har forskerne valgt å selv utvikle verktøyene for bruddeteksjon og ikke basert seg på for eksempel teknologiske bedrifters teknologi til det. Ettersom forskerne har valgt å utvikle verktøyene selv kan man stille spørsmål til om programmene ble så gode som de kunne blitt. I seg selv er det ikke nødvendigvis fordelaktig å ha kunnskaper om radiologi for å utvikle nettverket. Likevel er det gjennomgående for alle studiene at programmene som har blitt utviklet har hatt gode testegenskaper. Som nevnt i innledningen blir ikke de nevrale nettverkene bedre enn datagrunnlaget de får servert. Ulempen med det er at røntgenbildene for datagrunnlaget i første omgang ble vurdert av leger. Sensitivetet og spesifisiteten til legene som i vurderte røntgenbildene i datagrunnlaget vil derfor bli reflektert i programmets endelige algoritme. Man kan derfor stille seg kritisk til hvor godt datagrunnlaget har vært.

I «studie 1»; "Artificial intelligence for analyzing orthopedic trauma radiographs - Deep learning algorithms—are they on par with humans for diagnosing fractures?" har forfatterene brukt 256 458 røntgenbilder av hender, håndledd og ankler for å undersøke om et dypt nevralt nettverk er stand til å identifisere frakturer i ortopediske røntgenbilder. Datagrunnlaget for utviklingen av algoritmene er enormt. Som nevnt i innledningen er de nevrale nettverkene avhengige av store mengder data. Jo større datasettet er, desto bedre blir som regel resultatet. Studien peker på at nøyaktigheten til algoritmen på 83%, 1% bedre enn de ortopediske overlegene. Studien oppgir dessverre ikke sensitiviteten og spesifisiteten til hverken programmet eller ortopedene. Likevel gir den en pekepinn på hvor gode programmet har blitt for bruddiagnostikk. Hvis man tar det for gitt at programmet fungerer like godt som en ortopedisk overlege, er det flere bruksapplikasjoner i den kliniske hverdagen. For eksempel kan nyutdannede leger få støttet oppunder sine mistanker dersom legen er i tvil, eller få «bekreftet» en bruddiagnose dersom han/hun har en sterk mistanke. Spesielt i tilfeller der det mangler spesialister, eller der indikasjonen for å tilkalle en bakvakt ikke er stor nok, kan verktøyet tenkes å bli nyttig. Leger fra andre spesialiseringer som likevel er nødt til å vurdere røntgenbilder kan også få støtte fra et slikt program.

Idéen om å bruke programmet til å understøtte legers evne til frakturdeteksjon blir drøftet i «studie 2», «Deep neural network improves fracture detection by clinicians». Forfatterne har som formål i denne studien å undersøke til hvilken grad et dypt nevralt nettverk kan øke deteksjonsevnen, og ikke bare hvor godt det nevrale nettverket i seg selv fungerer. I tillegg til at sensitiviteten på bruddiagnostikk for akuttmedisinske leger gikk opp, er et annet interessant poeng at spesifisiteten gjorde det samme. Programmet ledet derfor udiskutabelt til bedre testegenskaper for legene. Studien viser til at de 135 409 røntgenbildene som dannet datagrunnlaget var vurdert av 18 ortopediske kirurger. Dette, samt det nevrale nettverkets prestasjon, kan argumentere for at det nevrale nettverket ble en refleksjon av de ortopediske kirurgenes evner. Idéen om å bruke et AI-basert verktøy til å støtte mindre erfarende leger styrkes av denne studien. Med tanke på at programmet ga legene en relativ reduksjon av feiltolkning på 47%, vil det være nærliggende å tro at det i framtiden vil bli uforsvarlig å ikke støttes av et slikt program.

Om programmet vil være en erstatning til å bringe inn spesialkompetanse i enkelte kliniske sammenhenger er derimot vanskeligere å vurdere. Juridiske aspekter og ansvarligtildeling er et stort tema i denne sammenhengen. Dersom spesialistkompetanse som følge av ansvarsfordeling må involveres uansett, kan man diskutere poenget med å støttes av et slikt program i første omgang. Juridiske aspekter og implementasjon ligger utenfor avgrensningen av denne oppgaven, men er likevel verdt å nevne.

Et annet poeng er at det nevrale nettverket lagt i denne studien er smalt. Det vil si at programmet kan løse akkurat de problemene det er lagd for å løse, men ingenting annet. Programmets evne til differensialdiagnostikk er derfor fraværende. Dersom problemstillingen strekker seg utover ren bruddiagnostikk, vil programmet kun være til marginal hjelp. Denne mangelen på generell kunstig intelligens er av avgjørende betydning for hvorfor å legge for stor vekt på programmenes vurdering er risikabelt. Det er også grunnen til at det å erstatte en lege i slik diagnostikk ikke er forsvarlig, til tross for gode, eller til og med bedre testegenskaper i programmene.

«Studie 4», “Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments” er veldig lik “studie 2”. Begge studiene undersøker testegenskapene til KI-baserte verktøy i diagnostikk av håndledds/distale radius-frakturer. Begge studiene baserer seg også på selvlagde nevrale

nettverk for bruddeteksjon. Til felles er også at røntgenbildene datasettene ble vurdert av ortopediske overleger. Studiene skiller seg ved at «studie 4» har et langt mindre datasett; 2 359 mot 135 409. 2 359 røntgenbilder er i minste laget for å gi det nevrale nettverket nok informasjon å bygge på. Det kan derfor tenkes at programmet med mer data kunne forbedret testresultatene sine. Til tross for et relativt lite datagrunnlag hadde nettverket bedre testegenskaper enn radiologene. At nettverket ikke presterte bedre enn de ortopediske legene må ses i lys av at datagrunnlaget ble utarbeidet av ortopediske leger. Et nevralt nettverk kan som nevnt i innledningen ikke prestere bedre enn det datagrunnlaget det bygger på.

Behovet for store, klassifiserte databaser er ofte til hinder for utviklingen av KI-baserte verktøy. Det kan skyldes den ressurskrevende oppgaven å samle inn tilstrekkelig med data, men også forhold som personvernsopplysninger og pasientsamtykke til at dataene deres blir brukt i forskning og potensielt i kommersielle bestrebelser. At forfatterne bak studien likevel har lykkes i å lage et nevralt nettverk som utkonkurrerer radiologer i bruddiagnostikk er imponerende. Etter hvert som datakravet til nevrale nettverk senkes som følge av bedre rammeverk, blir terskelen for å utvikle nye KI-baserte verktøy lavere. Spesielt sjeldne diagnoser, som medfører vanskeligere innsamling av data, kan da potensielt bli realistiske å diagnostisere med KI-baserte verktøy. Bredden av diagnoser i en database vil også bli reflektert i programmets evne til differensialdiagnostikk. For differensialdiagnostikk er det helt nødvendig at det foreligger tilstrekkelig med data for hver diagnose. Alle studiene som er brukt i denne oppgaven har hatt en smal tilnærming, programmert for én diagnose; fraktur. «Studie 1» har vært et delvis unntak, da det nevrale nettverket var i stand til å gjenkjenne ulike kroppsdeler; hender, håndledd og ankler, samt å diagnostisere frakturer i hver av dem.

«Studie 3», “Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs”, og «studie 5», “Automated detection and classification of the proximal humerus fracture by using deep learning algorithm”, har begge brukt en dyp lærings algoritme til å detektere henholdsvis hoftefrakturer og frakturer i proksimale humerus. De har til felles en særlig høy sensitivitet på 98- og 99%.

«Studie 5» hadde også en høyere spesifisitet enn allmennlegene og de ikke-spesialiserte ortopedene. Det nevrale nettverket hadde en spesifisitet på 97%, mens ortopedi-spesialistene hadde en spesifisitet på 98%. Skal man tro disse studiene er testegenskapene oppsiktsvekkende gode. Med et konfidensintervall mellom 99-100% med p-verdi 0,001 for

sensitivitet for det nevrale nettverket, er det grunn til å ha tillit til resultatene. Likevel må det nevnes at sammenligningen med de ikke-spesialiserte ortopedene og de spesialiserte ortopedene var statistisk signifikante.

Også i sammenligningen mellom «studie 3» og «studie 5» er det en stor differanse mellom størrelsen på datasettene brukt til å lage de nevrale nettverkene. I «studie 3» brukte forfatterne 25 505 røntgenbilder, mens det i «studie 5» ble brukt 1 891 bilder. I «studie 5» finner man igjen at det nevrale nettverket ikke når opp til testegenskapene til ortopedene med spesialisering i skulder. Ikke i noen av studiene som har sammenlignet de nevrale nettverkene med spesialister har nettverkene hatt bedre testegenskaper enn spesialistene. Som nevnt har det noe å gjøre med at nevrale nettverk ikke blir bedre enn kvaliteten på datasettene. Som et resultat av dette kan det vurderes at nevrale nettverk fremdeles har til gode å bli like dyktige som spesialister i frakturdiagnostikk.

5.0 Konklusjon

I dette litteraturstudiet har jeg undersøkt hvor gode testegenskapene til KI-baserte verktøy er på å detektere frakturer på røntgenbilder. Så langt det har latt seg gjøre er de nevrale nettverkene sammenlignet med leger fra forskjellige fagfelt. Innledningen inneholder en overordnet innføring i hva KI er og hvordan det kan anvendes i medisinske fagdisipliner. Studiene som ble brukt i oppgaven ble selektert med tanke på årstallet for publikasjon, ettersom fagfeltet er i rask utvikling. For å dekke formålet med oppgaven var det også nødvendig at studiene brukte KI-baserte nettverk til bruddiagnostikk, helst sammenlignet med leger.

Problemstilling for oppgaven var: «Hvor god er sensitiviteten og spesifisiteten blitt for verktøy basert på kunstig intelligens i bruddiagnostikk? Hvor gode er testegenskapene sammenlignet med leger, radiologer og ortopeder?» Resultatene viste svært gode testegenskaper for KI-baserte verktøy i bruddiagnostikk. Studiet av det nevrale nettverket som hadde de beste testegenskapene pekte på en sensitivitet på 99% og en spesifisitet på 97% for bruddeteksjon i proksimale humerus. Testegenskapene var dermed vel så gode som ortopedene med spesialisering i skulder. En annen studie viste at bruddeteksjonen på røntgenbilder av håndledd hadde en relativ reduksjon av feiltolking på 47% når legen ble støttet av et nevralt nettverk beslutningstakingen. I diskusjonsdelen ble KI-baserte verktøy drøftet og vurdert som et potensielt supplement til leger i diagnostikken.

Som konklusjon er dagens testegenskaper til verktøy basert på kunstig intelligens svært gode. Sensitiviteten og spesifisiteten er bedre enn leger uten spesialisering i ortopedi, og enkelte nevrale nettverk er like gode som spesialister i ortopedi. Samtidig tas det forbehold i de KI-baserte verktøyenes begrensninger, spesielt med tanke på differensialdiagnostikk.

1. Bjørkeng PK. Kunstig intelligens Den usynlige revolusjonen. Oslo: Vega Forlag; 2018. 24-30

p.
2. Bjørkeng PK. Kunstig intelligens Den usynlige revolusjonen. Oslo: Vega Forlag; 2018. 19 p.
3. Bjørkeng PK. Kunstig intelligens Den usynlige revolusjonen. Oslo: Vega Forlag; 2018. 565 p.
4. Mitchell T. Machine learning: McGraw-Hill Science/Engineering/Math; 1997. 2 p.
5. Tidemann A. Kunstig intelligens: Store Norske Leksikon; 2020 [Available from: https://snl.no/kunstig_intelligens.
6. Bansal H. How to get the Perfect start in AI & ML as Newbie?? Learn the Art in just 5 mins! 2019 [Available from: https://becominghuman.ai/how-to-get-the-perfect-start-in-ai-ml-as-newbie-learn-the-art-in-just-5-mins-cba28d2705e4.
7. Dvergsdal H. nevralt nettverk: Store Norske Leksikon; 2019 [cited 2020 04.06.]. Available from: https://snl.no/nevralt_nettverk.
8. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94-8.
9. Convolutional neural network: Wikipedia.com; [cited 2020 10.07.]. Available from: https://en.wikipedia.org/wiki/Convolutional_neural_network.
10. Ukjent. Deep Learning: How Will It Change Healthcare? : OrboGraph; [cited 2020 05.07.]. Available from: https://orbograph.com/deep-learning-how-will-it-change-healthcare/.
11. Fry H. Hallo, verden Hvordan være menneske i en verden styrt av datamaskiner. Oslo: Cappelen Damm; 2020. 107, 267 p.
12. Lhessani S. What is the difference between training and test dataset? : Medium.com; 2019 [cited 2020 14.08]. Available from: https://medium.com/@lhessani.sa/what-is-the-difference-between-training-and-test-dataset-91308080a4e8.
13. Laguipo ABB. Artificial Intelligence Enables Rapid COVID-19 Lung Imaging Analysis at UC San Diego Health

. New Medical Life Sciences 2020 9. april.
14. Watanabe AT, Lim V, Vu HX, Chim R, Weise E, Liu J, et al. Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography. J Digit Imaging. 2019;32(4):625-37.
15. Longoni C. AI Can Outperform Doctors. So Why Don't Patients Trust It? : Harvard Business Review; 2019 [Available from: https://hbr.org/2019/10/ai-can-outperform-doctors-so-why-dont-patients-trust-it.
16. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop. 2017;88(6):581-6.
17. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018;115(45):11591-6.
18. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol. 2019;29(10):5469-77.
19. Gan K, Xu D, Lin Y, Shen Y, Zhang T, Hu K, et al. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta Orthop. 2019;90(4):394-400.
20. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop. 2018;89(4):468-73.