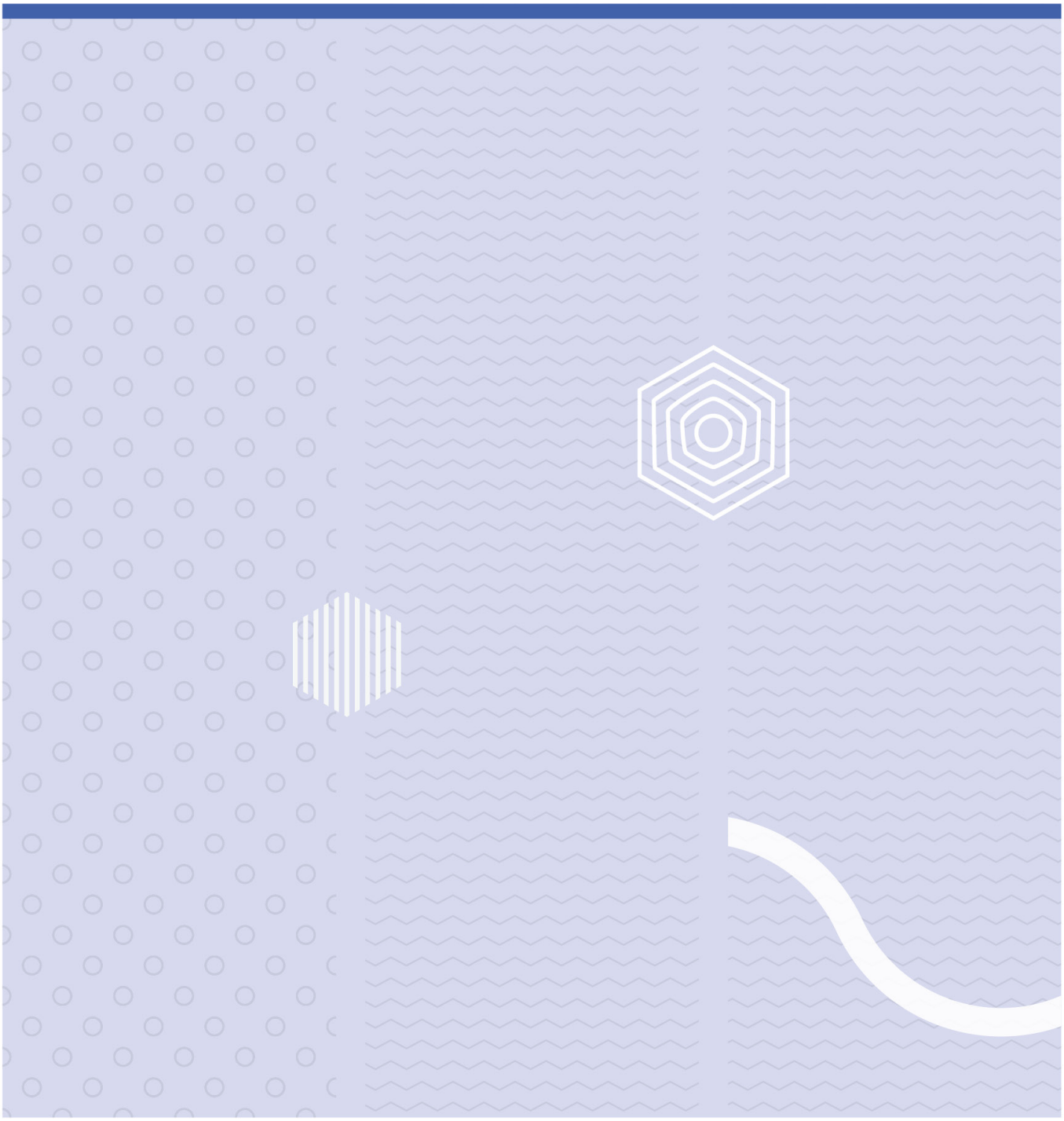
Universitetet i Sørøst-Norge
Fakultet for Helse- og sosialvitenskap
Institutt for optometri, radiografi og lysdesign
–

Bacheloroppgave
20/05/2021
RADFOR610-1 Radiografi – forskning og utvikling

Kandidatnummer: 6045, 6047 & 6057.

«Hvor nøyaktig er dyp læring i forhold til radiologer ved tolkning av skjelettalder-bilder?



Innholdsfortegnelse

Forord

Dette er en avsluttende bacheloroppgave ved radiografutdannelsen ved Universitetet i Sørøst-Norge. I denne oppgaven ønsket vi å se på hvordan dyp læringsmodeller er i forhold til radiologer, og hvor nøyaktige de er. Det er et spennende tema å jobbe med, da det er relativt nytt og gir økt kunnskap om hvordan disse modellene virker. Vi håper at oppgaven kan være nyttig for andre innen radiologiens verden.

Vi ønsker å takke vår veileder, Endre Grøvik, for god veiledning og gode tilbakemeldinger under arbeidet med oppgaven. Vi vil også takke CRAI ved OUS, for at vi fikk være tilskuere under spennende digitale foredrag innen kunstig intelligens. Til slutt vil takke bibliotekar Jana Myrvold for god hjelp under søkeprosessen og utforming av PICO-skjema.

Drammen, mai 2021

Kandidatnummer: 6045, 6047 & 6057.
RADFOR610, Universitetet i Sørøst-Norge.

Sammendrag

Mål

Målet med oppgaven er å se hvor nøyaktige dyp læringsmodeller er i forhold til radiologer som tolker skjelettalder bilder.

Metode

Det har blitt benyttet et systematisk litteraturstudie. Søk er gjennomført i databaser som Embase-Ovid og MEDLINE. Søkeordene som er benyttet er skjelettalder, dyp læring, maskinlæring og kunstig intelligens. Kvalitetssikring av artiklene ble gjort med hjelp av skjemaer fra helsebiblioteket.

Resultat

Det ble valgt ut fire artikler som alle sier noe om nøyaktighet mellom dyp læringsmodeller og radiologer. Artiklene nevner også begrensninger og hvordan modellene kan brukes, og bli bedre.

Konklusjon

Nøyaktigheten til de ulike dyp læringsmodellene er noe varierende, ut fra hvilken aldersgruppe som beskrives av modellen. Alle studiene tar opp forskjellige modeller, og at en av begrensningene til modellene er at det er for lite data for enkelte aldersgrupper.

Nøkkelord

Skjelettalder, Kunstig intelligens, Dyp Læring, Nøyaktighet, Radiolog og Radiograf.

Ordliste

Nøyaktighet	Det er benyttet begrep som RMS, MAD, MAE, mAP og RMSE i oppgaven. Dette er standardmål for nøyaktighet og måler hvor god dyp læringsmodellen er i forhold til radiologene.
RMS, Root Mean Square	Er det kvadratiske gjennomsnittet, og er en statistisk middelverdi av et sett med tall eller målserie av en variabel størrelse.
MAD, Mean Average Difference	Er utregning hvor man ser på de gjennomsnittlige variasjonene på dataene som blir utregnet frem til gjennomsnittet.(Khan Academy, u.å.)
MAE, Mean Average Error	Er gjennomsnittet over verifiseringsutvalget av de absolutte verdiene av forskjellene mellom prognosen og den tilsvarende observasjonen. MAE er en lineær poengsum som betyr av individuelle forskjeller vektes likt i gjennomsnittet (Eumetrain, u.å.)
RMSE, Root Mean Square Error	Er en kvadratisk målingsregel som måler den gjennomsnittlige størrelsen på feil. Forskjellen mellom prognose og tilsvarende observerte verdier hver kvadrat, og deretter gjennomsnittet over prøven. Til slutt tas kvadratroten av gjennomsnittet, og gir RMSE en relativ høy vekt til store feil. RMSE er mest nyttig når store feil er uønsket, og vil alltid være lik eller av større verdi som MAE, og jo større forskjell jo større varians i feilene. Dersom MAE og RMSE er like, er feilene like store. (Eumetrain, u.å.)
mAP, Mean Average Precision	Er en utregning av gjennomsnittlig gjennomsnitt av modellen og radiologene. Det gir en evaluering på presisjonen, og hvor effektive de er i gjennomsnitt på en skala fra 0 til 1,0. (Tan, 2019)
GP-metode	Tolkning basert på Greulich og Pyles atlas.
TW-metode	Tolkning basert på Tanner-Whitehouse metoden.

1 Innledning

I et moderne samfunn hvor enkle oppgaver erstattes av datamaskiner er det viktig å se hvor godt de fungerer, og hvor nøyaktige de er i forhold til de «gamle metodene». Det samme gjelder innen den radiologien, og for ikke mange år siden måtte man fremkalle røntgenbildene man hadde tatt. Nå i 2021 er det kunstig intelligens som er in, og særlig innen bildediagnostikk. Kan kunstig intelligens erstatte en del av arbeidet?

I denne oppgaven ønsker vi å se på bruken av dyp læring ved tolkning av skjelettalder hos barn. Dette viser seg å være en tidsbesparende metode, noe som kan gi radiologene mer tid til andre undersøkelser. I denne sammenhengen ønsker vi å se hvor nøyaktig dyp læringer i forhold til radiologer, samt belyse hvordan dette kan hjelpe radiologene i en hektisk hverdag.

Dette bunner da ut i vår problemstilling; Hvor nøyaktig er dyp læring i forhold til radiologer ved tolkning av skjelettalder-bilder?

Vi har valgt å begrense omfanget til nøyaktighet, og ikke se på effektiviteten ved bruk av dyp læring. Dette grunnet det store omfanget av studier som baserer seg på effektivitet og tidsbegrensninger. Likevel er effektivitet et viktig argument for å benytte seg av kunstig intelligens og dyp læring innen radiologi.

2 Teori

2.1 Skjelettalder

Skjelettalder er en billeddiagnostisk undersøkelse som gjøres for å avgjøre om barnet følger kronologisk alder (Tjønneland & Lagesen, 2013, s. 176). Ved denne undersøkelsen tas det røntgen av venstre hånd og håndledd, for å bestemme skjelettalderen og beregne slutthøyde (Helsebiblioteket, 2019). Det er ikke en fasit på barnets alder, men en veiledning i skjelettalder, da det kan være flere ting som spiller inn på skjelettmodningen. F.eks. kan ernæring, medikamenter, hormonforstyrrelser, malignitet og generelle oppvekstsvilkår være faktorer som spiller inn på skjelettmodningen (Tjønneland & Lagesen, 2013, s. 177). Samme metode har i lang tid blitt brukt for å bestemme skjelettalder, hvor radiologen ser på bildet og bruker et leksikon for å kunne angi skjelettalderen. Dette er en tidkrevende prosess som tar mye verdifull tid fra andre undersøkelser innen radiologi. I følge Bone Age: A Handy Tool for Pediatric Provides, tar det ca. 7,9 minutter på å fastsette skjelettalderen med TW-metoden (Creo & Schwenk, 2017)

Det finnes flere metoder for å beregne skjelettalder, og de vanligste metodene Tanner-Whitehouse metoden, og et atlas av Greulich og Pyle. TW-metoden baserer seg på 20 av beina i hånden og håndleddet, og setter en score ut fra dette. Disse er 7 av karpalbeina i hånden, mens de resterende 13 kalles RUS; Radius, Ulna og Short bones, som er de korte beina i tommelen, midtfinger og lillefinger (Kesimal & Jones, u.å.). Hvert av beinene får hver sin score basert på modenhet og kjønn, og legges sammen for å danne en totalscore, som settes inn i en graf for å bestemme alderen (Kesimal & Jones, u.å.). GP-metode baserer seg på et atlas. Atlaset inneholder referansebilder for begge kjønn fra alderen 0-19 år. Alderen måles ved å sammenlikne bilder fra atlaset og røntgenbildet (Mughal, Hassan & Ahmed, 2014).

2.2 Kunstig Intelligens

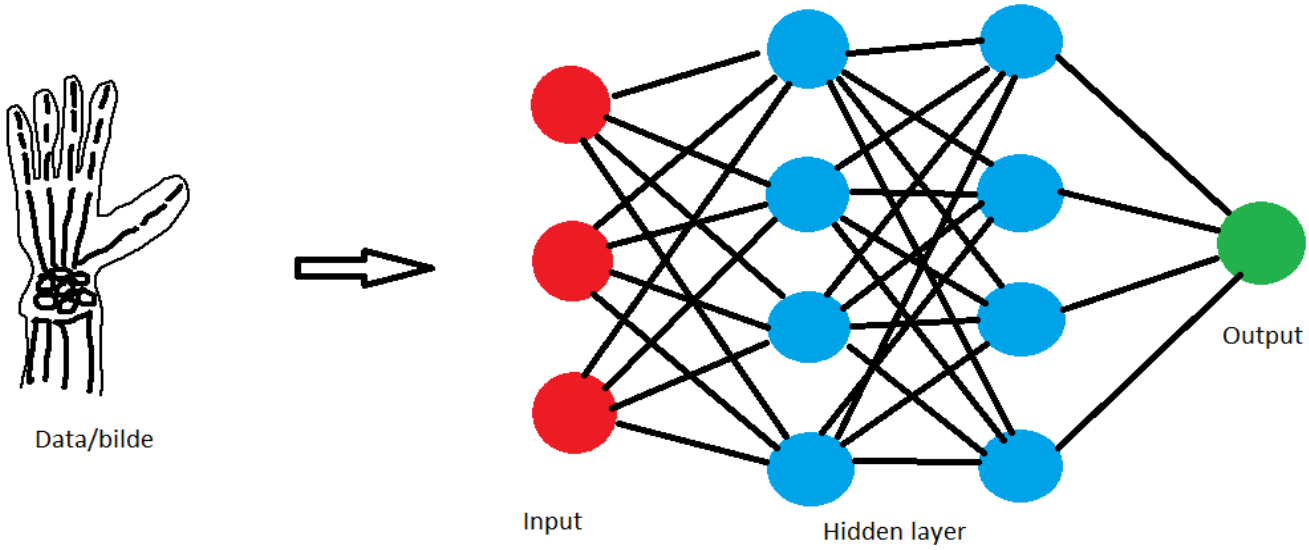
Kunstig intelligens er en type informasjonsteknologi, som kan justere sin egen aktivitet. Det er det som gjør at det fremstår som intelligent (Tidemann, 2017). Et viktig trekk ved kunstig intelligens er autonomi, at den kan operere selvstendig (Bergsjø & Bergsjø,

2019, s. 51). Innen kunstig intelligens er det flere underkategorier som maskinlæring og dyp læring.

2.3 Dyp Læring

Dyp læring er en prosess som trener opp «dype kunstige nevrale nettverk», og er en sentral metode innen maskinlæring. Det går under prinsippet at datamaskiner tilegner seg ny kunnskap, om noe den ikke kan fra før (Tidemann, 2017). Metoden bygger et nevralt nettverk som har mange lag, og for hvert lag som legges til, øker læringskapasiteten fra de lavere lagene (Tidemann 2020). Det er først et input-lag, hvor data som skal vurderes er, og dette går da gjennom flere lag, frem til output-laget, hvor resultatet kommer ut (Bjørkeng, 2018, s. 18). Det kan f.eks. være fire-fem lag, hvor hvert av lagene har hver sin prosess med gjenkjennelse. Hvis det første laget f.eks. kan gjenkjenne enkle former, kan det femte laget kjenne igjen hele hånden med alle strukturer. (Tidemann, 2020). Jo mere data det dype nevrale nettverket har, jo bedre blir modellen. I dette tilfellet vil det da si at jo flere bilder modellen har fra undersøkelser, jo bedre blir modellen til å tolke skjelettalder.

Figur 1: Illustrasjon av et lite, kunstig nevralt nettverk med flere lag.



3 Metode

Denne studien er en litteraturstudie, og for å kunne finne ut av hvordan dyp læring kan brukes til å måle skjelettalder må man finne artikler og tidligere studier. I dette tilfellet ble det funnet flere artikler etter systematiske søk i databaser. Alle artiklene vi fant måtte gjennom en prosess for å kunne se hvilke av de vi kunne benytte.

3.1 Identifisering av studier

Det ble gjennomført systematiske søk i forskjellige databaser, både MEDLINE og Embase-Ovid. Søkene ble gjennomført mellom 12. mars til 6. april, 2021. For å kunne søke mer effektivt ble det laget PICO-skjema basert på MeSH-termer, og gir en indikasjon på søkeordene vi benyttet. PICO-skjema er presentert i Vedlegg 1.

Dette ga oss en søkestrategi som presenteres i figur 2, og vi startet å søke i Embase-Ovid. På MEDLINE måtte ordet Bone Age Determination endres til Age Determination by Skeleton, for at vi skulle få treff.

Figur 2: Embase-Ovid søkestrategi. Her ser man ulike søkeord, og hvordan de ulike ordene ble brukt sammen for å danne søket vårt.

☐	8	5 and 7
☐	7	Age Determination by Skeleton/
☐	6	4 and 5
☐	5	1 or 2 or 3
☐	4	bone age determination/
☐	3	Deep Learning/
☐	2	Artificial Intelligence/
☐	1	Machine Learning/

Hovedkomponentene i søket vårt var dyp læring og skjelettalder. Det ble også benyttet ord som maskinlæring og kunstig intelligens for å sørge for at man fant alle relevante artikler og studier som omhandler temaet.

3.2 Utvelgelse av artiklene

For at artiklene skulle inkluderes i oppgaven skulle de omhandle bruken av dyp læring og at dette var satt opp mot radiologer som tolker skjelettalder-bilder. Bruken av dyp læring skal ha vært testet, og artiklene skal belyse hvor nøyaktig det er i forhold til radiologer. Vi fant flere artikler som tok opp dette, og valgte ut fire artikler. I den ene artikkelen ble det henvist til en annen studie, gjennomført av Larson et al, som vi valgte å benytte i stedet for den opprinnelige artikkelen. De artiklene som ble ekskludert er ikke relevante, og andre artikler svarte på det samme bare mer detaljert.

3.3 Kvalitetsvurdering

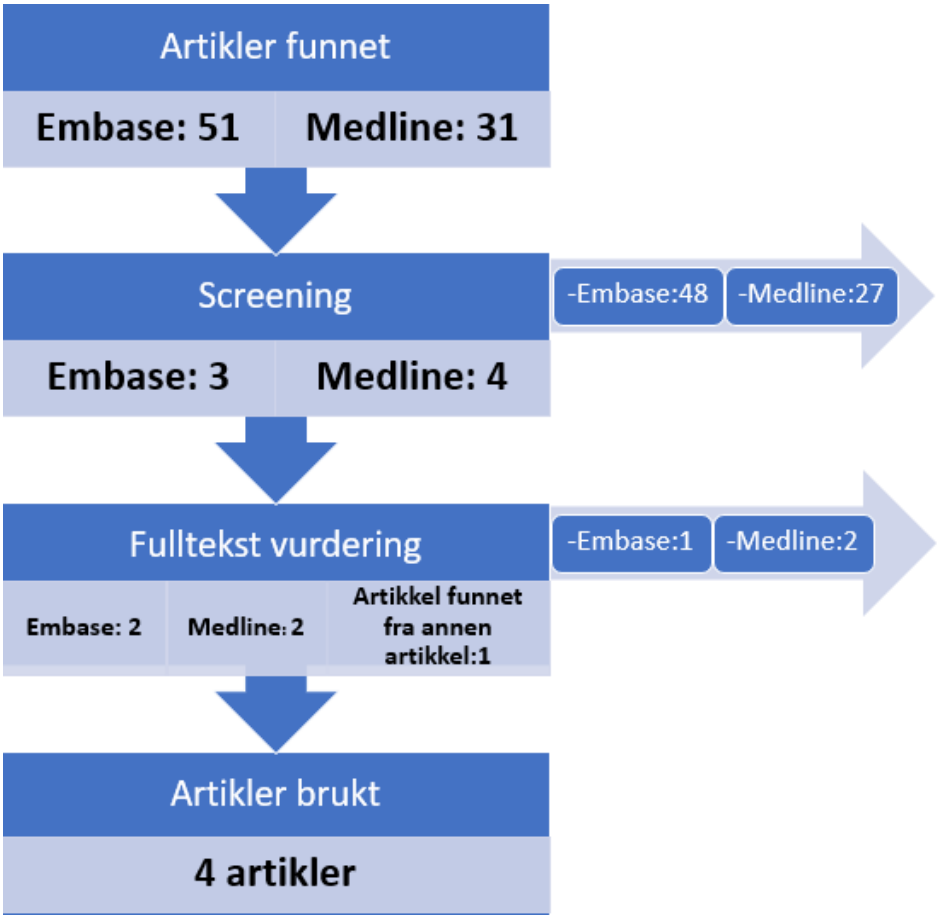
For å vurdere kvaliteten på artiklene brukte vi en sjekkliste utarbeidet av Helsebiblioteket (Helsebiblioteket, 2016). Vi brukte «Sjekkliste for vurdering av en kvalitativ studie», og gikk gjennom flere punkter for å vurdere de inkluderte artiklene.

I disse sjekklistene er det flere punkter som skal gjennomgås ved hver artikkel. Sjekklisten er i tre deler, hvor del A er innledende vurdering, del B er hva er resultatene og del C er om resultatene kan være til hjelp. Det kan f.eks. være spørsmål om artikkelen svarer på egen problemstilling, og om resultatene er til å stole på. Det er også viktig å se om resultatene kan benyttes i praksis. Noen av de andre spørsmålene er f.eks. om det kommer tydelig frem hvordan analysen er gjennomført, om det er noen etiske forhold eller lignende.

4 Resultat

Etter gjennomført litteratursøk, og funnet flere artikler i databasene, måtte artiklene vurderes. Gruppen leste gjennom artiklene fra screening-prosessen, og gjorde seg opp hver sin mening om artiklene. Deretter gikk vi sammen for å diskutere artiklene, og videre velge ut hvilke vi skulle bruke. Noen av artiklene ble valgt bort da de ikke var relevante for vår problemstilling, eller at andre artikler kunne gi mer detaljert informasjon. De ekskluderte artiklene er i Vedlegg 2. Prosessen med utvelgelse er beskrevet i flytskjemaet i figur 3. Vi endte opp med fire artikler, da disse svarte godt på vår problemstilling, og finnes i tabell 1.

Figur 3: Flytskjema fra søkeprosessen vår. Her vises det hvor mange artikler som ble funnet med de aktuelle søkeordene fra PICO-skjemaet, hvor mange som ble med videre til screening-prosessen, og hvor mange artikler vi valgte å ha med etter en fulltekst vurdering.



Tabell 1: Oversikt over inkluderte artikler, med informasjon om problemstilling eller mål for studien, antall bilder som er med i de enkelte studiene, antall radiologer som er med å tolke bildene, og hvilken journal de er publisert i.

Artikkel	Forfatter, År	Problemstilling/Mål	Metode	Antall bilder	Antall Radiologer	Journal
Automated Bone Age Assessment Using Artificial Intelligence: The Future of Bone Age Assessment	Lee, B. L. & Lee, M. S. 2020	Ser på flere dyp lærings-modeller, og hvordan de fungerer og hvor nøyaktige de er.	Greulich & Pyle, Tanner-Whitehouse metode.	VUNO Med-BoneAge: 18940 HH-boneage.io: MediAI-BA:	0	Korean Journal of Radiology
Diagnostic Performance Of Convolutional Neural Network-based Tanner-Whitehouse 3 Bone Age Assessment	Zhou, X-L., Wang, E-G., Lin, Q., Dong, G-P., Wu, W., Huang, K., Lai, C., Yu, G., Zhou, H-C., Ma, X-H., Jia, X., Shi, L., Zheng, Y-S., Liu, L-X., Ha, D., Ni, Hao., Yang, J & Fu, J-F. 2020	Ser på nøyaktigheten av en dyp-læringsmodell, som er basert på Tanner-Whitehouse 3 metoden.	Tanner-Whitehouse 3 metode.	9059	2 Radiologer, 4 Endokrinologer	Quantitative Imaging in Medicine and Surgery

Fully Automated Deep Learning System for Bone Age Assessment	Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B. A., Alkasab, T. K., Choy, G. & Do, S. 2017	Ser på nøyaktigheten ved bruk av en dyp lærings-modell basert på Greulich & Pyle metoden, hos begge kjønn. Bruker fire forskjellige dyp lærings-metoder.	Greulich & Pyle Metode	8325	0	Journal of Digital Imaging
Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs	Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V. & Langlotz, C. P 2017	Sammenlikne hvor godt dyp lærings-modellen for skjelettalder er i forhold til erfarne radiologer, og eksisterende system.	Greulich & Pyle Metode	14036	4 Radiologer	Radiology

Den ene artikkelen vi valgte å bruke var et gjennomført litteratursøk, som hadde sett på flere forskjellige dyp lærings-modeller. De andre artiklene var kvalitative studier som hadde benyttet bilder fra tidligere undersøkelser for å lære opp det dype nevrale nettverket i modellen, og som sammenliknet nøyaktigheten med legene.

4.1 Resultater fra artiklene

4.1.1 Datamengde

I artikkelen Performance of a Deep-Learning Neural Network Model in Assesing Skeletal Maturity on Pediatric Hand Radiographs, benyttes bilder fra 14036 undersøkelser, hvor 200 bilder ble benyttet til det første test-settet (Larson, Chen, Lungren, Halabi, Stence & Langlotz, 2017, s. 313). Det første test-settet av bilder ble satt sammen ved at modellen satte en gjennomsnittlig skjelettalder, samtidig som tre radiologer skulle se på de samme bildene. Disse radiologene hadde mellom 2 og 9 års erfaring (Larson et al, 2017, s. 315). Dette utgjorde referansestandarden til modellen for videre opplæring. Alle bilder som benyttes i opplæringen var tolket av radiolog, og rapportene ble automatisk hentet inn i modellen. Radiologene hadde basert tolkningene ved hjelp av GP-metoden. De inkluderte bilder av pasienter med skjelettdysplasier (Larson et al, 2017, s. 314). I det neste test-settet ble bildene delt inn i opplærings- og validerings-sett. 90% av bildene ble benyttet til opplæring, og 9% ble benyttet til validering. Bildene ble også nedskalert til 224 x 224 piksler, for at alle bilder skulle ha en standardstørrelse (Larson et al, 2017, s. 316). Det ble fjernet bilder fra undersøkelser av jenter fra 0 til 2,5 år og 17 til 19 år, grunnet lite data. Bilder for gutter mellom 0-2 år og 15-19 år ble fjernet for å muliggjøre sammenligning med tidligere arbeid.

I artikkelen Diagnostic Performance of Convolutional Neural Network-based Tanner-Whitehouse 3 Bone Age Assessment System benyttes bilder fra 9059 undersøkelser (Zhou et al, 2020, s. 658). Undersøkelsene var også tolket av radiolog før bildene ble lagt inn i modellen. Dataene ble delt inn i et opplæringssett på 8005 bilder, og et validerings-sett på 804 bilder. Det ble brukt bilder av kvinner fra 0-18 år, og menn fra 0-17 år. I det første testsettet på 250 ble det brukt 50/50 av kvinner og menn (Zhou et al, 2020, s. 661). Etter at disse bildene ble lagt inn ble det igjen brukt 250 bilder for å verifisere nøyaktigheten til modellen, sammenliknet med to radiologer og fire

endokrinologer. Bildene ble nedskalert til 256 x 256 piksler (Zhou et al, 2020, s.659). Det kommer frem flere begrensninger som påvirket modellen, blant annet at den ikke registrerte pasienter med skjelettsykdommer. Modellen hentet sine data fra ett sykehus. (Zhou et al, 2020, s. 665)

I artikkelen Fully Automated Deep Learning System for Bone Age Assessment benyttes 8325 bilder fra tidligere undersøkelser. Det ekskluderes bilder fra barn mellom 0 til 4 år, høyrehåndsbilder, deformerte bilder, og ikke-lesbare radiologirapporter (Do et al, 2017, s. 428). Her ble bildene delt inn etter kjønn, og det ble dannet egne test-sett for både jenter og gutter. Det ble tilfeldig valgt ut 15% av bildene for opplæringssett blant begge kjønn, 15% av bildene for validerings-sett, og de resterende 70% ble brukt som opplæringssett for jentegruppen og guttegruppen (Do et al, 2017, s. 428). Bildene her ble også nedskalert til 512 x 512 piksler. Her er det allerede aldersbegrensninger, ettersom alle bilder fra barn mellom 0-4 ble ekskludert. I denne artikkelen har de brukt fire forskjellige dyp læringsmodeller for å tolke skjelettalderbildene, og setter disse modellene opp mot hverandre.

Lee. & Lee. har sett på flere forskjellige dyp læringsmodeller, og diskutert hvor effektive og nøyaktige de er. Den første dyp læringsmodellen som blir tatt opp er BoneXpert. Ifølge Lee. & Lee er denne modellen mer effektiv i forhold til en erfaren radiolog ved bruk av GP-metoden. Effektiviteten og nøyaktigheten til denne modellen faller drastisk dersom det er færre enn 8 ben inkludert i bildet. Samtidig har dyp læringsmodellen problemer med tolkning dersom det er dårlig bildekvalitet, samt avvikende benutvikling. Disse avvikene kan imidlertid bli overkommelige ved jevnlige oppdateringer.

Den neste modellen som blir tatt opp av Lee. & Lee. er VUNO Med-BoneAge. Dette er en modell som ble trent med 18940 venstre hånds bilder til å bli analysert ved hjelp av GP metoden. VUNO foreslår tre mest sannsynlige skjelettaldere basert på treningsmaterialet modellen har mottatt. Ved det første bildeforslaget er nøyaktigheten 69.5%, men denne øker til 93% når alle tre forslagene er sammensatt. Dette er ikke en helautomatisk prosess siden det kreves at en person skal velge en av de tre forslagene. Modellen viste en 29% reduksjon i tiden tolkningen tar.

HH-boneage.io bruker TW3 metoden til å bestemme skjelettalderen, og finner ROI av 13 forskjellige bein i hånden og skaper et resultat ut fra disse. Denne modellen, sammenlignet med radiologer, viste 97.6% nøyaktighet, med en gjennomsnittlig feilmargin på ca. 0.62 år (Lee. & Lee., 2020, s. 797). Dette ble målt etter 1 år verdt med data.
MedAI-BA bruker også TW3 metoden for å finne skjelettalderen, og har en gjennomsnittlig feilmargin på 0.59 år. (Lee. & Lee., 2020, s. 797).

4.1.2 Nøyaktighet

Nøyaktighet i forhold til radiologene ble målt med RMS/RMSe, MAD og MAE/mAP. Alle verdiene er vist i tabell 2. Enkelte av artiklene har ikke sammenliknet dyp læringsmodellen med radiologer, men sammenliknet flere forskjellige modeller.

Tabell 2: Oversikt over RMS og MAD hos modellen og hos radiologene som er med i studiene.

Artikkel	RMS[1] Radiologer		RMS/RMSe[2] Modell		MAD[3] Radiologer	MAD/MAE[4] Modell
Automated Bone Age Assessment Using Artificial Intelligence: The Future of Bone Age Assessment			VUNO Med-BoneAge: 0,62. MediAI-BA: HH-boneage.io: 0,62			VUNO Med-BoneAge: 0,46. MediAI-BA: 0,59 HH-boneage.io: 0,46
Diagnostic Performance of Convolutional Neural Network-based Tanner-Whitehouse 3 Bone Age Assessment	TW3-Carpal: R1: 0,74 R2: 0,80 R3: 0,91 R4: 0,78 R5: 1,18 R6: 0,93 Gjennomsnitt: 0,89	TW3-RUS: R1: 0,73 R2: 0,85 R3: 0,91 R4: 0,78 R5: 1,15 R6: 1,03 Gjennomsnitt: 0,91	TW3-Carpal: 0,54 0,56 0,56 0,58 0,36 0,37 Gjennomsnitt: 0,50	TW3-RUS: 0,57 0,57 0,57 0,58 0,35 0,38 Gjennomsnitt: 0,50		
Fully Automated Deep Learning System for Bone Age Assessment			RMSe: M1: 1,51 (F) 1,45 (M) M2: 1,08 (F) 0,93 (M) M3: 1,07 (F) 0,91 (M) M4: 0,93 (F) 0,82 (M)			mAP[5]: M1: 33,8% (F) 32,4% (M) M2: 47,5% (F) 49,5% (M) M3: 48,3% (F) 50,6% (M) M4: 53,3% (F) 55,8% (M)

		Gjennomsnitt: 1,14 (F) 1,02 (M)		Gjennomsnitt: 45,7% (F) 47,07% (M)
Performance of a Deep Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs.	Radiolog 1: 0,73 Radiolog 2: 0,73 Radiolog 3: 0,95 Radiologirapport: 0,87 Gjennomsnitt: 0,82	0,67 0,67 0,68 Radiologirapport: 0,6 Gjennomsnitt: 0,67	Radiolog 1: 0,55 Radiolog 2: 0,53 Radiolog 3: 0,69 Radiologirapport: 0,65 Gjennomsnitt: 0,61	0,53 0,53 0,53 Radiologirapport: 0,51 Gjennomsnitt: 0,52
1 – RMS = Root Mean Square 2 – RMSE = Root Mean Square Error 3 – MAD = Mean Absolute Difference 4 – MAE = Mean Absolute Error 5 – mAP = Mean Average Precision				

I den første artikkelen i tabellen er det målt RMS og MAD for de ulike modellene. Dette skyldes at artikkelen sammenlikner flere dyp læringsmodeller, og ser på nøyaktigheten. Her benytter de RMSE og MAE, for å måle nøyaktigheten. Den første modellen, VUNO Med-BoneAge, har henholdsvis 0,62 RMSE og 0,46 MAE, som vil si at det er variasjon mellom feilene. For modellen MediAI-BA er det kun oppgitt en MAE lik 0,59, som er gjennomsnittet av alle feil i modellen. For den siste modellen, HH-boneage.io, oppgis det RMSE på 0,62 og MAE på 0,46, som er like verdier som den første modellen.

I artikkel nummer 2 i tabellen sammenliknes RMS mellom radiologer og to dyp-læringsmodeller. For den første modellen, TW3-Carpal, er gjennomsnitts RMS for radiologer/endokrinologer 0,89, mens for TW3-RUS er gjennomsnitts RMS 0,91. For dyp læringsmodellen er gjennomsnitts RMS for TW3-Carpal og TW3-RUS 0,50. Dette vil si at TW3-RUS og TW3-Carpal er like gode ved dyp læringsmodellen, mens ved radiologene er det en liten forskjell.

I artikkel nummer 3 bruker de RMSE og mAP, Her gjøres det en vurdering mellom modellene som brukes på menn og kvinner med 4 modeller til hvert kjønn. M1 bruker originale bilder, M2 bruker prosesserte bilder, M3 bruker prosesserte og data argumentasjon, mens M4 bruker en forhåndstrent modell, samt prosesserte bilder og data argumentasjon. Alle modellene blir testet på RMSE i hvor stort sprik i alder som modellene gjettet og mAP i løpet av 2 år.

I den siste artikkelen er det oppgitt RMS og MAD for både radiologer og modellen. Gjennomsnitts RMS for radiologene var 0,82, og skjelettalderen fra radiologirapportene er tatt med i beregningene. For modellen var gjennomsnitts RMS 0,67, som er lavere enn radiologenes, som vil si at modellen er mer nøyaktig når det kommer til å fastsette skjelettalder. Gjennomsnitts MAD for radiologene er 0,61 og for modellen er det 0,52. Her er det lavere verdi for modellen som sier at den er mer nøyaktig.

5 Diskusjon

Det er varierende nøyaktighet mellom radiologene og dyp læringsmodellene. Dette skyldes blant annet aldersbegrensningene som er i de ulike artiklene. Alle artiklene vi har brukt, sier at aldersbegrensningene skyldes for lite data. Som nevnt i teoridelen krever dyp læringsmodeller data for å kunne tolke bildene. Et eksempel er fra artikkelen av Larson hvor det er oppført en figur som viser mengde data fra de ulike aldersgruppene (Larson et al, 2017, s. 318). Det er lite data fra gruppen fra 0 til 5 år, og mellom 16 til 19 år, både for test- og valideringssettet. De har hentet ut data fra to forskjellige sykehus. Dersom de hadde hentet ut data fra flere sykehus, kunne de ha fått mer data i disse aldersgruppene, og økt nøyaktigheten. Det samme gjelder for artikkelen av Do (Do et al, 2017, s. 428) hvor undersøkelser av barn mellom 0 til 4 år var ekskludert.

Vedrørende disse aldersgruppene, og hvilken standardmetode benyttes, er det naturlige å gå tilbake til både GP's atlas og TW-metoden, alt ettersom hva som er aktuelt i landet.

Skjelettets utseende påvirkes også av etnisitet. Dermed vil skjelettalderen bli påvirket av hvilken etnisitet man tilhører (Lee & Lee, 2020, s.797). Dette medfører at modellen scorer dårligere ved forskjellige etnisiteter, da modellen mangler data fra forskjellige folkegrupper, som igjen påvirker nøyaktigheten. For å kunne forbedre modellen ved slike tilfeller vil det igjen kreve mer data, slik at modellen kan skille mellom forskjellene, og øke nøyaktigheten.

For at modellen etter hvert skal kunne klare å oppdage sykdommer og andre tilstander i skjelettet hos barn, må man igjen ha mer data. Radiologer har flere muligheter når det kommer til å oppdage dette, da de har et trent øye og har flere kilder å basere seg på enn det modellen har. Når man kan hente ut flere data fra andre verdensdeler og data fra flere etnisiteter, kan man også inkludere data fra pasienter med allerede oppdaget skjelettforandringer og dysplasier, for at modellene skal kunne oppdage disse selv etter opplæring. Modellen kan bli trent opp ved hjelp av «augmented data», altså fabrikkert data som har hensikt å lære maskinen kjennetegn ved et skjelettalderbilde. Motargumentet til dette kan være at de fabrikkerte dataene som blir laget ikke er tilstrekkelig for en reell situasjon, samt at maskinen ikke lærer av spesielle kasus som kommer frem.

Ved bruk av dyp læringsmodeller i klinikken bør det være en form for sikkerhetsventil, i tilfeller hvor dyp læringsmodellen ikke er sikker på tolkningen. Da vil en radiolog kunne komme inn for å kontrollere svaret modellen har gitt. Sykehuset velger selv hvilken prosent dette skal være på. Det kan f.eks. være tilfeller der modellen er >95% sikker på tolkningen. Sikkerhetsventilen er til for å sørge at nøyaktigheten til modellene opprettholdes, og kan forbedres ved at den blir gitt mer data hvor usikkerheten er høy. Dette vil være en kontinuerlig prosess med forbedringer som fører til at det ikke lenger vil være behov for en sikkerhetsventil.

Som nevnt trenger modellen store mengder data for å kunne trenes opp til å gjøre gode vurderinger. Innen personvern er det noen begrensninger, og derfor har modellene kun basert seg på bilder fra ett eller to sykehus. Dette medfører at datamengden er begrenset, og kan påvirke generaliserbarheten til modellen. Dette kan løses ved at modellen f.eks. sendes til flere sykehus, i stedet for at sykehusene må dele data. Dette forhindrer at sensitiv informasjon om pasientene sendes ut, og at modellen kan bli bedre innen flere aldersgrupper.

Et slikt system vil naturligvis ha innvirkning på radiografer, og mange av dyp læringsmodellene i dag har problemer med artefakter og dårlig bildekvalitet. Dette kan sees ved f.eks. BoneXpert (Lee & Lee, 2020, s. 794) der man ofte ser at modellen sliter med å prestere på et godt nivå ved slike feil. Dette fører til at radiografer må gjennomføre sitt arbeid presist slik at modellen kan gjenkjenne strukturene. Kravet til kompetanse forblir slik at automatikken ved en dyp læringsmodell forblir.

Når det gjelder effektivitet, er det et viktig argument for å benytte seg av kunstig intelligens og dyp læringsmodeller i radiologiens verden. Dette viser seg å være tidsbesparende, og kan være med på å avlaste radiologene som tolker bilder flere timer om dagen. Det er også viktig å nevne at det er mye arbeid som skal til for å kunne implementere kunstig intelligens i helsesektoren, men at etter hvert som tiden går vil det være mer av det.

5.1 Metodekritikk

Det er viktig å opplyse om at relevant informasjon kan ha blitt ekskludert, og dette kan skyldes valg tatt under uthenting og utvelgelse av artikler/studier. Det kan også ha oppstått feil under tolkning og oversettelse, da alle artiklene er på engelsk. Alle resultatene kommer fra utenfor Europa. I tillegg ser ikke alle artiklene på sammenlikning mellom dyp læringsmodellen og radiologer, noe som kan påvirke resultatet av nøyaktighet.

6 Konklusjon

Dyp læringsmodellene har per dags dato en noe begrenset nøyaktighet, dette kommer av manglende data til å lære opp slike modeller. Men det kan fungere som et enestående hjelpemiddel til radiologer, og føre til at radiologer kan utføre arbeidet sitt raskere og kan bruke spart tid på andre viktigere arbeidsoppgaver. Modellene kan også fungere alene med tilstrekkelige sikkerhetsventiler. Noe av modellene er veldig nøyaktige på «standard» kasus, men nøyaktigheten faller av når det kommer kasus utenom det vanlige.

7 Litteraturliste

Bersgjø, L. O & Bergsjø, H. (2019) *Digital Etikk: Big Data, Algoritmer og Kunstig Intelligens.* Universitetsforlaget.

Bjørkeng, P. K. (2018) *Kunstig Intelligens: Den Usynlige Revolusjonen.* Vega Forlag.

Creo, A. L. & Schwenk, W. F. (2017) Bone Age: A Handy Tool for Pediatric Providers. *Pediatrics: Official Journal of The American Academy of Pediatrics.* 140 (6), 2. https://doi.org/10.1542/peds.2017-1486

Eumetrain. (u.å.) *Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).* Hentet 3.mai fra http://www.eumetrain.org/data/4/451/english/msg/ver_cont_var/uos3/uos3_ko1.htm

Helsebiblioteket. (2019) *Pediatriveiledere: 2.2 Kortvoksthet og vekstretardasjon.* Hentet fra: https://www.helsebiblioteket.no/pediatriveiledere?key=144404&menuitemkeylev1=5962&menuitemkeylev2=5964

Helsebiblioteket. (2016) *Sjekklister.* Hentet fra: https://www.helsebiblioteket.no/kunnskapsbasert-praksis/kritisk-vurdering/sjekklister

Kesimal, U. & Jones, J. (u.å.) *Tanner-Whitehouse Method.* Hentet 8.april 2021 fra https://radiopaedia.org/articles/tanner-whitehouse-method

Khan Academy. (u.å.) *Mean Absolute Deviation (MAD).* Hentet 2. Mai fra https://www.khanacademy.org/math/cc-sixth-grade-math/cc-6th-data-statistics/cc-6-mad/v/mean-absolute-deviation

Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V. & Langlotz, C. P. (2017) Performance of a Deep Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. *Radiology.* 287: 313-322. https://doi.org/10.1148/radiol.2017170236

Lee, B-D. & Lee M. S. (2020) Automated Bone Age Assessment Using Artificial Intelligence: The Future of Bone Age Assessment. *Korean Journal of Radiology.* 22(5):792-800 https://doi.org/10.3348/kjr.2020.0941

Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B. A., Alkasab, T. K., Choy, G. & Do, S. (2017) Fully Automated Deep Learning System for Bone Age Assessment. *Journal of Digital Imaging* 30:427-441 https://link.springer.com/article/10.1007/s10278-017-9955-8

Mughal, A. M., Hassan, N. & Ahmed, A. (2014) *Bone Age Assessment Methods: A Critical Review.* 30 (1): 211-215. Hentet fra: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3955574/

Tan, R. J. (2019, 24. Mars) *Breaking Down Mean Average Precision (mAP)*. Towards Data Science. Hentet fra: https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52

Tidemann, A. (2017, 28. November) *Dyp Læring.* Store Norske Leksikon. Hentet fra: https://snl.no/dyp_l%C3%A6ring

Tidemann, A. (2020, 8. Januar) *Kunstig Intelligens.* Store Norske Leksikon. Hentet fra: https://snl.no/kunstig_intelligens

Tjønneland, R. M. & Lagesen, B. (2013) *Barneradiografi: En veiledning i praksis.* Fagbokforlaget.

Zhou, X-L., Wang, E-G., Lin, Q., Dong, G-P., Wu, W., Huang, K., Lai, C., Yu, G., Zhou, H-C., Ma, X-H., Jia, X., Shi, L., Zheng, Y-S., Liu, L-X., Ha, D., Ni, H., Yang, J. & Fu, J-F. (2020) Diagnostic Performance of Convolutional neural Network-based Tanner- Whitehouse 3 Bone Age Assessment System. *Quantitative Imaging in Medicine and Surgery.* 2020;10(3):657-667. https://doi.org/10.21037/qims.2020.02.20

8 Vedlegg

Vedlegg 1: PICO-skjema

P	I	CO
Skjelettalder	Dyp læring	Nøyaktighet
	Maskinlæring	
	Kunstig intelligens	

Vedlegg 2: Ekskluderte artikler

Forfatter	År	Artikkel	Ekskluderingsgrunnlag
Hardy, M & Harvey, H.	2019	Artifical intelligence in diagnostic imaging: impact on the radiography profession	Forklarer hvordan AI kan implementeres i yrket. Ikke relevant for vår problemstilling.
Tajmir, S. H., Lee, H., Shailam, R., Gale, H. I., Nguyen, J. C., Westra, S. J., Lim, R., Yune, S., Gee, M. S & Do, S.	2018	Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability	Veldig overfladisk og kort artikkel. Relevant for problemstillingen, men andre artikler ga samme informasjon, men mer detaljert.
Steenkiste, T. V., Ruyssinck, J., Janssens, O., Vandersmissen, B., Vandecasteele, F., Devolder, P., Achten, E., Van Hoecke, S., Deschrijver, D. & Dhaene, T.	2018	Automated Assessment of Bone Age Using Deep Learning and Gaussian Process Regression	For kort og lite detaljer i forhold til vår problemstilling.

Wang, F., Cidan, W., Gu, X., Chen, S., Yin, W., Liu, Y., Shi, L., Pan, H. & Jin, Z.	2021	Performance of an artificial intelligence system for bone age assessment in Tibet.	Ikke helt relevant for problemstillingen vår, da vi mener den fokuserer for mye på en lokal befolkning.