Bacheloroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for medisin og helsevitenskap
Institutt for helsevitenskap i Gjøvik

Inger-Marie Hoel
Kristina Marie Adamsen
Mikela Dawn Estabrook

Kunstig intelligens ved frakturdeteksjon: Har radiologiens nye paradigmeskifte startet?

Artificial intelligence in fracture detection: Has the new paradigm shift in radiology begun?

Bacheloroppgave i Radiografi
Veileder: Astrid Berntsen
Mai 2024

NTNU
Kunnskap for en bedre verden

Inger-Marie Hoel
Kristina Marie Adamsen
Mikela Dawn Estabrook

Kunstig intelligens ved frakturdeteksjon: Har radiologiens nye paradigmeskifte startet?

Artificial intelligence in fracture detection: Has the new paradigm shift in radiology begun?

Bacheloroppgave i Radiografi
Veileder: Astrid Berntsen
Mai 2024

Norges teknisk-naturvitenskapelige universitet
Fakultet for medisin og helsevitenskap
Institutt for helsevitenskap i Gjøvik

NTNU
Kunnskap for en bedre verden

Forord

Denne oppgaven er det avsluttende arbeidet for vår bachelorgrad ved NTNU i Gjøvik. Disse tre årene har vært en utrolig reise som har ført til nye vennskap og gode minner som vil vare livet ut, men har også bydd på utfordringer og tøffe tider. Til gjengjeld har vi opparbeidet oss verdifull kunnskap og erfaringer som vi tar med oss videre.

Det har vært en givende og interessant prosess å arbeide med bacheloroppgaven. Fra høsten 2023 til våren 2024 har det blitt lagt ned utallige timer i planlegging og gjennomføring, noe vi håper gjenspeiles i oppgaven.

Først og fremst vil vi takke hverandre for det gode samarbeidet. Underveis i prosessen har vi møtt på utfordringer og uenigheter, men gjennom diskusjoner har vi sammen funnet løsninger. Vi ønsker å takke vår veileder Astrid Berntsen for støtte og hjelp med bacheloroppgaven vår, dine innspill og tilbakemeldinger har vært til stor hjelp. Vi retter også en takk til dere som har lest gjennom oppgaven og kommet med konstruktiv kritikk. Det er med stor glede og stolthet at vi kan presentere en ferdigstilt bacheloroppgave.

"It always seems impossible until it's done."

– Nelson Mandela

Gjøvik, 08.mai 2024

Inger-Marie Hoel, Kristina Marie Adamsen og Mikela Dawn Estabrook

21BRADIO, NTNU i Gjøvik

Innhold

Sammendrag 8

Abstract 9

Forkortelser 10

Ordforklaringer 11

1.0 Innledning 12

1.1 Bakgrunn for valg av tema 12

1.2 Formål med oppgaven 13

1.3 Forskningsspørsmål 13

1.4 Radiograffaglig relevans 14

1.5 Avgrensning 14

1.6 Oppgavens oppbygging 15

2.0 Teori 16

2.1 Hva er kunstig intelligens? 16

2.1.1 Ulike retninger innen kunstig intelligens 16

2.2 Kunstig intelligens innenfor radiologi 18

2.2.1 Kunstig intelligens innen konvensjonell røntgen i dag 18

2.3 Sensitivitet og spesifisitet 19

2.4 Generelt om frakturer i hånd og håndledd 20

2.4.1 Distal radiusfraktur 20

2.4.2 Scaphoidfraktur 21

3.0 Metode 22

3.1 Litteraturstudie 22

3.1.1 Begrunnelse for valg av metode 22

3.1.2 Fremgangsmåte 22

3.2 PICO-skjema 23

3.3 Søkeprosessen og datautvalg 23
3.4 Inklusjons- og eksklusjonskriterier 24
3.5 Flytskjema 26
3.6 Kvalitetsvurdering av utvalgte studier 28
3.7 Analyse 28
3.8 Etiske implikasjoner 30
4.0 Resultater 31
4.1 Utvalgte studier 31
4.2 Sammenfatning av resultatene 41
4.2.1 Sensitivitet 41
4.2.2 Spesifisitet 44
5.0 Diskusjon 47
5.1 Sensitivitet og spesifisitet 47
5.2 Forskjell mellom distale radius og scaphoid 50
5.3 Yrker og erfaring 51
5.4 Metodekritikk 53
6.0 Konklusjon 56
7.0 Litteraturliste 57
Studier inkludert i oppgaven 57
Annet litteratur 58
Figurer 61
Vedlegg 61
Vedlegg 1: Søkehistorikk 61
Vedlegg 2: Utregning av gjennomsnittet av Se og Sp til utvalgte studier 64
Vedlegg 3: Utregning av gjennomsnitt for yrkeserfaring i Li et al. (2023) 64

Sammendrag

Forskningsspørsmål: Kunstig intelligens versus bildetolkende leger: kan KI oppnå like høy sensitivitet og spesifisitet som en bildetolkende lege ved granskning av distale radius- og scaphoidfrakturer?

Formål: Hensikten med oppgaven er å sammenlikne evnen til KI-algoritmer og ulike legegrupper som tolker bilder, når det kommer til å detektere frakturer i distale radius og scaphoid. De ulike legegruppene inkluderer radiologer, ortopeder, kirurger og en akuttmottakslege, som også har forskjellig års erfaring med å granske konvensjonelle røntgenbilder. Faktorene som undersøkes er sensitiviteten og spesifisiteten innenfor de to anatomiske områdene, der dette skal knyttes opp mot KI og de ulike yrkesgruppene.

Metode: For å besvare oppgavens forskningsspørsmål er det benyttet en kvalitativ litteraturstudie. Grunnlaget for oppgaven er ni studier som undersøker sensitiviteten og/eller spesifisiteten til KI og bildetolkende leger separat.

Resultat: Resultatene for sensitivitet og spesifisitet fra de ulike studiene er presentert i to tabeller, og verdiene er oppgitt i prosenter. Basert på disse verdiene er det mulig å observere både større og mindre forskjeller mellom KI-algoritmen og de bildetolkende legene. Det kan ses at KI presterer på et nivå som tilsvarer legene. I tillegg kan det ses en lavere evne til å detektere scaphoidfrakturer i forhold til distale radiusfrakturer, både hos KI og de bildetolkende legene.

Konklusjon: Oppgaven konkluderer med at KI har evne til å detektere og utelukke frakturer i distale radius og scaphoid, tilsvarende likt eller bedre enn de bildetolkende legene. I tillegg kan KI brukes som et verktøy for leger med mindre erfaring. Det trengs fortsatt mer forskning innen temaet.

Nøkkelord: Fraktur, Scaphoid, Distale Radius, Håndledd, Sensitivitet, Spesifisitet, Kunstig Intelligens, Radiolog, Ortoped, Kirurg, Akuttmottakslege, Konvensjonell Røntgen

Antall ord: 9719

Abstract

Research topic: Artificial Intelligence versus image-interpreting physicians: can AI achieve the same sensitivity and specificity as an image-interpreting physician, during interpretation of distal radius and scaphoid fractures?

Purpose: The purpose of this thesis is to compare the ability of AI algorithms and different physician groups who interpret images, when detecting fractures located in the distal radius and scaphoid. The various groups of physicians include radiologists, orthopedics, surgeons and an ED physician, who all have differing years of experience regarding interpretation of conventional x-ray images. This thesis examines the factors of sensitivity and specificity, where these are compared within the two anatomical areas, between the various occupational groups and AI.

Method: To address the research topic, a literature study was conducted. A total of nine studies serves as the foundation for this thesis. These studies examine the sensitivity and/or specificity of AI and image-interpreting physicians separately.

Results: The results of the sensitivity and specificity from the various studies are presented in two tables, where the values are in percentages. Based on these values, it is possible to observe both major and minor differences between the AI-algorithms and the image-interpretating physicians. It can also be observed that AI performs at a level equivalent to the physicians. In addition, a lower ability to detect scaphoid fractures compared to distal radius fractures can be seen, for both the AI and the image- interpretating physicians.

Conclusion: This thesis concludes that AI is able to detect and rule out fractures within the distal radius and scaphoid, with an equivalent or greater ability than the image-interpretating physicians. Additionally, AI can be used as a tool to aid physicians with less experience. Further research is required on this subject.

Keywords: Fracture, Scaphoid, Distal Radius, Wrist, Sensitivity, Specificity, Artificial Intelligence, Radiologist, Orthopedic, Surgeon, ED physician, Conventional X-ray

Number of words: 9719

Forkortelser

Forkortelser:	Forklaring:
AUC	Area under the curve
CNN	Convolutional neural network / Konvensjonelt nevralt nettverk
DRF	Distal radiusfraktur
IRR	Initial radiology reports
KI	Kunstig intelligens
LIS	Leger i spesialisering
NPV	Negative predictive values / Negativ prediktive verdier
PPV	Positive predictive values / Positiv prediktiv verdi
ROI	Region of interest
Se	Sensitivitet
Sp	Spesifisitet

Tabell 1: Ordforklaring på forkortelser benyttet i oppgaven.

Ordforklaringer

Ordforklaring:	
Accuracy	Er målingen på andelen riktige vurderinger blant de totale granskede bilder (OECD.AI, u.å.)
Bildetolkende leger (eget definert ord)	Yrker som tolker bilder. I denne oppgaven omfatter det: radiologer, ortopeder, kirurger og akuttmottakslege
CE-merket	«Conformité Européenne» er et merke som brukes hos produkter på markedet i det Europeiske Økonomiske Samarbeidsområdet (EØS). Det betyr at produktene som blir solgt i EØS, er godkjent og vurdert til å ha oppfylt de strenge kravene innen helse, miljøvern og sikkerhet (European commission, u.å.)
FDA-godkjent	«Food and Drug Administration», har som ansvar å beskytte folkehelsen i USA ved å kvalitetssikre humane og veterinære legemidler, medisisnsk utstyr, kosmetikk, biologiske produkter og produkter som avgir stråling (FDA, u.å.)
Falsk negativ (False negative, FN)	En bildediagnostikk test vurderes som negativ, men pasienten har fraktur (Morgan *et al.*, 2015)
Falsk positiv (False positiv, FP)	En bildediagnostikk test vurderes som positiv, men pasienten har ikke fraktur (Morgan *et al.*, 2015)
MURA	Et stort datasett med røntgenbilder av muskel og skjelett (Rajpurkar *et al.*, 2018)
Sann negativ (True negativ, TN)	En bildediagnostikk test vurderes som negativ, og pasienten har ikke fraktur (Morgan *et al.*, 2015)
Sann positiv (True positiv, TP)	En bildediagnostikk test vurderes som positiv, og pasienten har fraktur (Morgan *et al.*, 2015)

Tabell 2: Ordforklaring på begreper benyttet i oppgaven.

1.0 Innledning

Kunstig intelligens (KI) er i ferd med å bli en større del av hverdagen, også innen radiologi. En metaanalyse utført av Jung *et al.* (2024), viser at KI innen bildediagnostikk hovedsakelig er utprøvd i Asia, Europa og Nord-Amerika. Denne teknologien er tatt i bruk på flere modaliteter, blant annet CT, ultralyd, MR, nukleærmedisin, mammografi, stråleterapi og konvensjonell røntgen (Malamateniou *et al.*, 2021).

På konvensjonell røntgen er fraktur i hånd og håndledd noen av de vanligere skadene som avbildes (Hamblen & Simpson, 2007, s. 191). I framtiden vil forekomsten av disse frakturene sannsynligvis ikke avta, særlig med tanke på økt befolkningstetthet og en stadig høyere gjennomsnittsalder (Hernæs & Skyrud, 2022, s. 26). Antall leger som gransker bilder øker ikke i takt med denne samfunnsutviklingen, da radiologi er et fagområde som allerede har utfordringer når det gjelder rekruttering av spesialister (Aase *et al.*, 2022, s. 8). Dette resulterer i mange røntgenbilder som fordeles på et begrenset antall leger, med en forventning om at effektiviteten skal opprettholdes. Det økte presset på legene kan føre til flere feildiagnoser (Cohen *et al.*, 2023). I tillegg kan andre faktorer som mangel på erfaring og fatigue påvirke diagnostiseringsevnen (Zhang *et al.*, 2023).

KI er nå tatt i bruk på konvensjonell røntgen, blant annet til å detektere frakturer. Denne implementeringen kan potensielt være med på å redusere arbeidsbelastningen for legene, og samtidig øke effektiviteten (Vestre Viken, 2024). Det kan derimot undres om KI har evne til å detektere og utelukke frakturer på lik linje med legene. Denne bacheloroppgaven skal derfor undersøke hvor sensitiv og spesifikk kunstig intelligens er, sammenliknet med leger som tolker bilder, ved deteksjon av distale radius- og scaphoidfrakturer.

1.1 Bakgrunn for valg av tema

Bakgrunnen for valget av temaet oppstod som følge av forelesninger om kunstig intelligens, i tillegg til det pågående prosjektet i Vestre Viken som har blitt omtalt mye i nyhetene. Dette prosjektet heter «Bruk av kunstig intelligens i bildediagnostikk», og har som hensikt å undersøke muligheter og utfordringer ved implementering av KI til å diagnostisere frakturer. For eksempel undersøkes det om KI kan forbedre kapasiteten til sykehusene, samt

pasientflyten ved å redusere ventetid på svar, i tillegg til å lette arbeidshverdagen for de ansatte. Per den 10.04.2024, er det fem sykehus i Vestre Viken som har KI i klinisk drift: Bærum sykehus, Kongsberg sykehus, Drammen sykehus, Ringerike sykehus og Hallingdal sjukestugu (Vestre Viken, 2024). Dette medførte til økt interesse for å undersøke KI-teknologiens prestasjon, og om det er hensiktsmessig å implementere denne teknologien på bildediagnostisk avdeling.

1.2 Formål med oppgaven

I denne oppgaven sammenliknes sensitiviteten og spesifisiteten til kunstig intelligens og leger som tolker bilder. Formålet er å undersøke potensialet til KI-algoritmene ved granskning av røntgenbilder for fraktur i scaphoid og distale radius.

1.3 Forskningsspørsmål

Med utgangspunkt i temaet har vi kommet frem til følgende forskningsspørsmål:

Kunstig intelligens versus bildetolkende leger: kan KI oppnå like høy sensitivitet og spesifisitet som en bildetolkende lege ved granskning av distale radius- og scaphoidfrakturer?

For å utdype forskningsspørsmålet, skal det undersøkes om KI har evne til å oppdage frakturer i røntgenbilder på lik linje med legene som til vanlig tolker disse bildene. Bildetolkende leger er vår fellesbetegnelse for leger som gransker bilder. Dette inkluderer radiologer, ortopeder, kirurger og en akuttmottakslege, ettersom studiene benyttet i denne oppgaven har sammenliknet KI med én eller flere av disse yrkesgruppene. I tillegg har vi undersøkt andre faktorer som kan påvirke diagnostiseringsevnen, blant annet tilgang til kliniske opplysninger, antall plan avbildet, samt yrke og erfaring.

1.4 Radiograffaglig relevans

Temaet og forskningsspørsmålet er av radiograffaglig relevans, da KI innen frakturdeteksjon er ny teknologi som enda er i utprøvelsesstadiet, og vil på sikt bli implementert på bildediagnostiske avdelinger (Vestre Viken, 2024). Denne teknologien vil derfor fortløpende bli en større del av radiografers hverdag (Abilgaard *et al.*, 2018). Ved spørsmål om fraktur er det vanlig at radiografen får svar av legen som gransker røntgenbildene, enten direkte eller gjennom journalsystemer. Dersom fraktur-detekterende KI innføres på bildediagnostisk avdeling, vil svaret potensielt kunne komme direkte fra algoritmen, og dermed føre til at radiografene forholder seg mer til KI enn de bildetolkende legene.

1.5 Avgrensning

For å avgrense oppgaven har vi valgt å se på faktorene sensitivitet og spesifisitet til kunstig intelligens og bildetolkende leger, ved deteksjon av fraktur i distale radius og scaphoid. Kunstig intelligens kan brukes til å detektere frakturer i alle skjelettstrukturene i kroppen. Oppgaven er derfor avgrenset til et par strukturer i samme anatomiske område, for å få sammenliknbare resultater som kan settes opp mot hverandre. Vi har også valgt å inkludere studier som kun tar for seg voksne pasienter, det vil si personer over 18 år. Årsaken til dette er fordi det var få studier som hadde testet KI-algoritmene sin evne til å detektere slike frakturer hos barn. I tillegg har barn vekstplater som i enkelte tilfeller kan etterligne en frakturlinje, og dermed føre til at KI-algoritmen stiller flere falske positive diagnoser (Suzuki *et al.*, 2022). Dette kunne derfor ha ført til store variasjoner i resultatene, og gjort det vanskelig å sammenlikne evnen til KI og de bildetolkende legene. Flere av studiene tok for seg faktorer som «accuracy», «AUC», positive prediktive verdier (PPV) og negative prediktive verdier (NPV). Disse faktorene ble ekskludert grunnet oppgavens omfang.

1.6 Oppgavens oppbygging

Denne bacheloroppgaven er delt inn i flere ulike kapitler med underkapitler. Teorikapittelet presenterer informasjon som belyser sentrale deler av oppgaven, slik at man får en helhetlig forståelse. Videre tar metodekapittelet for seg litteraturstudie som valgt metode, søkeprosessen, innsamling og utvelgelse av data, i tillegg til benyttet analysemetode. I resultatkapittelet blir hovedfunnene fra de utvalgte studiene presentert i oversiktlige tabeller, før de sammenfattes. I diskusjonsdelen blir funnene diskutert opp mot forskningsspørsmålet og teorien som er presentert, i tillegg til metodekritikk av egen studie. Avslutningsvis fremlegges det en konklusjon.

2.0 Teori

I dette kapittelet presenteres relevant teori for denne oppgaven. Her redegjøres hva kunstig intelligens er, de ulike retningene innenfor KI, kunstig intelligens innenfor radiografi, for så å gå mer i dybden på KI innen konvensjonell røntgen. Videre forklares også begrepene sensitivitet og spesifisitet. Til slutt presenteres litt bakgrunnsinformasjon om frakturer i distale radius og scaphoid.

2.1 Hva er kunstig intelligens?

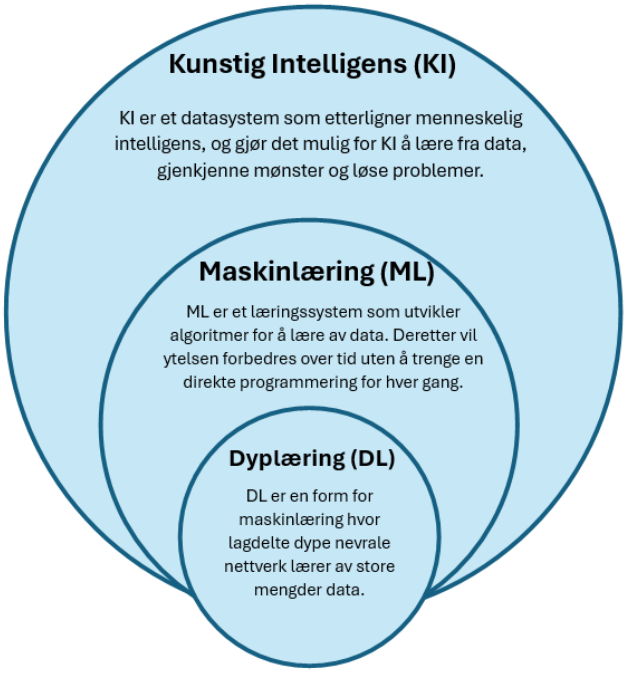
I 1956 var forskere innen automatteori, nevrale nettverk og studie av intelligens, samlet på Dartmouth College for et arbeidsseminar (Tørresen, 2013, s. 12). Her ble det presentert ulike modeller av det nevrale nettverket til hjernen. I ettertid har dette seminaret blitt sett på som grunnleggelsen av begrepet og fagfeltet kunstig intelligens, da det førte til samarbeid mellom ulike sentrale forskere (Tørresen, 2013, s. 12-13).

Kunstig intelligens (KI) er et begrep som refererer til systemer hvor oppgaver blir utført, enten digitalt av applikasjoner eller fysisk ved robotikk. Utførelsen av oppgavene tar utgangspunkt i tolkning og bearbeiding av store data- og informasjonsmengder. Slike systemer baserer seg på at KI har muligheten til å tilpasse seg, ved å nøye analysere og ta hensyn til hvordan effekten av tidligere handlinger har på omgivelsene og beslutningene. Systemet får oppgitt et sluttresultat, hvor det selv må finne den mest effektive løsningen for å nå målet, innenfor gitte kriterier (European commission, 2018, s. 1).

2.1.1 Ulike retninger innen kunstig intelligens

KI-baserte dataprogrammer kan opptre etter fastsatte regler, slik at oppgavene løses på en identisk måte hver gang. Derimot finnes det også avanserte programmer som har egenskapen til å lære. Disse vil tilpasse seg slik at utførelsen av oppgaver kan bli bedret for hver gang (Abilgaard *et al.*, 2018). Et sentralt begrep innenfor KI er «maskinlære». Dette baserer seg på at maskinene lærer selv, slik at de ikke trenger detaljerte instrukser for hvordan de skal løse oppgaven (Barua, 2023, s. 37-38). Se figur 1 for en visuell oversikt over underkategoriene til kunstig intelligens.

Maskinlære er som regel sammensatt av nevrale nettverk, som etterlikner strukturen til nevronettverket i den menneskelige hjerne. Det enkelte nevron utfører en grunnleggende beregning av signalene som kommer inn. Gjennom flere ulike kanaler vil signalene prosesseres og derav avgi en respons. Her sendes det et signal videre til alle nevronene den har en kobling til. Slik som i hjernen vår, vil KI koble sammen flere enkle prosesseringsenheter i et nettverk. Dette gjelder spesielt metoden dyplæring, også kjent som «deep learning». Her bygges det et nevralt nettverk med mange lag. Hvert lag lærer seg en annerledes og unik måte å tolke informasjonen på, og for hvert lag som legges til, vil systemets beregningskraft øke (Abilgaard *et al.*, 2018). Det er flere ulike typer algoritmer innenfor maskinlæring. Konvensjonelle nevrale nettverk (CNN) er dyplæringsalgoritmen som oftest brukes ved analyse av bilder (Erickson *et al.*, 2017). Når et dyplæringsnettverk skal forsøke å identifisere et objekt i et bilde, vil de nedre lagene gjenkjenne grunnleggende egenskaper som kanter, mens de øvre lagene legger merke til mer komplekse strukturer som for eksempel et ansikt (Abilgaard *et al.*, 2018).



Figur 1: Sammenheng mellom Kunstig Intelligens, Maskinlæring og dyplæring. Bearbeidet fra *What is artificial intelligence (AI)?,* av Interaction Design Foundation, 2016, Interaction Design Foundation (https://www.interaction-design.org/literature/topics/ai). CC BY – SA 4.0

2.2 Kunstig intelligens innenfor radiologi

Innenfor radiologi vil det nevrale nettverket kunne læres opp til å oppdage abnormaliteter i et radiologisk bilde, og deretter foreslå forskjellige diagnoser. I studier som er gjort til nå, er KI innenfor radiologi begrenset og ikke mye brukt i det kliniske. Det er derimot noe som er under utvikling, og kan etter hvert bli brukt til mer avanserte oppgaver (Gore, 2020; Hardy & Harvey, 2020). For at kunstig intelligens skal kunne brukes til å stille diagnoser, må det trenes opp ved å analysere store mengder bilder hvor det er en gitt fasit. Treningen starter med at nettverket bruker tilfeldige filtre og utgangsverdier, hvor disse justeres dersom diagnosen blir feil. Når algoritmene er optimalisert, testes de på helt nye bilder for å sikre at nettverket fungerer slik det er ment.

Forskning viser at granskningsarbeidet til bildetolkende leger i framtiden ikke vil bli erstattet av kunstig intelligens, men at algoritmen kan være et supplement når det kommer til kvalitetsforbedring (Hardy & Harvey, 2020). For eksempel har radiologene andre oppgaver i tillegg til å beskrive bilder, som intervensjon, kommunikasjon, undervisning og forskning, som kunstig intelligens ikke kan overta (Abilgaard *et al.*, 2018).

2.2.1 Kunstig intelligens innen konvensjonell røntgen i dag

Dyplæring har ifølge Gan *et al.* (2019) blitt gradvis mer implementert innen bildediagnostikk siden 2012. Det er et økende antall studier som undersøker bruken av dyplæring innen medisinsk bildeanalyse, til å diagnostisere blant annet lungetuberkulose og lungefortetninger på røntgen thorax (Gan *et al.*, 2019). I tillegg ble det i 2009 utviklet en programvare kalt BoneXpert, som analyserer alle skjelettstrukturene i hånd og håndledd, og deretter kalkulerer en alder for hver enkelt struktur. Denne programvaren hadde en hensikt med å kunne erstatte radiologen fullstendig (Thodberg *et al.*, 2022). I en studie om bruken av BoneXpert gjennomført av Thodberg *et al.* (2022), er det inkludert 18 land fra Europa, noe som tyder på at programvaren er godt integrert i denne verdensdelen. I en rapport fra Helse Sør-Øst fra 2021, står det at det er seks helseforetak i Norge som benytter seg av BoneXpert (Helsedirektoratet, 2021, s. 6).

De siste årene har dyplæring også blitt stadig mer implementert når det kommer til deteksjon av brudd, og antallet studier på temaet har økt betraktelig siden 2016 (Meena &

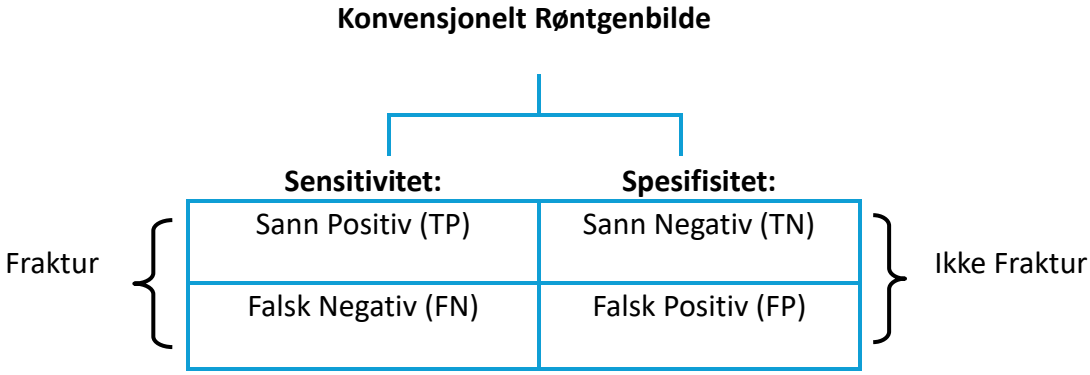
Roy, 2022). Det er flere programvarer som i dag blir bruk til å detektere brudd (Hayashi *et al.*, 2022). BoneView er en av mange slike KI-algoritmer, og er godkjent i både USA og Europa, ettersom den er FDA-godkjent og CE-merket. Den er tatt i bruk i 30 land, og på over 800 institusjoner (Gleamer, u.å.) I tillegg er andre KI-algoritmer brukt i både Europa, Nord- og Sør-Amerika, Asia og Australia, ifølge en metaanalyse om KI og frakturdeteksjon gjennomført av Jung *et al.* (2024). Dette indikerer at KI innen deteksjon av frakturer er på god vei til å bli utbredt i ulike verdensdeler.

2.3 Sensitivitet og spesifisitet

Innen diagnostisering av frakturer, sykdommer og andre tilstander, har medisinsk avbildning en sentral rolle. Måleenhetene som oftest brukes for å vurdere diagnostiseringsevnen, er sensitivitet (Se) og spesifisitet (Sp). Disse målingene er med på å teste egenskapen til å kunne korrekt og nøyaktig identifisere om pasienten har fraktur eller ikke har fraktur (Filleron, 2017).

Sensitivitet defineres som sannsynligheten for at en diagnostisk test er positiv hos en pasient med en skade eller sykdom (Drake & Levine, 2005). Dette handler derfor om de bildetolkende legene og KI sin evne til å kunne oppdage en fraktur hos pasienter som har dette (Shreffler & Huecker, 2023). Resultatet defineres dermed som sann positiv (TP). Om det blir vurdert til at pasienten ikke har en fraktur, men faktisk har det, er resultatet falskt negativt (FN) (Morgan *et al.*, 2015).

Spesifisitet defineres derimot som sannsynligheten for at en diagnostisk test er negativ hos en pasient uten skade eller en sykdom (Drake & Levine, 2005). Dette kan forklares som de bildetolkene legene og KI sin evne til å vurdere at pasienten ikke har noen fraktur (Shreffler & Huecker, 2023). Et slikt resultat er derfor sann negativ (TN). Dersom det vurderes til at pasienten har fraktur, men pasienten faktisk ikke har dette, er resultatet falsk positivt (FP) (Morgan *et al.*, 2015).



Figur 2: Sammenheng mellom sensitivitet og spesifisitet ved TP, FN, FP og TN. Bearbeidet fra *Derivations of sensitivity and specificity* av Matt A. Morgan, 2015, Radiopaedia.org (https://radiopaedia.org/cases/receiver-operating-characteristic-roc-curve). CC BY-NC 4.0 DEED

2.4 Generelt om frakturer i hånd og håndledd

Frakturer i hånd- og håndledd er uten tvil en av de vanligere skadene som avbildes på røntgen (Hamblen & Simpson, 2007, s. 191). Konvensjonell røntgen er den foretrukne modaliteten ved diagnostisering av slike frakturer, ettersom det er raskt og kostnadseffektivt. I tillegg gir det en lavere stråledose sammenliknet med CT (Suzuki *et al.*, 2022; Zhang *et al.*, 2023).

2.4.1 Distal radiusfraktur

Distale radiusfrakturer (DRF) utgjør ca. 20% av alle brudd (Matre & Hole, 2015, s. 164; Suzuki *et al.*, 2022; Zhang *et al.*, 2023), hvor 15 000 nordmenn får en slik fraktur hvert år (Matre & Hole, 2015, s. 164). Ved klinisk mistanke om DRF, henvises pasienten til røntgen der det tas AP- og laterale projeksjoner. Eventuell behandling vil baseres på funn fra røntgenbildene og bruddets klassifisering. Som regel behandles DRF enten konservativt med gips, eller operativ med innsettelse av osteosyntesemateriale (Metodebok, 2023). Etter operasjon og/eller gipsing skal nye røntgenbilder tas i begge plan, for å kvalitetssikre reponeringen (Hamblen & Simpson, 2007, s. 181). Flesteparten av pasientene oppnår god tilheling, og gjenvinner normal funksjon av hånden. Sett i forhold til antallet som får påvist håndleddsfrakturer er det relativt få som opplever komplikasjoner, men eldre er mer utsatte sammenliknet med andre aldersgrupper. Ukorrekt tilheling av frakturen kan føre til deformitet av armen, som igjen kan

hemme funksjonaliteten (Hamblen & Simpson, 2007, s. 183). Sekundærtilstander som artrose, kan oppstå etter en DRF og føre til redusert funksjon i håndleddet. Dette kan dermed påvirke pasientens selvstendighet og deres livskvalitet (Suzuki *et al.*, 2022; Zhang *et al.*, 2023). For å forebygge dette, er det viktig at frakturen oppdages tidlig og immobiliseres i korrekt posisjon, innen de to første ukene fra bruddet oppsto (Hamblen & Simpson, 2007, s. 181).

2.4.2 Scaphoidfraktur

Forekomst av scaphoidfraktur i håndroten er mer sjeldent, og opptrer oftest som en del av en mer omfattende skade i hånden. Disse frakturene er vanligst hos yngre voksne, og rammer i to av tre tilfeller unge menn. Scaphoidfrakturer utgjør nærmere 60% av alle håndrotsfrakturene og ca. 11% av frakturene i hånden (Matre & Hole, 2015, s. 186). Frakturer i scaphoidbeinet blir ofte oversett, enten grunnet et manglende røntgenbilde av området, eller at bruddet ikke ble oppdaget på røntgenbildet. Dette kan skyldes at symptomene ofte er utydelige og ikke fremtredende, som gjør det vanskelig for legen å detektere abnormalitet basert på de kliniske funnene (Hamblen & Simpson, 2007, s. 192-193). Ved diagnostisering gjennomføres røntgen med scaphoidprojeksjoner, som innebærer AP-, lateral-, og skråprojeksjon med ulnar deviasjon. Om det er negativt funn på røntgen, men fortsatt mistanke om fraktur, tas det en CT-undersøkelse. Eventuelt legges det på gips, og deretter tas nye røntgenbilder om to uker. Dette er grunnet at det tilkommer callusdannelse rundt frakturen, og den kommer derfor mer til syne (Matre & Hole, 2015, s. 186).

Ubehandlede scaphoidfrakturer kan potensielt være problematiske, og forekomsten av komplikasjoner er høy (Hamblen & Simpson, 2007, s. 195). Det kan føre til osteoartritt med manglende eller forsinket sammenvoksing, avaskulær nekrose, vedvarende håndleddssmerter og funksjonstap (Hamblen & Simpson, 2007, s. 195; Ozkaya *et al.*, 2022). Derfor vil det være essensielt med tidlig diagnostisering, slik at behandling kan settes i gang så fort som mulig for at håndleddet kan opprettholde sin normale fysikk og funksjon (Ozkaya *et al.*, 2022)

3.0 Metode

Til denne oppgaven ble det benyttet en kvalitativ forskningsmetode. Innenfor dette ble en systematisk litteraturstudie tatt i bruk, der tidligere forskning innen temaet danner grunnlaget for besvarelsen av forskningsspørsmålet. Under dette kapitlet forklares hvordan vi har gått frem ved å presentere den valgte metoden, innsamlingen av data, analysemetode, og kritikk til metoden.

3.1 Litteraturstudie

Litteraturstudie er en metode som belyser en problemstilling ved å ta utgangspunkt i forskningslitteratur som allerede eksisterer (Grønseth & Jerpseth, 2019, s. 80). Det vil dermed ikke skapes ny kunnskap, men det danner en oversikt over nyttig forskning innen et fagområde (Støren, 2010, s. 18). En litteraturstudie innebærer å gjennomføre systematiske søk etter litteratur i ulike relevante databaser, deretter kritisk vurdere disse funnene, og til slutt sammenfatte alt skriftlig (Grønseth & Jerpseth, 2019, s. 80). Formålet med denne metoden er å undersøke de utvalgte studiene som en helhet, hvor sammenfatningen har som hensikt å skape en systematisk oversikt (Malterud, 2017, s. 23).

3.1.1 Begrunnelse for valg av metode

Litteraturstudie som metode er mest relevant for å hente informasjon og besvare oppgavens forskningsspørsmål. Til oppgaven egner det seg best å benytte kvantitativ data, for å kunne sammenlikne sensitivitet og spesifisitet hos bildetolkende leger og kunstig intelligens. Tidligere forskning på temaet er begrenset i Norge, og det finnes foreløpig ingen publiserte verdier. Internasjonalt er det mer utbredt, og det er derfor naturlig å benytte datamateriale fra disse landene. Basert på dette, ble det bestemt at litteraturstudie var metoden som egnet seg best for denne oppgaven.

3.1.2 Fremgangsmåte

Til å begynne med ble det avklart at temaet skulle omhandle kunstig intelligens. Deretter ble det gjennomført flere utforskende litteratursøk på databaser som Sciencedirect, Scopus,

PubMed og MEDLINE (Ovid), for å få en oversikt over omfanget av litteratur og forskning rundt valgt tema hos de ulike databasene. Dette dannet utgangspunktet for videre innhenting av informasjon. De utforskende søkene inspirerte til det nåværende forskningsspørsmålet, og var nyttig for videre strukturert søk da nødvendige nøkkelord ble identifisert i litteraturen.

3.2 PICO-skjema

PICO-skjema er en strukturell måte å utarbeide og formulere forskningsspørsmålet. Det benyttes som et verktøy for å presisere spørsmålet før litteratursøket starter. Dette skjemaet er med på å vise til hvem og hva forskningsspørsmålet handler om, avklare tiltakene som skal undersøkes, og hvilke resultater som er interessante. Funksjonen til et PICO-skjema vil også inkludere identifisering og organisering av søkeord, i tillegg til å finne fram inklusjons- og eksklusjonskriterier til litteraturen (Grønseth & Jerpseth, 2019, s. 85). Dette var med på å gi struktur til søkeprosessen, med gode definerte søkeord og klargjøring av aktuelle kriterier til litteraturen.

P	Population/problem	Mistanke om fraktur i distale radius og scaphoid
I	Intervention	Bruk av kunstig intelligens til granskning ved konvensjonell røntgen
C	Comparison	Bildetolkende leger
O	Outcome	Sensitivitet og spesifisitet ved identifisering og utelukking av frakturer

Tabell 3: PICO-skjema

3.3 Søkeprosessen og datautvalg

Søkeprosessen ble gjennomført som et strukturert søk, i perioden 26.februar til 22.april. Dette ble gjort i databasene Pubmed og Scopus, med bruk av søkeordene som ble utformet fra PICO-skjemaet. Søkene fra Scopus hadde flere duplikater, men ble beholdt grunnet noen nye artikler med relevans til oppgaven. Sciencedirect og MEDLINE (Ovid) ble også forsøkt, men de resulterte i svært få treff hvor flertallet også var duplikater. Ingen av de nye artiklene var egnet til å besvare forskningsspørsmålet, ettersom at de ikke oppfylte

inklusjonskriteriene. Artikler fra databasene Sciencedirect og MEDLINE (Ovid) er derfor ikke tatt i bruk.

For å få relevante treff under søket, ble det blant annet benyttet engelske søkeord for å oppnå flest mulige treff i databasene. Resultatene ble deretter avgrenset til engelsk og de skandinaviske språkene for å fjerne artikler av andre språk. I tillegg ble søkene avgrenset til en spesifikk tidsperiode for å sikre at de utvalgte artiklene er faglig oppdaterte. Noen av søkeordene som ble tatt i bruk er «Artificial intelligence, X-ray, Fracture, Wrist og Scaphoid». Disse ble brukt i kombinasjon med «AND» eller «OR» for å spesifisere søket. I tillegg ble det benyttet flere synonymer for å ikke ekskludere relevante artikler. Av den grunn, ble det for eksempel brukt «wrist fracture OR distal radius fracture» under søket.

For å få oversikt over søkeprosessen ble det utformet en tabell over søkehistorikken, som inneholder informasjon om søkestreng, database, antall treff og dato (Vedlegg 1). Antall treff på de ulike databasene er illustrert i et flytskjema (Figur 3).

3.4 Inklusjons- og eksklusjonskriterier

For å passe på at søket ble mer rettet mot det utvalgte temaet og forskningsspørsmålet, samt sørge for at artiklene var av god kvalitet, ble det utformet inklusjons- og eksklusjonskriterier (Grønseth & Jerpseth, 2019, s. 89). Ved bruk av disse kriteriene ble irrelevante artikler utelukket.

Inklusjonskriterier	Eksklusjonskriterier
Artiklene er skrevet på engelsk, norsk, svensk eller dansk.	Artikler skrevet på andre språk.
Artiklene er skrevet mellom 2018 og mars 2024.	Artikler skrevet før 2018.
Artikler som har tilgang til hele artikkelen/fulltekst (uten kostnader eller registrering av bruker).	Artikler som ikke har fullstendig tekst tilgjengelig, finnes i en lukket database eller er bak en betalingsmur.
Artikkelen omhandler kunstig intelligens innenfor konvensjonell røntgen.	Artikler som tar utgangspunkt i andre modaliteter enn konvensjonell røntgen, for eksempel CT, MR og ultralyd.
Artikkelen tar for seg sensitivitet og/eller spesifisitet til KI og bildetolkende leger hver for seg.	Artikler som tar for seg sensitivitet og spesifisitet kun hos KI eller bildetolkende lege. Artikler som tar for seg bildetolkende leger kun i kombinasjon med KI, altså med KI som hjelpemiddel.
Artiklene omhandler håndleddsfrakturer, dette inkluderer distale radius og scaphoid.	Artikler som omhandler andre tilstander enn håndleddsfrakturer, for eksempel osteoporose. Artikler som omhandler radiusfrakturer hos andre arter enn mennesker (eks. hund, hest, katt).
Artikkelen er fagfellevurdert.	Artikkelen er ikke fagfellevurdert.

Tabell 4: Inklusjons- og eksklusjonskriterier

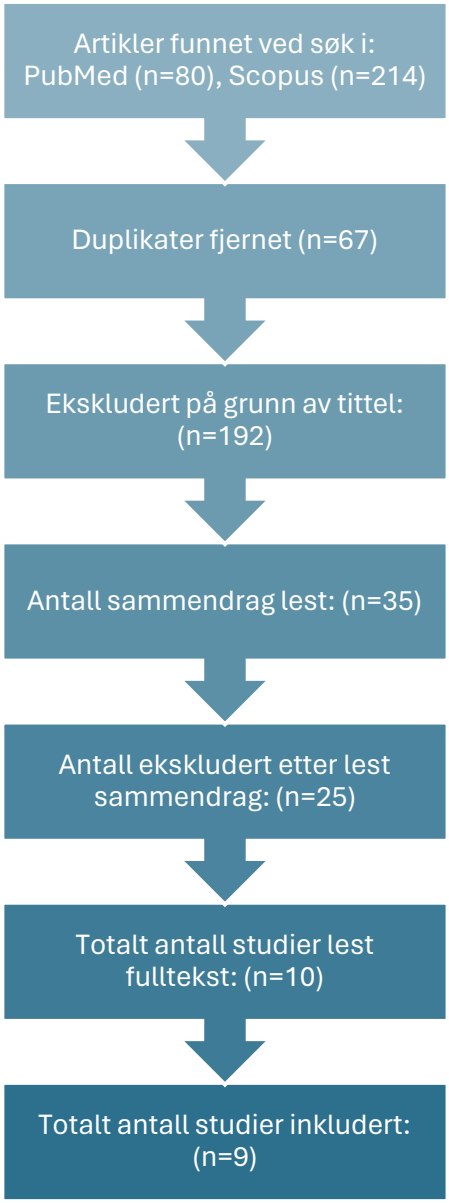
Ettersom at oppgavens tema og forskningsspørsmål er basert på nyere teknologi, hvor det har skjedd stor framgang de siste årene, ble det bestemt at artikler skrevet mellom 2018 og mars 2024 skulle inkluderes. Som følge av denne utviklingen, ble artikler skrevet før dette ekskludert for å sikre at studiene var oppdaterte og mest mulig relevante. Det ble ikke

begrenset at artiklenes opprinnelse skulle være innenfor et visst geografisk område, da det finnes et fåtall publikasjoner innenfor temaet.

Ettersom at innholdet og resultatene i studiene skal tolkes, ble det besluttet at artiklene måtte være skrevet på engelsk eller et av de skandinaviske språkene. De utvalgte studiene sammenlikner sensitiviteten og spesifisiteten mellom KI og bildetolkende leger når det kommer til tolkning av ulike anatomiske strukturer i hånd og håndledd. Flertallet av studiene spesifiserer seg enten innen distale radius eller scaphoid, men det er også et par som omtaler flere anatomiske strukturer i samme område. Derfor vil kun resultatene fra distale radius og scaphoid benyttes, mens de resterende resultatene blir ekskludert. Underveis ble det oppdaget at flere artikler omhandlet det valgte temaet innenfor veterinærmedisin, og tok for seg distal radiusfraktur hos bla. hunder, hester og katter. Publikasjoner knyttet til dette ble derfor ekskludert.

3.5 Flytskjema

For å visualisere hvordan søkeprosessen er gjennomført, ble det tatt i bruk et flytskjema (Figur 3). Dette skjemaet er basert på inklusjons- og eksklusjonskriteriene.



Figur 3: Flytskjema

Flytskjemaet inneholder søk gjennomført i både PubMed og Scopus. Totalt ble det lest 35 sammendrag, hvor 25 av disse studiene ble ekskludert grunnet irrelevans. Videre ble det lest fulltekst til de resterende ti studiene, hvor ni ble inkludert i oppgaven på bakgrunn av inklusjon- og eksklusjonskriteriene. Den ekskluderte studien viste seg å være en metaanalyse som benyttet flere av de allerede inkluderte studiene, og er derfor ikke brukt. Alle de utvalgte studiene er kvantitative studier, og består av syv artikler fra PubMed og to artikler fra Scopus.

3.6 Kvalitetsvurdering av utvalgte studier

For å kvalitetssikre våre utvalgte studier, har vi tatt i bruk en sjekkliste fra Helsebiblioteket (2021). Der var det listet opp en rekke spørsmål, som vi kritisk vurderte opp mot studiene. Spørsmålene i sjekklisten handlet hovedsakelig om studiene hadde et tydelig forskningsspørsmål og passende metode til å besvare denne, samt om resultatene var troverdige og kunne benyttes videre (Helsebiblioteket, 2021). Dersom man svarte nei på selv ett av disse spørsmålene, var det grunnlag nok til å forkaste studien (Helsebiblioteket, 2021). De utvalgte studiene vurderte vi til å oppfylle alle kravene i sjekklisten.

Når en vitenskapelig artikkel skal publiseres, vil den bli vurdert og godkjent av fagpersoner både internt i tidsskriftet, men også av eksterne uavhengige eksperter innenfor fagfeltet (Dalland, 2021, s. 145). Dette skal bidra til å sørge for at artikkelen holder seg til standarden for vitenskapelige artikler, og samtidig skille de fra fagartikler (Dalland, 2021, s. 145). Dermed vurderer vi det til at studiene som er hentet fra databasene er av god faglig kvalitet. For å være på den sikre siden har vi gått gjennom hver av tidsskriftene som de utvalgte studiene er publisert i, og undersøkt at hver av de inkluderte artiklene er fagfellevurdert.

3.7 Analyse

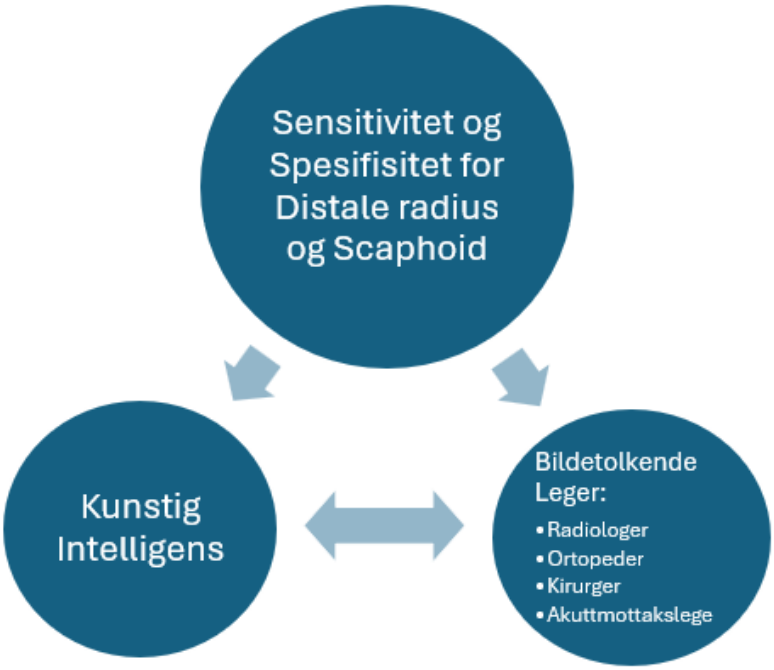
Ifølge Forsberg og Wengström (2016, s. 41), handler analysefasen om hvordan data skal behandles og analyseres. Den har også som mål å forstå, beskrive og tolke dataen. Analysen gjennomføres ved å dele opp informasjonen i mindre deler, hvor innholdet undersøkes hver for seg. Til slutt settes funnene sammen igjen, hvilket danner en ny forståelse (Forsberg & Wengström, 2016, s. 152). Analysen skal kunne svare på spørsmålet oppgaven undersøker. Dette gjennomføres ved å systematisere, ta i bruk sjekklister og kode resultatene opp mot forskningsspørsmålet (Grønseth & Jerpseth, 2019, s. 91).

De inkluderte studiene i oppgaven er analysert ved hjelp av Evans systematiske innholdsanalyse. Denne analysemetoden syntetiserer datamaterialet fra de ulike studiene, ved å systematisere resultatene og deretter sammenfatte de (Evans, 2002). Analyseprosessen til Evans deles inn i fire ulike faser (Evans, 2002, s. 25).

Den første fasen handler om å samle inn data. I denne fasen avklarte vi hvilke studier som skulle brukes i oppgaven. For å finne relevante studier ble det utformet inklusjons- og eksklusjonskriterier. Vi gjennomførte søk i ulike databaser, hvor kriteriene ble benyttet for å få et homogent utvalg (Evans, 2002, s. 25).

Fase to tar utgangspunkt i å identifisere hovedfunn i studiene. For å få en helhetlig forståelse, leste vi gjennom de utvalgte studiene i flere omganger. Hovedfunnene innen sensitiviteten og spesifisiteten til KI og bildetolkende leger, ble markert og notert under lesing. På denne måten ble den relevante informasjonen identifisert (Evans, 2002, s. 25).

I fase tre er hovedmålet å finne felles temaer på tvers av studiene. Vi identifiserte likheter og forskjeller mellom studiene, for å finne gjennomgående temaer. Hovedtemaene som ble utvalgt, er sensitivitet og spesifisitet for distale radius og scaphoid. I tillegg ble det også dannet subtemaer som underbygger hovedtemaene (Evans, 2002, s. 25). Subtemaene er KI og de ulike yrkesgruppene som faller inn under bildetolkende leger. Hovedtemaene og deres subtemaer er illustrert i figur 4.



Figur 4: Illustrasjon av hovedtemaene og subtemaene

I den fjerde og siste fasen samlet vi funnene, som er beskrevet i fase tre, i oversiktlige tabeller (Tabell 14 og 15). Disse brukte vi for å danne en beskrivelse av hovedtemaene og subtemaene. Funnene sammenliknet vi med eksempler fra de opprinnelige studiene, for å sørge for at beskrivelsen korresponderer med originalen.

3.8 Etiske implikasjoner

En litteraturstudie benytter eksisterende forskning for å besvare oppgaven. Etikken er derfor ikke like sentralt, slik det er ved for eksempel intervju eller spørreundersøkelser. Det er likevel viktig å passe på at de utvalgte studiene til oppgaven er gjennomført på en etisk forsvarlig måte, da disse mest sannsynlig har innhentet personopplysninger. De kan enten ha hentet inn opplysninger direkte som for eksempel personnummer, eller indirekte som kjønn, alder og sted. Eventuelt kan opplysningene være hentet via en koblingsnøkkel med et løpenummer, som samsvarer med en navneliste (Dalland, 2021, s. 169). Det kan dermed være lurt å se på om studiene er godkjent av en etisk komité, som vurderer om studiene overholder de etiske retningslinjene. Eventuelt kan det undersøkes om forskerne har skrevet noe om hvordan de har tatt etikken i betraktning (Dalland, 2021, s. 170). Alle de inkluderte studiene i denne oppgaven er godkjent av en etisk komité. Til slutt er en annen viktig etisk betraktning, at vi krediterer forfatterne underveis og ikke selv tar æren for deres arbeid.

4.0 Resultater

Kapittelet tar for seg resultatene og hovedfunnene til de utvalgte studiene, som er fremstilt i litteraturmatriser under kapittel 4.1. Denne matrisen er tatt i bruk grunnet en ryddigere og mer oversiktlig presentasjon av informasjonen til de ni ulike studiene. Videre sammenfattes resultatene i kapittel 4.2. Her har vi valgt å fremstille resultatene i to ulike tabeller (Tabell 14 og 15). Disse presenteres i egne delkapitler under sammenfatningen. I tillegg har vi valgt å beskrive feildiagnoser i sammenfatningen, ettersom dette påvirker sensitiviteten og spesifisiteten.

4.1 Utvalgte studier

Studie 1	Deep learning assisted diagnosis system: improving the diagnostic accuracy of distal radius fractures
Forfatter(e)	Zhang *et al.* (2023)
Tidsskrift	Frontiers in medicine vol.10 (2023)
Land	Kina
Formål	Utforske intelligent deteksjonsteknologi basert på dyplæringsalgoritmer for å hjelpe den kliniske diagnosen av distal radiusfraktur (DRFs), og videre sammenlikne den med menneskelig evne.
Nøkkelbegrep	Artificial intelligence, deep learning, distal radius fractures, computer-assisted diagnosis, elderly population groups
Studiedesign	Studiedesign ikke oppgitt. Studien inkluderte 3240 pasienter (1620 med fraktur og 1620 uten fraktur). Dette resulterte i 3276 AP og 3260 laterale bilder. KI-algoritmen sammenliknes med 3 ortopeder og 3 radiologer, hvorav alle har minst 3 års erfaring. Legene har ikke hatt tilgang til kliniske opplysninger.
Resultat	KI-algoritmen hadde en sensitivitet på 95.70% og en spesifisitet på 98.37%. Gjennomsnittet til ortopedene var en sensitivitet på 91.94% og en spesifisitet på 95.44%, mens gjennomsnittet til radiologene var en sensitivitet på 90.44% og en spesifisitet på 94.62%.
Konklusjon	Studien konkluderer med at modellen hadde utmerket evne til å oppdage distale radiusfrakturer (DRF) på konvensjonelle røntgenbilder. Ved å bruke KI-algoritmen som en sekundær ekspert for å assistere klinisk diagnose, forventes det å forbedre nøyaktigheten (sensitivitet/spesifisitet) ved diagnostisering av DRFs, og forbedre klinisk arbeidseffektivitet ved å redusere arbeidsbelastningen til bildetolkende leger.
Relevans	Studien er relevant for oppgaven på grunn av at den tar for seg sensitivitet og spesifisitet til både KI og bildetolkende leger hver for seg, når det kommer til å detektere fraktur i distale radius.

Tabell 5: Studie 1

Studie 2	Detection and localization of distal radius fractures: Deep learning system versus radiologist
Forfatter(e)	Blüthgen *et al.* (2020)
Tidsskrift	European Journal of Radiology vol.126 (2020)
Land	Sveits
Formål	Evaluere evnen til et dyplæringsprogram som analyserer bilder for å detektere og lokalisere frakturer på distale radius, og deretter sammenlikne det med radiologers evne.
Nøkkelbegrep	Artificial intelligence, Deep learning, Fracture detection, Musculoskeletal radiology, Radiographs
Studiedesign	Et enkelt-senter, retrospektiv kohortstudie. Studien inkluderte 258 pasienter, og hadde 624 bilder i ulike projeksjoner (208 med fraktur og 416 uten). I tillegg ble det benyttet en ekstern database (MURA) med 200 AP og laterale bilder av 100 pasienter (50 med fraktur og 50 uten). To modeller av KI-algoritmen sammenliknes med to radiologer og en 2. års lege i spesialisering (LIS). Radiologene har 16 og 7 års erfaring.
Resultat	*Internt datasettet*: Sensitiviteten til modell 1 var 81% og modell 2 hadde 90%. Radiolog 1 og 2 fikk 86%, og radiolog 3 hadde 90%. Ved spesifisitet fikk den første modellen 100% og den andre 97%. Radiolog 1 hadde et nivå på 86%, radiolog 2 på 97% og radiolog 3 fikk 62%. *Eksternt datasettet fra MURA:* Sensitiviteten til modell 1 var 80% og modell 2 fikk 82%. Radiolog 1, 2 og 3 hadde 98%. Ved spesifisitet fikk den første modellen 86% og den andre 78%. Radiolog 1 hadde et nivå på 94%, radiolog 2 på 90% og til slutt radiolog 3 med 76%.
Konklusjon	Studien konkluderte med at KI har evnen til å detektere og lokalisere distale radiusfrakturer med høy Se og Sp, selv med et lite datasett. KI detekterte/lokaliserte DRFs på lik linje med mindre erfarne radiologer.
Relevans	Studien er relevant da den tar utgangspunkt i sensitivitet og spesifisitet for KI og bildetolkende leger hver for seg, ved deteksjon av DRF.

Tabell 6: Studie 2

Studie 3	Detecting Distal Radial Fractures from Wrist Radiographs Using a Deep Convolutional Neural Network with an Accuracy Comparable to Hand Orthopedic Surgeons
Forfatter(e)	Suzuki *et al.* (2022)
Tidsskrift	Journal of Digital Imaging vol. 35 (2022) s. 39-46
Land	Japan
Formål	Evaluere evnen til CNN når det kommer til å diagnostisere distale radiusfrakturer, ved å bruke røntgenbilder i frontal- og lateralplan. Dette ble deretter sammenliknet med håndortopedenes evne til å diagnostisere frakturene.
Nøkkelbegrep	Distal radial fractures, Convolutional neural network, Deep learning, Radiograph
Studiedesign	Studiedesign ikke oppgitt. Studien inkluderte 792 pasienter, som ga 1633 bilder i ulike projeksjoner (503 med fraktur og 289 uten). KI-algoritmen sammenliknes med 3 håndortopeder, som har 7, 8 og 10 års erfaring. Legene har ikke hatt tilgang til kliniske opplysninger.
Resultat	CNN-modellen hadde en sensitivitet på 98.7% og en spesifisitet på 100%. Håndortoped 1, 2 og 3 hadde en lik sensitivitet på 96%. Når det kommer til spesifisitet, hadde håndortoped 1 en spesifisitet på 98.7%, håndortoped 2 hadde 93.3%, og håndortoped 3 hadde 97.3%.
Konklusjon	Studien konkluderte med at KI-algoritmen hadde evne til å diagnostisere distale radiusfrakturer på lik linje med eller bedre enn de tre håndortopedene, under de samme forholdene.
Relevans	Studien er relevant for oppgaven ettersom den omhandler sensitivitet og spesifisitet hos KI og ortopediske håndkirurger alene, ved granskning av distal radiusfraktur.

Tabell 7: Studie 3

Studie 4	Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments
Forfatter(e)	Gan *et al.* (2019)
Tidsskrift	Acta Orthopaedica vol.90(4). (2019) s. 394-400
Land	Kina
Formål	Evaluere evnen til CNN med en «fast object detection alogritm» som identifiserer en «Region of Interest» (ROI) for å oppdage DFRs på AP håndleddsbilder. Deretter sammenliknes dette med radiologer og ortopeder.
Nøkkelbegrep	Ingen oppgitt
Studiedesign	Studiedesign ikke oppgitt. Studien inkluderte 2340 pasienter (1491 med fraktur og 849 uten fraktur). Dette ga 2340 AP bilder, ettersom at studien ikke benyttet andre projeksjoner KI-algoritmen sammenliknes med 3 ortopeder og 3 radiologer, hvorav ortopedene har mer enn 5 års erfaring, og radiologene har minst 3 års erfaring.
Resultat	Inception-v4 modellen hadde en sensitivitet på 90% og en spesifisitet på 96%. Ortopedene hadde en sensitivitet på 93% og en spesifisitet på 95%, mens radiologene hadde en sensitivitet på 81% og en spesifisitet på 87%.
Konklusjon	Studien konkluderte med at modellen hadde en diagnostisk evne som tilsvarte ortopedene sin, og var bedre enn radiologene på å skille AP håndleddsbilder med og uten DFRs.
Relevans	Studien er relevant for oppgaven fordi den ser på sensitiviteten og spesifisiteten ved deteksjon av distale radiusfrakturer, både for KI og bildetolkende leger hver for seg.

Tabell 8: Studie 4

Studie 5	Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs
Forfatter(e)	Cohen *et al.* (2023)
Tidsskrift	European Radiology vol. 33. (2023) s. 3974-3983
Land	Frankrike
Formål	Sammenlikne ytelsen av KI opp mot radiologer, ved deteksjon av håndleddsfrakturer på røntgenbilder.
Nøkkelbegrep	Artificial intelligence, Fractures, Bones, Radiography, Wrist
Studiedesign	Et enkelt-senter, retrospektiv, diagnostisk studie. Studien inkluderte 637 pasienter, som ga 1917 bilder med ulike projeksjoner fra forskjellige anatomiske områder. Av disse var det 166 radiusfrakturer og 25 scaphoidfrakturer. KI-algoritmen sammenliknes med IRR (Initial radiology reports), hvilket ble laget av 41 radiologer, hvorav 29 LIS leger på sitt 4./5.år, 8 radiologer og 4 overleger. Legene har ikke hatt tilgang til kliniske opplysninger.
Resultat	Ved deteksjon av radiusfraktur, hadde KI-algoritmen «BoneView» en sensitivitet på 89%, mens radiologene hadde en sensitivitet på 83%. Ved deteksjon av scaphoidfraktur, hadde KI-algoritmen en sensitivitet på 84%, mens radiologene hadde en sensitivitet på 80%. Både KI-algoritmen og radiologene hadde en spesifisitet på 96%, men disse verdiene representerer gjennomsnittet av alle de anatomiske områdene.
Konklusjon	Studien konkluderte med at ytelsen av kunstig intelligens ved deteksjon av håndleddsfraktur på røntgenbilder, er bedre enn ikke-spesialiserte radiologer. En kombinasjonsanalyse som inkluderte KI og radiolog sammen, ga den best ytelsen.
Relevans	Studien er relevant for oppgaven på grunn av at den inneholder sensitivitet for både KI og radiologer hver for seg, ved deteksjon av fraktur i distale radius og scaphoid.

Tabell 9: Studie 5

Studie 6	Evaluation of a convolutional neural network to identify scaphoid fractures on radiographs
Forfatter(e)	Li *et al.* (2023)
Tidsskrift	Journal of Hand Surgery. vol.48(5), (2023) s. 445-450
Land	Kina
Formål	Avklare om en CNN-modell kan oppnå tilsvarende nivå som eksperter ved identifisering av scaphoidfrakturer.
Nøkkelbegrep	Scaphoid fracture, convolutional neural network, computer assisted diagnosis, radiography
Studiedesign	Studiedesign ikke oppgitt. Studien inkluderte 600 pasienter, hvor 1139 bilder i AP-projeksjon og scaphoidserie ble benyttet. KI-algoritmen sammenliknes med 4 håndkirurger, som har henholdsvis 3, 3, 13 og 14 års arbeidserfaring.
Resultat	CNN-modellen hadde en sensitivitet på 82%, og en spesifisitet på 94%. Håndkirurgene hadde til sammen en sensitivitet på 76%, og en spesifisitet på 96%.
Konklusjon	Studien konkluderer med at CNN sin evne til å identifisere scaphoidfrakturer tilsvarte evnen hos flertallet av kirurgene.
Relevans	Studien er relevant for oppgaven da den tar for seg sensitivitet og spesifisitet hos KI og håndkirurger ved deteksjon av scaphoidfraktur.

Tabell 10: Studie 6

Studie 7	Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography
Forfatter(e)	Ozkaya *et al.* (2022)
Tidsskrift	European Journal of Trauma and Emergency Surgery. Vol. 48 (2022) s. 585-592
Land	Tyrkia
Formål	Fastslå den diagnostiske ytelsen til kunstig intelligens ved hjelp av CNN, for å oppdage scaphoidfrakturer på AP røntgenbilder av håndledd. Dette ble deretter sammenliknet med en lege på akuttmottaket og to ortopediske spesialister med ulik erfaring.
Nøkkelbegrep	Scaphoid, Fracture, Deep learning, Artificial intelligence, Radiography
Studiedesign	Studiedesign ikke oppgitt. Studien inkluderte 390 pasienter (192 med fraktur og 198 uten fraktur). Dette resulterte i 390 AP bilder, ettersom at studien ikke benyttet andre projeksjoner. KI-algoritmen sammenliknes med en lege i akuttmottaket og 2 ortopeder. Studien deler ortopedene i en erfaren ortoped, og en mindre erfaren ortoped.
Resultat	CNN hadde en sensitivitet på 76% og en spesifisitet på 92%. Akuttmottakslegen hadde en sensitivitet på 62% og en spesifisitet på 90%. Den mindre erfarne ortopeden hadde en sensitivitet på 72% og en spesifisitet på 92%, mens den mer erfarne ortopeden hadde en sensitivitet på 86% og en spesifisitet på 98%.
Konklusjon	Studien konkluderte med at dyplæringsalgoritmen har potensial til å diagnostisere scaphoidfrakturer. I tillegg kan den være et nyttig verktøy for uerfarne ortopeder eller håndkirurger.
Relevans	Studien er relevant for oppgaven ettersom den omhandler sensitivitet og spesifisitet for KI og bildetolkende leger ved deteksjon av fraktur i scaphoid.

Tabell 11: Studie 7

Studie 8	Commercially-available AI algorithm improves radiologists' sensitivity for wrist and hand fracture detection on X-ray, compared to a CT-based ground truth.
Forfatter(e)	Jacques *et al.* (2023)
Tidsskrift	European Radiology (2023)
Land	Frankrike
Formål	Undersøke radiologens ytelse til å detektere frakturer i håndledd og hånd, med hjelp av KI. For å kontrollere, ble resultatene sammenliknet med CT bilder.
Nøkkelbegrep	Artificial intelligence, Wrist fractures, Fractures (multiple), Trauma, Diagnostic imaging
Studiedesign	Et enkelt-senter, retrospektiv, diagnostisk studie. Studien inkluderte 296 pasienter, som ga 788 bilder av hånd og håndledd i ulike projeksjoner. Av disse hadde 118 bilder radiusfraktur, og 35 hadde scaphoidfraktur. Resterende bilder tar for seg andre anatomiske områder, og omtales derfor ikke videre i denne oppgaven. KI-algoritmen sammenliknes med 23 radiologer, med erfaring fra 2-25 år og et gjennomsnitt på 5.6 år. 9 av radiologene var LIS, og 14 var erfarne. Legene har ikke hatt tilgang til kliniske opplysninger.
Resultat	Ved deteksjon av fraktur i radius, hadde KI-algoritmen en sensitivitet på 94.9%, mens radiologene til sammen hadde en sensitivitet på 88.0%. Dersom KI og radiologene kombineres, hadde de en sensitivitet på 91.8%. Ved deteksjon av fraktur i scaphoid hadde KI-algoritmen en sensitivitet på 85.7%, mens radiologene til sammen hadde en sensitivitet på 63.9%. Kombinert hadde de en sensitivitet på 77.0%.
Konklusjon	Studien konkluderte med at KI bedrer sensitiviteten og falske negative svar for radiologen, uten å påvirke spesifisiteten eller falske positive svar.
Relevans	Studien er relevant for oppgaven fordi den inneholder sensitivitet for KI og radiologer hver for seg, både ved deteksjon av radiusfrakturer og scaphoidfrakturer.

Tabell 12: Studie 8

Studie 9	Musculoskeletal radiologist-level performance by using deep learning for detection of scaphoid fractures on conventional multi-view radiographs of hand and wrist
Forfatter(e)	Hendrix *et al.* (2023)
Tidsskrift	European Radiology. Vol 33 (2023) s. 1575-1588
Land	Nederland
Formål	Å vurdere KI-algoritmen opp mot fem erfarne radiologer, ved diagnostisering av scaphoidfrakturer på konvensjonell røntgen med bilder i flere plan.
Nøkkelbegrep	Scaphoid bone, Fractures, Bone, Artificial intelligence, Multicenter study, Clinical decision support system
Studiedesign	Et retrospektivt, multi-senter og eksperimentell studie. Studien har flere datasett, hvorav datasett 4 ble benyttet for frakturdeteksjon systemet. Denne inkluderte 209 pasienter fordelt på 219 cases (65 med fraktur og 154 uten fraktur), hvilke resulterte i 688 bilder i ulike projeksjoner. KI-algoritmen sammenliknes med 5 radiologer, som har henholdsvis 5, 7, 22, 24 og 26 års arbeidserfaring.
Resultat	InceptionV3 hadde en sensitivitet på 72%, og oppnådde en spesifisitet på 93%. Når det gjelder sensitivitet hos radiologene, fikk radiolog 1 og 2 en sensitivitet på 75%. Radiolog 3 og 5 hadde 83% og radiolog 4 oppnådde 80%. Ved spesifisitet hadde radiolog 1 den høyeste på 94%. Radiolog 2 hadde 84%, radiolog 3 fikk 88%, radiolog 4 oppnådde 91% og til slutt hadde radiolog 5 en spesifisitet på 81%.
Konklusjon	Studien konkluderer med at KI-algoritmen er i stand til å oppnå tilsvarende nivå som erfarne radiologer, når det gjelder deteksjon av scaphoidfrakturer.
Relevans	Studien er relevant for oppgaven på grunn av at den inneholder sensitivitet og spesifisitet for både KI og radiologer hver for seg, ved deteksjon av fraktur i scaphoid.

Tabell 13: Studie 9

4.2 Sammenfatning av resultatene

Etter uthenting av hovedfunnene fra de utvalgte studiene, med utgangspunkt i forskningsspørsmålet, ble disse resultatene presentert i to ulike tabeller. Dette gjør resultatene enklere å sammenlikne, og skaper en god oversikt. Tabellene tar for seg sensitivitet og spesifisitet innenfor granskning for fraktur i distale radius og scaphoid, hvor tallene tar utgangspunkt i prestasjonen til KI, radiologer, ortopeder, kirurger og/eller en akuttsmottakslege.

4.2.1 Sensitivitet

	Studie	Anatomisk struktur	KI	Radiologer	Ortopeder	Kirurger	Akuttmott-akslege
1	Zhang *et al.* (2023)	Distale radius	95.7%	90.44%	91.94%	-	-
2	Blüthgen *et al.* (2020)	Distale radius	83.25%*	92.67%*	-	-	-
3	Suzuki *et al.* (2022)	Distale radius	98.7%	-	96%*	-	-
4	Gan *et al.* (2019)	Distale radius	90%	81%	93%	-	-
5	Cohen *et al.* (2023)	Distale radius	89%	83%	-	-	-
		Scaphoid	84%	80%	-	-	
6	Li *et al.* (2023)	Scaphoid	82%	-	-	76%	-
7	Ozkaya *et al.* (2022)	Scaphoid	76%	-	79%*	-	62%
8	Jacques *et al.* (2023)	Distale radius	94.9%	88%	-	-	-
		Scaphoid	85.7%	63.9%	-	-	-
9	Hendrix *et al.* (2023)	Scaphoid	72%	79.2%*	-	-	-

Tabell 14: Sensitivitet ved granskning for fraktur i distale radius og scaphoid hos KI, radiologer, ortopeder, kirurger og akuttmottakslege.

*Verdien er gjennomsnittet av resultatene til den aktuelle studien. For utregning, se vedlegg 2.

Cellene marker med «-» har ikke brukt disse yrkene i studiet.

I tabell 14 presenteres sensitiviteten ved granskning for fraktur i distale radius og scaphoid. Sensitivitet innebærer å kunne stille korrekt diagnose der det foreligger en fraktur.

I studien til Zhang *et al.* (2023) var det KI som oppnådde høyest sensitivitet basert på prosenttallene. KI fikk en sensitivitet på 95.7%, ortopedene hadde 91.94% og radiologene hadde 90.44% ved deteksjon av fraktur i distale radius. Studien konkluderer med at KI-algoritmen presterte bedre enn de bildetolkende legene når det gjelder sensitivitet.

Studien til Blüthgen *et al.* (2020) sammenliknet sensitiviteten mellom KI og radiologer ved deteksjon av distal radiusfraktur. I denne studien var det radiologene som oppnådde høyest resultat med 92.67%, noe som utgjorde nesten 10 prosentpoeng i forskjell fra KI-algoritmen som hadde en sensitivitet på 83.25%. Blüthgen *et al.* (2020) skriver at frakturene som KI ikke greide å detektere, hadde et mer uvanlig utseende enn det algoritmen har trent med.

I Suzuki *et al.* (2022) blir sensitiviteten til KI-algoritmen og ortopeder sammenliknet ved granskning for distal radiusfraktur. KI oppnådde høyest sensitivitet på 98.7%, mens ortopedene hadde 96%. Suzuki *et al.* (2022) skriver at det var tilfeller med falske negative (FN) funn, hos både ortopedene og KI-algoritmen. De oppgir derimot ingen konkrete verdier for antall FN. Studien konkluderer med at KI har lik eller bedre evne til å diagnostisere DRF enn ortopedene, under like forhold.

Studien til Gan *et al.* (2019) tar også for seg sensitivitet hos KI, radiologer og ortopeder ved deteksjon av fraktur i distale radius. I denne studien fikk ortopedene best resultat med en sensitivitet på 93%, deretter KI med et resultat på 90% og til sist radiologene med 81%. Ifølge studiet viser KI-algoritmen en relativ lik diagnostiseringsevne som ortopedene, med utgangspunkt i resultatene. Videre skriver Gan *et al.* (2019) at KI har en overlegen evne til å diagnostisere distale radiusfrakturer sammenliknet med radiologene.

I studien til Cohen *et al.* (2023) er både distale radius og scaphoid omtalt, der sensitiviteten til KI og radiologer undersøkes. Ved granskning av begge anatomiske strukturer, kan det observeres at KI oppnår best resultat. Sensitiviteten for fraktur i distale radius er 89% for KI og 83% hos radiologene. For scaphoid er sensitiviteten hos KI 84% mens den er 80% hos radiologene. Cohen *et al.* (2023) rapporterer at ved de 25 scaphoidfrakturene, hadde radiologene 5 tilfeller av FN, mens KI utgjorde 4 FN. Av de 166 distale radiusfrakturene,

produserte radiologene 27 FN, og KI 17 FN. Studien konkluderer med at KI har bedre evne til å detektere frakturer i håndleddet enn det radiologer i praksis har.

Studien til Li *et al.* (2023) tar for seg sensitivitet til KI og kirurger ved deteksjon av fraktur i scaphoid. Det var en forskjell på 6 prosentpoeng, hvor KI-algoritmen presterte best med en sensitivitet på 82%, og kirurgene 76%. Totalt detekterte KI-algoritmen 41 av 50 pasienter med fraktur (TP). Studien konkluderer med at det ikke var noen signifikant forskjell mellom KI og kirurgene.

I Ozkaya *et al.* (2022), blir sensitiviteten til KI, ortopeder og en akuttmottakslege sammenliknet ved deteksjon av scaphoidfraktur. Ortopedene presterte best med en sensitivitet på 79%, etterfulgt av KI med 76% og akuttmottakslegen som hadde en betydeligere lavere sensitivitet på 62%. Ozkaya *et al.* (2022) rapporter at av totalt 50 røntgenbilder med fraktur, var antall tilfeller av FN 19 for akuttmottakslegen, 14 for den mindre erfarne ortopeden, 12 for KI og 7 for den mer erfarne ortopeden. Studien konkluderte med at det ikke var noen signifikant forskjell mellom KI og den mindre erfarne ortopediske spesialisten, men derimot presterte algoritmen bedre enn akuttmottakslegen.

Jacques *et al.* (2023), er den andre studien som tar for seg deteksjon av fraktur i både distale radius og scaphoid. Denne studien sammenlikner sensitiviteten til KI og radiologer, der det er KI som har de beste resultatene for begge de anatomiske områdene. Ved fraktur i distale radius er sensitiviteten for KI 94.9% mens radiologene har 88%. Sensitiviteten ved scaphoidfrakturer er hos KI 85.7%, mens hos radiologene er den på 63.9%. Studien konkluderer med at algoritmen presterer på lik linje med, eller bedre enn de fleste radiologene i studien.

Til slutt har studien til Hendrix *et al.* (2023) tatt for seg sensitivitet ved granskning av scaphoidfrakturer, og sammenliknet KI-algoritmen med radiologer. Algoritmen hadde en sensitivitet på 72%, mens gjennomsnittet til radiologene ble 79.2%. I studien rapporteres det at KI-algoritmen hadde 29 feildiagnoser, mens radiologene hadde 23. Hendrix *et al.* (2023) skriver at det var 3 frakturer som algoritmen diagnostiserte som FN, men radiologene diagnostiserte riktig. Derimot diagnostiserte også radiologene 3 frakturer som FN, hvor KI diagnostiserte de riktig. Hendrix *et al.* (2023) skriver at granskningsevnen til algoritmen tilsvarte radiologene sin evne.

Sammenlikningene i tabell 14 viser at det er størst forskjell mellom yrkesutøverne og algoritmen i studiene til Jacques *et al.* (2023) og Blüthgen *et al.* (2020). I tillegg er sensitiviteten lavest ved deteksjon for fraktur i scaphoid både hos KI og de bildetolkende legene.

4.2.2 Spesifisitet

	Studie	Anatomisk struktur	KI	Radiologer	Ortopeder	Kirurger	Akuttmott-akslege
1	Zhang *et al.* (2023)	Distale radius	98.37%	94.62%	95.44%	-	-
2	Blüthgen *et al.* (2020)	Distale radius	90.25%*	84.17%*	-	-	-
3	Suzuki *et al.* (2022)	Distale radius	100%	-	96.43%*	-	-
4	Gan *et al.* (2019)	Distale radius	96%	87%	95%	-	-
5	Cohen *et al.* (2023)	Distale radius	Ikke oppgitt	Ikke oppgitt	-	-	-
		Scaphoid	Ikke oppgitt	Ikke oppgitt	-	-	-
6	Li *et al.* (2023)	Scaphoid	94%	-	-	96%	-
7	Ozkaya *et al.* (2022)	Scaphoid	92%	-	95%*	-	90%
8	Jacques *et al.* (2023)	Distale radius	Ikke oppgitt	Ikke oppgitt	-	-	-
		Scaphoid	Ikke oppgitt	Ikke oppgitt	-	-	-
9	Hendrix *et al.* (2023)	Scaphoid	93%	87.6%*	-	-	-

Tabell 15: Spesifisitet ved granskning for fraktur i distale radius og scaphoid hos KI, radiologer, ortopeder, kirurger og akuttmottakslege.

*Verdien er gjennomsnittet av resultatene til den aktuelle studien. For utregning, se vedlegg 2.
Cellene marker med «-» har ikke brukt disse yrkene i studiet.

Tabell 15 presenterer spesifisiteten ved granskning for fraktur i distale radius og scaphoid. Spesifisitet handler om å stille korrekt diagnose der det ikke foreligger en fraktur. Studiene som ikke oppgir isolerte tall for de utvalgte anatomiske områdene, er markert som «ikke oppgitt» i tabellen.

I Zhang *et al.* (2023) sin studie blir spesifisiteten til KI, radiologer og ortopeder undersøkt, ved granskning av distal radiusfraktur. KI-algoritmen oppnådde det høyeste prosenttallet med en spesifisitet på 98.37%. Deretter hadde ortopedene et resultat på 95.44%, som tilsier kun noen få prosentpoeng lavere enn resultatet til KI. Til slutt hadde radiologene den laveste spesifisiteten på 94.62%. På bakgrunn av resultatene i tabellen, konkluderer studien med at KI presterte bedre enn både ortopedene og radiologene når det gjelder spesifisitet.

I studien til Blüthgen *et al.* (2020) blir det tatt for seg spesifisitet til KI og radiologer, ved diagnostisering av distal radiusfraktur. I dette studiet fikk KI det høyeste resultatet for spesifisitet på 90.25%. Deretter fikk radiologene omtrent 6 prosentpoeng lavere enn KI, med en spesifisitet på 84.17%. Blüthgen *et al.* (2020) konkluderer med at dyplæringsbaserte modeller har mulighet til å lokalisere og detektere DRF med høy spesifisitet.

Studien til Suzuki *et al.* (2022) undersøker spesifisitet ved granskning av distal radiusfraktur, og sammenlikner KI med tre ortopeder. KI-algoritmen hadde en spesifisitet på 100%, og ortopedene hadde 96.43%. Suzuki *et al.* (2022) skriver om tilfeller hvor to pasienter ble diagnostisert korrekt av KI, mens radiologene feildiagnostiserte dem som falsk positiv (FP). Studien konkluderer med at forskjellen mellom KI og ortopedene ikke blir beregnet som signifikant.

Ved studiet til Gan *et al.* (2019) ble KI, radiologer og ortopedene sin spesifisitet sammenliknet, ved diagnostisering av distale radiusfrakturer. Her var det KI som oppnådde det høyeste resultatet på 96%. Videre har ortopedene en spesifisitet på 95% og deretter radiologene på 87%. Studien konkluderer med at KI-algoritmen har en ytelse på lik linje med ortopedene, men en bedre ytelse enn radiologene.

Li *et al.* (2023) undersøkte spesifisiteten til KI og kirurger ved granskning av scaphoidfrakturer. KI-algoritmen fikk en spesifisitet på 94%, og kirurgene hadde 96%. I studien rapporteres det at algoritmen detekterte 47 av 50 pasienter som ikke hadde fraktur (TN). Li *et al.* (2023) skriver at KI-algoritmen hadde høy spesifisitet, og at det ikke var en signifikant forskjell mellom deres KI-algoritme og kirurgene.

I studiet til Ozkaya *et al.* (2022) var spesifisiteten for scaphoidfrakturer sammenliknet mellom KI, ortopeder og en akuttmottakslege. Ortopedene hadde høyest spesifisitet med et gjennomsnitt på 95%, etterfulgt av KI-algoritmen med 92%, og til slutt akuttmottakslegen

med den laveste spesifisiteten på 90%. Ozkaya *et al.* (2022) oppgir at deres KI-algoritme diagnostiserte 46 av 50 pasienter korrekt uten fraktur (TN), og 4 tilfeller som positive der det ikke var fraktur (FP).

Hendrix *et al.* (2023) sammenliknet spesifisiteten til KI-algoritmen og radiologer ved granskning for scaphoidfrakturer. KI-algoritmen oppnådde 93%, og radiologene 87.6%. I studien skrives det at 9 røntgenbilder som var diagnostisert til å være FP av radiologene, ble diagnostisert riktig av KI. Radiologene greide derimot å riktig diagnostisere 3 av røntgenbildene som algoritmen hadde klassifisert som FP.

Til slutt valgte vi å ekskludere tallene for spesifisitet ved studiene til Cohen *et al.* (2023) og Jacques *et al.* (2023). Årsaken til dette er fordi studiene kombinerte flere anatomiske strukturer i resultatene for spesifisitet. Det er dermed ikke mulig å få representative verdier for scaphoid og distale radius.

Sammenlikningene i tabell 15 viser størst forskjell mellom algoritmen og yrkesutøverne i studiene til Blüthgen *et al.* (2020), Gan *et al.* (2019) og Hendrix *et al.* (2023). Det kan også observeres at det ikke er noen store forskjeller mellom resultatene for distale radius og scaphoid.

5.0 Diskusjon

Diskusjonskapittelet er organisert i to deler. Først diskuteres de utvalgte studienes resultater, hvor det fokuseres på sensitivitet og spesifisitet, forskjellen mellom distale radius og scaphoid, samt de ulike yrkesgruppene og deres arbeidserfaring. Deretter tar kapittelet for seg metodekritikk. Her vil den valgte metoden, gjennomføring av litteratursøk, innsamling av data og studienes innhold, kritisk vurderes.

5.1 Sensitivitet og spesifisitet

I teorien ble det redegjort at sensitivitet er evnen til å kunne oppdage frakturer der det er en reel fraktur, mens spesifisitet er evnen til å kunne vurdere at det ikke er en fraktur, der det faktisk ikke foreligger en fraktur. I tabell 15 kan det i flere av studiene observeres at KI-algoritmene har høyere resultat på spesifisitet enn de bildetolkende legene. Som sett i tabell 14, er sensitiviteten for både KI og de bildetolkende legene lavere enn det spesifisiteten er i tabell 15. Dette kan tyde på at det er vanskeligere å fastslå om det er en fraktur, enn om det ikke er en fraktur.

Sensitiviteten og spesifisiteten påvirkes av hvor mange tilfeller falske negative og falske positive som oppstår. Flere tilfeller av FN vil senke sensitiviteten, mens flere tilfeller av FP vil senke spesifisiteten. Når det gjelder antallet feildiagnostiseringer som opptrer mellom bildetolkende leger og KI-algoritmene, er det fire studier som tar for seg dette. Suzuki *et al.* (2022) nevner et par tilfeller hvor flertallet av ortopedene hadde feildiagnostisert røntgenbildene, og KI-algoritmen diagnostiserte dem riktig. De nevner også at den rette diagnosen ikke alltid ble satt av verken algoritmen eller ortopedene. I tabell 14 om sensitivitet, kan det ses at KI-algoritmen hadde høyere prosenttall enn ortopedene. Dette betyr at de bildetolkende legene hadde flere tilfeller hvor de feildiagnostiserte som falsk negativ, enn det KI hadde. Når det kommer til spesifisiteten i tabell 15, oppnådde KI i Suzuki *et al.* (2022), et resultat på 100%, noe som tilsier at algoritmen ikke stilte noen falske positive diagnoser. Ortopedene i studien feildiagnostiserte derimot noen frakturer som falske positive, ettersom resultatet for spesifisiteten var 96.43%. Cohen *et al.* (2023) oppgir i deres studie at antallet falske negative diagnoser ved distale radius var 17 hos KI, og 27 for radiologene. Ved scaphoid utgjorde KI 4 falske negative resultater, mens radiologene hadde

5. At KI har færrest tilfeller av FN kommer tydelig fram i tabell 14, ettersom algoritmen oppnår høyere sensitivitet enn radiologene.

Ozkaya *et al.* (2022) rapporterer at 12 av 50 scaphoidfrakturer ble feildiagnostisert som FN av KI-algoritmen. Hos akuttmottakslegen var det 19 tilfeller med FN, den mindre erfarne ortopeden hadde 17 og til slutt hadde den erfarne ortopeden 7. Dette kan observeres av resultatene i tabell 14, ettersom ortopedene til sammen har høyest sensitivitet, etterfulgt av KI og akuttmottakslegen. Selv om det ikke oppgis eksakte tall for falske positive resultater, kan det ses i tabell 15 at det var slike tilfeller hos alle tre gruppene i studien. Til slutt oppgir Hendrix *et al.* (2023) at KI-algoritmen feildiagnostiserte 29 bilder, mens radiologene feildiagnostiserte 23 bilder. Det oppgis ikke eksakte tall for antall falske negative og falske positive. Dersom tallene i tabell 14 undersøkes, kan det oppdages at radiologene hadde høyere sensitivitet enn KI, som tilsier at algoritmen diagnostiserte flere frakturer som falske negative. I tabell 15 har derimot KI-algoritmen høyest spesifisitet, som betyr at radiologene stilte flere falske positive diagnoser. Resultatene av disse fire studiene viser dermed en klar sammenheng mellom antall FN og sensitivitet, og antall FP og spesifisitet.

Det er flere faktorer som kan føre til feildiagnostiseringer, som igjen vil påvirke sensitiviteten og spesifisiteten. For eksempel oppfatter ikke KI-algoritmer bilder på samme måte som mennesker, og er derfor ikke i stand til å se på skjelettet som en struktur, slik bildetolkende leger gjør (Blüthgen *et al.*, 2020). I tilfeller hvor KI-algoritmen er i stand til å korrekt diagnostisere frakturer, men legene ikke, kan det tenkes at frakturene var subtile. En kan derfor undre om det var tilfeldigheter som førte til den korrekte diagnosen, eller om KI-algoritmen har bedre evne til å detektere slike frakturer. I tillegg skriver Blüthgen *et al.* (2020) at frakturer som er falske negative, ofte har et uvanlig utseende, som kan føre til at KI-algoritmen feildiagnostiserer disse frakturene. Samtidig skriver de videre at det er usannsynlig at disse frakturene ville bli oversett av et menneske. Dette kan være en årsak til at KI feildiagnostiserte bilder der legene ikke gjorde det.

Til vanlig er opplysninger om skademekanisme noe som er tilgjengelige for leger som tolker bilder, slik at de bedre kan vurdere sannsynligheten for at det foreligger en fraktur. Ifølge Jacques *et al.* (2023), er det å ha tilgang til denne informasjonen vist å bedre deteksjon av frakturer. I studiene til Zhang *et al.* (2023), Suzuki *et al.* (2022), Jacques *et al.* (2023) og Cohen *et al.* (2023), står det spesifikt at radiologene og ortopedene ikke har fått tilgang til

opplysninger som skademekanisme eller hvor pasienten har vondt. På denne måten stiller legene på lik linje med KI, som heller ikke har disse opplysningene (Zhang et al., 2023). Ved at de bildetolkende legene ikke har tilgang til denne informasjonen, går de dermed inn blindt i forhold til det de vanligvis gjør. Istedenfor å kunne fokusere på kun ett område, er legene nødt til å granske hele bildet nøyere. Det er mulig at legene derfor ser raskere over hver enkelt skjelettstruktur, fremfor å nøyere granske noen få strukturer. Det kan dermed tenkes at det er lettere å overse frakturer som ikke er like åpenbare, noe som kan føre til et økt antall falske negative resultater.

Dersom de overnevnte studiene trekkes frem, det vil si Zhang *et al.* (2023), Suzuki *et al.* (2022), Jacques *et al.* (2023) og Cohen *et al.* (2023), oppdages det i alle fire studiene at algoritmen både har bedre sensitivitet og spesifisitet, der det oppgis, enn de bildetolkende legene. Det kan dermed indikere at mangel på opplysninger for de bildetolkende legene har hatt en innvirkning på evnen til å fastslå om det er eller ikke er fraktur til stede. Samtidig er det vanskelig å si om dette er en sammenheng. De resterende studiene nevner ikke at de kliniske opplysningene var utilgjengelige, og har resultater som varierer om det er KI eller de bildetolkende legene som gjør det best.

En annen ting som kan påvirke sensitiviteten og spesifisiteten er, ifølge studiene til Suzuki *et al.* (2022) og Li *et al.* (2023), antall plan røntgenbildene er tatt i. Dette skyldes at en fraktur kan synes i ett plan, men ikke i det andre. Ved spørsmål om distale radius- og scaphoidfrakturer er det standard å ta røntgenbilder i AP- og lateralprojeksjon, eller en scaphoidserie (Ozkaya *et al.*, 2022). I studiene til Gan *et al.* (2019) og Ozkaya *et al.* (2022), er det derimot kun gransket røntgenbilder i AP-plan. Ozkaya *et al.* (2022) skiller seg ut fra flertallet av de andre studiene når det kommer til sensitivitet, da både KI, ortopeder og akuttmottakslegen har betydeligere lavere resultater. Gan *et al.* (2019) sin studie har derimot ingen betydelige forskjeller i resultatene for sensitivitet hos verken KI, radiologene eller ortopedene, sammenliknet med de andre studiene. Dersom spesifisiteten i disse studiene undersøkes, ses det at begge studiene har resultater som ikke skiller seg betydelig ut fra de andre studiene. Basert på resultatene til disse to studiene, er det vanskelig å fastslå om antall plan har hatt en innvirkning for deteksjon av frakturene.

Som diskutert over kan det være flere ulike årsaker til at sensitiviteten og spesifisiteten påvirkes. Ulike forutsetninger som mangel på klinisk informasjon og tilgang til bilder i flere

plan, kan bidra til å påvirke granskningsevnen. Dette vil igjen påvirke sensitiviteten og spesifisiteten, ved at en økning i antall FN resulterer i lavere sensitivitet, og flere FP fører til lavere spesifisitet.

5.2 Forskjell mellom distale radius og scaphoid

I tabell 14 er det mulig å observere en forskjell innen sensitivitet ved detektering av fraktur i distale radius og scaphoid. Dette gjelder både mellom KI algoritmene og de bildetolkende legene. I tabell 15 er det derimot mindre forskjell i prosenttallene mellom de to anatomiske områdene.

Dersom prosenttallene fra alle studiene sammenliknes, ses det at sensitiviteten ved distale radius for de bildetolkende legene varierer fra 81%-96%, mens ved scaphoid varierer det fra 62%-80%. For KI varierer sensitiviteten ved distale radius fra 83.24%-98.7%, mens ved scaphoid varierer det fra 72%-85.7%. Dermed er prosenttallene hos både KI og de bildetolkende legene i flertallet av studiene lavere ved scaphoid, enn ved distale radius. Ifølge studiet til Cohen *et al.* (2023) er scaphoidbilder vanskeligere å vurdere, selv for erfarne radiologer. Videre skriver de at frakturer i scaphoid er sjeldnere sammenliknet med andre anatomiske områder. Distale radiusfrakturer er vanligere, og derfor er det også økt erfaring ved diagnostisering av dette området. Dette kan forklare forskjellene mellom sensitiviteten ved scaphoid og distale radius. Det kan derfor tenkes at sannsynligheten for å få falske negative resultater vil være høyere enn falske positive, ved granskning for fraktur i scaphoid sammenliknet med distale radius. En økning i falske negative tilfeller, vil redusere sensitiviteten. Dermed kan dette forklare hvorfor det stort sett observeres en lavere sensitivitet for både KI og de bildetolkende legene ved scaphoid, enn distale radius.

Når det kommer til spesifisitet, er det derimot vanskelig å si hvorfor det ikke er like stor forskjell i prosenttallene mellom de to anatomiske områdene. I tabell 15 kan det ses at spesifisiteten for KI ved scaphoid varierer fra 92%-94%, og ved distale radius fra 90.25%-100%. De bildetolkende legene har resultater for scaphoid som varierer fra 87.6%-96%, og for distale radius er forskjellen fra 84.17%-96.43%. En mulig forklaring er at begge parter synes det er lettere å vurdere et bilde som sant negativt ved begge strukturene, enn som sant positivt.

5.3 Yrker og erfaring

I de utvalgte studiene blir det tatt i bruk flere forskjellige bildetolkende leger. Dette inkluderer radiologer, ortopeder, kirurger og en akuttmottakslege. I tillegg har de forskjellige legene ulik års erfaring med å granske bilder for frakturer.

I resultatene er det mulig å observere både større og mindre forskjeller mellom disse yrkesgruppene. I studiene til Zhang *et al.* (2023), Gan *et al.* (2019) og Ozkaya *et al.* (2022) benyttes ortopeder, og radiologer eller en akuttmottakslege. Det kan ses et mønster i disse studiene, hvor ortopedene oppnår høyest resultat både for sensitivitet og spesifisitet, i forhold til den andre legegruppen i hver enkelt studie. I Zhang *et al.* (2023) er det få prosentpoeng som skiller radiologene fra ortopedene, mens i Gan *et al.* (2019) er det litt flere prosentpoeng som skiller de to gruppene. I dette tilfellet kan arbeidserfaringen til de ulike yrkesgruppene være av betydning. Både radiologene og ortopedene i Zhang *et al.* (2023) har minst 3 års erfaring, mens i Gan *et al.* (2019) har radiologene minst 3 års erfaring og ortopedene har mer enn 5 år med erfaring. Dette kan indikere at ortopedene i Gan *et al.* (2019) har noe mer yrkeserfaring enn radiologene, og forklare hvorfor det er større forskjell mellom yrkesgruppene i denne studien enn hos Zhang *et al.* (2023). I tillegg kan det også forklare hvorfor ortopedene i studien til Gan *et al.* (2019) men ikke i Zhang *et al.* (2023), var de som fikk bedre resultat enn KI ved sensitivitet, og tilsvarende likt som KI ved spesifisitet.

Den største forskjellen er derimot sensitiviteten til ortopeden og akuttmottakslegen i studiet til Ozkaya *et al.* (2022), der det er 17 prosentpoeng som skiller disse to gruppene. Denne studien er også den eneste som tar i bruk en akuttmottakslege til å granske røntgenbilder. Dette kan tyde på at denne yrkesgruppen brukes mindre til å tolke bilder enn andre yrkesgrupper, som radiologer og ortopeder, og kan forklare hvorfor akuttmottakslegen hadde det dårligste resultatet for sensitivitet. I tillegg er Li *et al.* (2023) også den eneste studien som benytter håndkirurger til å granske bilder, noe som tyder på at denne yrkesgruppen heller ikke brukes like aktivt til dette. Dersom sensitiviteten til kirurgene sammenliknes med radiologene i Jacques *et al.* (2023), ved granskning for scaphoidfraktur, kan det derimot observeres at kirurgene i studien til Li *et al.* (2023) presterer best av de to yrkesgruppene. Yrkeserfaring kan i dette tilfellet hatt en betydning, ettersom Jacques *et al.* (2023) oppgir at gjennomsnittet for yrkeserfaringen til de 23 bildetolkende legene er 5.6 år. Dersom gjennomsnittet av erfaringen til de fire kirurgene i Li *et al.* (2023) regnes ut, er den 8.25 år

(Vedlegg 3). Dette kan forklare hvorfor resultatet for sensitivitet i studien til Jacques *et al.* (2023) er betydeligere lavere enn i Li *et al.* (2023), og hvorfor disse gruppene presterte dårligere enn KI-algoritmene.

Det er ikke alltid erfarne radiologer eller ortopeder som har mulighet til å granske bildene når skaden oppstår akutt. Dette ansvaret faller da på ikke-ortopediske kirurger, uerfarne radiologer og ortopeder, eller akuttmottaksleger, som må ta en rask avgjørelse om det er brudd eller ikke. Antallet brudd som overses kan øke, noe som fører til at flere pasienter går ubehandlet (Gan et al., 2019; Zhang et al., 2023).

Noen av studiene omtaler større forskjeller innen yrkeserfaring for radiologene. Dette gjelder studiene til Cohen *et al.* (2023), Jacques *et al.* (2023) og Hendrix *et al.* (2023). Cohen *et al.* (2023) dannet et granskningsteam bestående av 41 radiologer. Studien opplyser at radiologene deres var hovedsakelig yngre og mindre erfarne, og hadde ikke noe form for spesialisering. I denne gruppen var det 29 leger i spesialisering (LIS), som var på sitt 4. eller 5. år av spesialiseringsutdannelsen, åtte radiologer og fire overleger. Et tilsvarende oppsett ble benyttet i studiet til Jacques *et al.* (2023), hvor de hadde 23 radiologer med 2-25 års yrkeserfaring innen radiologi. Ni av radiologene var LIS-leger med erfaring innen granskning av traumebilder, og 14 var mer erfarne radiologer med spesialisering innen muskel- og skjelett. Til slutt benyttet Hendrix *et al.* (2023) fem erfarne muskel- og skjelett radiologer. Disse hadde 5, 7, 22, 24 og 26 år med erfaring. KI-algoritmen presterte stort sett bedre enn radiologene i alle de tre overnevnte studiene, hvor forskjellen varierte med 4-7 prosentpoeng. Et unntak er sensitiviteten ved gransking av scaphoidbilder, hvor radiologene i studien til Hendrix *et al.* (2023) gjorde det noen få prosentpoeng bedre enn KI. En mulig underliggende årsak til dette kan være forskjeller i radiologenes arbeidserfaring. I Hendrix *et al.* (2023) ble det benyttet radiologer med mange års erfaring. Derimot blir det hos Cohen *et al.* (2023) og Jacques *et al.* (2023) oppgitt at radiologene ikke var like erfarne.

Forskjellen i erfaring hos radiologene kan være en faktor som påvirket de endelige resultatene presentert hos disse studiene. Ettersom at det ble benyttet radiologer med mindre yrkeserfaring, kan resultatene ha blitt lavere enn om studiene hadde tatt i bruk mer erfarne radiologer. I et slikt tilfelle, er det mulig at sensitiviteten og spesifisiteten til radiologene hos disse studiene hadde blitt høyere. Samtidig er det mulig at studiene bedre representerer yrket i klinisk praksis, ved at de bruker radiologer med forskjellig erfaring

innenfor granskning for fraktur. Ikke alle sykehus har erfarne radiologer tilgjengelig døgnet rundt, og ved at studiene benytter leger med variert arbeidserfaring, får de bedre representert den virkelige situasjonen.

Avslutningsvis er det mulig å se en sammenheng mellom yrke og erfaring blant de bildetolkende legene. Disse faktorene kan påvirke resultatene for sensitivitet og spesifisitet. Som tidligere diskutert, har yrkeserfaring en påvirkning på diagnostiseringsevnen til legene. Et eksempel på dette er studiet til Jacques *et al.* (2023). I tillegg er yrke en sentral faktor som kan påvirke hvor gode legene er på å diagnostisere frakturer. Dette kommer godt frem i studiene til Ozkaya *et al.* (2022) og Li *et al.* (2023).

5.4 Metodekritikk

For å samle inn studier til denne oppgaven, ble det benyttet to databaser. Årsaken til dette er at de andre forsøkte databasene kun resulterte i duplikater, og ikke ga nye relevante treff. Det forekom noen forskjeller i søkeordene som ble benyttet mellom PubMed og Scopus. I begynnelsen hadde vi ikke en fullstendig forståelse for hvordan søkemotorene fungerte. PubMed var den første databasen som ble brukt, hvor søket ble snevret inn ved å ta i bruk færrest mulige søkeord. Dette er på bakgrunn av at enhver ny kombinasjon med «OR», endte opp med å legge til flere tusen irrelevante studier i søket. Dette var mest sannsynlig brukerfeil fra vår side, og kan ha ført til at noen relevante studier ikke ble oppdaget. I Scopus ble derimot søkestrengen spesifisert. Som et forsøk på å inkludere flere artikler, ble det benyttet flere synonymer for medisinsk bildetolkning og anatomisk område i kombinasjonen «OR», noe som ikke ble gjort i PubMed. Det er derfor mulig at vi gikk glipp av relevante artikler til oppgaven, spesielt med tanke på antall databaser benyttet, og søkene utført i PubMed.

Grunnet manglende forskning på temaet i Norge, ble det benyttet artikler fra andre land og verdensdeler. Dermed er artiklene skrevet på engelsk. En svakhet med dette kan være misforståelse eller feiltolkning av informasjon og resultater, ettersom studiene benytter begreper og uttrykk vi ikke kjenner til. For å minimere risikoen for mistolkning, ble uklarheter diskutert i fellesskap for å oppnå en homogen forståelse.

Oppgavens datamateriale er innhentet fra kvantitative studier. Dermed baserer den seg på målinger, tall og statistikk, som bør styrke oppgavens troverdighet. Det må derimot tas forbehold om at det sammenliknes ulike situasjoner og tall fra studiene, og de kan derfor ha forskjellige forutsetninger. I enkelte studier var det nødvendig at vi selv regnet ut gjennomsnittet av de undersøkte faktorene, for å få verdier som kunne sammenliknes. Derfor må det også tas forbehold om at disse verdiene kan inneholde regnefeil. Utregninger er lagt til som vedlegg, slik at det er etterprøvbart. I tillegg er det en mulighet for at data har blitt feiltolket, som videre kan føre til en ukorrekt sammenfatning av våre resultater. For å forhindre dette ble alle studiene lest nøye gjennom individuelt, og deretter diskutert i fellesskap for å forsikre at en felles forståelse av materialet er oppnådd.

Relevans og gyldighet er ifølge Dalland (2021, s. 42) essensielt for validiteten av oppgaven. På grunn av få relevante artikler som tok for seg alle de undersøkte faktorene, ble det inkludert totalt ni studier for å være sikker på at dataene om sensitivitet og spesifisitet som ble innhentet var representative. Dette mener vi bidrar til å øke oppgavens validitet. Reliabiliteten av oppgaven er også viktig, da dette handler om at målinger er utført riktig slik at resultatene kan stoles på. Resultatene til de utvalgte studiene ble undersøkt, og det kunne ses likheter på tvers av studiene. Blant annet var sensitiviteten konsekvent lavere enn spesifisiteten i hovedsakelig alle studier. Dette var en likhet som var med på å støtte at verdiene er representative, og ikke tilfeldige.

Under studieseleksjonsprosessen ble det inkludert noen studier som skiller seg fra de andre studiene benyttet i oppgaven. Dette gjelder totalt fire studier, hvorav to brukte færre projeksjoner og de siste to oppga ikke isolerte verdier for spesifisitet i sine resultater. Gan *et al.* (2019) og Ozkaya *et al.* (2022) tar kun i bruk AP projeksjon ved deteksjon av frakturer. Derimot tar de resterende studiene i bruk flere projeksjoner, som lateral og scaphoidserie. Det var også et par studier som ikke hadde brukbare resultater for spesifisitet. Dette gjaldt studiene til Cohen *et al.* (2023) og Jacques *et al.* (2023), som hadde spesifisitet til KI og bildetolkende leger for flere anatomiske strukturer sammenslått. Disse faktorene kan derfor hatt en innvirkning på dataene, og dermed resultatet for sensitivitet og spesifisitet. Det ble tatt en beslutning om å inkludere disse studiene på tross av deres mangler, ettersom de inneholder verdier som kan brukes i resultatet. Hensikten med dette er at flere verdier kan bidra til å styrke oppgaven. De oppfylte også inklusjons- og eksklusjonskriteriene, og

ettersom det var et begrenset utvalg av artikler ble de valgt å inkluderes. Jacques *et al.* (2023) skriver også at studien deres ikke er ment til å sammenlikne KI med radiolog, men heller undersøke forskjellen mellom å ikke bruke versus å bruke KI som et verktøy når det kommer til granskning av bilder. Vi valgte å inkludere denne studien til tross for dette, fordi studien hadde verdier som kunne sammenliknes i tillegg til at den oppfylte inklusjons- og eksklusjonskriteriene.

6.0 Konklusjon

Denne oppgaven har undersøkt og sammenliknet sensitiviteten og spesifisiteten til KI og leger som tolker røntgenbilder av distale radius og scaphoid. Flere av studiene som er benyttet i oppgaven har selv konkludert med at KI har en tilsvarende lik, eller bedre evne til å detektere og utelukke frakturer enn de bildetolkende legene.

Flertallet av våre resultater viser at KI presterte på nivå med legene eller bedre, både ved sensitivitet og spesifisitet. I enkelte studier har algoritmen derimot en diagnostiseringsevne som er noen få prosentpoeng lavere enn hos de bildetolkende legene. Differansen som skiller KI og legene utgjør hovedsakelig få prosentpoeng. Disse forskjellene er såpass små, at vi vurderer det til at algoritmene presterer tilsvarende likt som legene. På bakgrunn av resultatene våre, kan vi i likhet med studiene konkludere med at KI har evne til å detektere og utelukke frakturer i distale radius og scaphoid, tilsvarende likt eller bedre enn de bildetolkende legene.

Flere av studiene tar for seg feildiagnoser, både for KI og de bildetolkende legene. Begge gruppene produserer feil, både på samme og forskjellige bilder. I studien til Ozkaya *et al.* (2022), produserte KI algoritmen færre feildiagnoser enn akuttmottakslegen og den mindre erfarne ortopeden, men flere enn den erfarne ortopeden. Disse resultatene virker lovende, men det kan konkluderes med at KI trenger å trene på større datasett med økt variasjon i frakturenes utseende.

I oppgaven undersøkes også hva yrker og erfaring kan ha å si for sensitiviteten og spesifisiteten. Det observeres at radiologene og ortopedene, som er de mest benyttede yrkene i studiene, hovedsakelig har bedre resultater enn kirurgene og akuttmottakslegen. I enkelte tilfeller er det de to førstnevnte yrkesgruppene som gjør det bedre enn KI. Dermed kan det konkluderes med at KI ikke vil erstatte legene per nå, men på sikt kan det være mulig at algoritmen får en større rolle innen diagnostisering. I dag kan algoritmen brukes som et verktøy for mindre erfarne bildetolkende leger, noe som også kan føre til reduksjon av feildiagnoser. Avslutningsvis kan det konkluderes med at det trengs ytterligere forskning innen temaet.

7.0 Litteraturliste

Studier inkludert i oppgaven

Blüthgen, C., Becker, A. S., Vittoria de Martini, I., Meier, A., Martini, K. & Frauenfelder, T. (2020). Detection and localization of distal radius fractures: Deep learning system versus radiologists. *European Journal of Radiology*, *126*, 108925. https://doi.org/10.1016/j.ejrad.2020.108925

Cohen, M., Puntonet, J., Sanchez, J., Kierszbaum, E., Crema, M., Soyer, P. & Dion, E. (2023). Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs. *European Radiology*, *33*(6), 3974-3983. https://doi.org/10.1007/s00330-022-09349-3

Gan, K., Xu, D., Lin, Y., Shen, Y., Zhang, T., Hu, K., Zhou, K., Bi, M., Pan, L., Wu, W. & Liu, Y. (2019). Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. *Acta Orthopaedica*, *90*(4), 394-400. https://doi.org/10.1080/17453674.2019.1600125

Hendrix, N., Hendrix, W., van Dijke, K., Maresch, B., Maas, M., Bollen, S., Scholtens, A., de Jonge, M., Ong, L. S., van Ginneken, B. & Rutten, M. (2023). Musculoskeletal radiologist-level performance by using deep learning for detection of scaphoid fractures on conventional multi-view radiographs of hand and wrist. *European Radiology*, *33*(3), 1575-1588. https://doi.org/10.1007/s00330-022-09205-4

Jacques, T., Cardot, N., Ventre, J., Demondion, X. & Cotten, A. (2023). Commercially-available AI algorithm improves radiologists' sensitivity for wrist and hand fracture detection on X-ray, compared to a CT-based ground truth. *European Radiology*. https://doi.org/10.1007/s00330-023-10380-1

Li, T., Yin, Y., Yi, Z., Guo, Z., Guo, Z. & Chen, S. (2023). Evaluation of a convolutional neural network to identify scaphoid fractures on radiographs. *Journal of Hand Surgery [European Volume]*, *48*(5), 445-450. https://doi.org/10.1177/17531934221127092

Ozkaya, E., Topal, F. E., Bulut, T., Gursoy, M., Ozuysal, M. & Karakaya, Z. (2022). Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography. *European Journal of Trauma and Emergency Surgery*, *48*(1), 585-592. https://doi.org/10.1007/s00068-020-01468-0

Suzuki, T., Maki, S., Yamazaki, T., Wakita, H., Toguchi, Y., Horii, M., Yamauchi, T., Kawamura, K., Aramomi, M., Sugiyama, H., Matsuura, Y., Yamashita, T., Orita, S. & Ohtori, S. (2022). Detecting Distal Radial Fractures from Wrist Radiographs Using a Deep Convolutional Neural Network with an Accuracy Comparable to Hand Orthopedic Surgeons. *Journal of Digitital Imaging*, *35*(1), 39-46. https://doi.org/10.1007/s10278-021-00519-1

Zhang, J., Li, Z., Lin, H., Xue, M., Wang, H., Fang, Y., Liu, S., Huo, T., Zhou, H., Yang, J., Xie, Y., Xie, M., Lu, L., Liu, P. & Ye, Z. (2023). Deep learning assisted diagnosis system: improving the diagnostic accuracy of distal radius fractures. *Fronteirs in Medicine*, *10*, 1224489. https://doi.org/10.3389/fmed.2023.1224489

Annet litteratur

Abilgaard, A., Hopp, E., Sakinis, T., Roterud, H., Bjørnerud, A., Beyer, M. & Smith, H. (2018). Vil radiologer bli erstattet av kunstig intelligens? *Tidsskriftet*, *138*(17). https://doi.org/10.4045/tidsskr.18.0587

Barua, I. (2023). *Kunstig intelligens redder liv: AI er legenes nye superkrefter* (1. utg.). Cappelen damm.

Dalland, O. (2021). *Metode og oppgaveskriving* (7. utg.). Gyldendal.

Drake, C. & Levine, R. A. (2005). Sensitivity, specificity and other diagnostic measures with multiple sites per unit. *Contemporary Clinical Trials*, *29*(2), 252-259. https://doi.org/10.1016/j.cct.2004.12.008

Erickson, B. J., Korfiatis, P., Akkus, Z. & Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, *37*(2), 505-515. https://doi.org/10.1148%2Frg.2017160130

European commission. (2018). *A definition of AI: Main capabilities and disciplines*. Hentet 18.desember fra https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

European commission. (u.å.). *CE marking*. Hentet 1.mai fra https://single-market-economy.ec.europa.eu/single-market/ce-marking_en

Evans, D. (2002). Systematic reviews of interpretive research: Interpretive data synthesis of processed data. *Australian Journal of Advanced Nursing*, *20*(2), 22-26. https://www.ajan.com.au/archive/Vol20/Vol20.2-4.pdf

FDA. (u.å.). *FDA*. Hentet 1.mai fra https://www.fda.gov/

Filleron, T. (2017). Comparing sensitivity and specificity of medical imaging tests when verification bias is present: The concept of relative diagnostic accuracy. *European Journal of Radiology*, *98*, 32-35. https://doi.org/10.1016/j.ejrad.2017.10.022

Forsberg, C. & Wengström, Y. (2016). *Att göra systematiska litteraturstudier: Värdering, analys och presentation av omvårdnadsforskning* (4. utg.).

Gleamer. (u.å.). *BoneView*. Gleamer.AI. https://www.gleamer.ai/solutions/boneview

Gore, J. C. (2020). Artificial intelligence in medical imaging. *Magnetic resonance imaging*, *68*, A1-A4. https://doi.org/10.1016/j.mri.2019.12.006

Grønseth, R. & Jerpseth, H. (2019). *Bacheloroppgaven i sykepleie: Praktiske råd i skriveprosessen* (1. utg.). Fagbokforlaget.

Hamblen, D. L. & Simpson, A. H. R. W. (2007). *Adam's Outline of fractures, Including joint injuries* (12. utg.). Elsevier.

Hardy, M. & Harvey, H. (2020). Artificial intelligence in diagnostic imaging: impact on the radiography profession. *The British journal of radiology*, *93*(1108), 1-7. https://doi.org/10.1259/bjr.20190840

Hayashi, D., Kompel, A. J., Ventre, J., Ducarouge, A., Nguyen, T., Regnard, N. E. & Guermazi, A. (2022). Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. *Skeletal Radiology*, *51*(11), 2129-2139. https://doi.org/10.1007/s00256-022-04070-0

Helsebiblioteket. (2021, 17.09.2021). *4.1 Sjekklister*. Helsebiblioteket. https://www.helsebiblioteket.no/innhold/artikler/kunnskapsbasert-praksis/kunnskapsbasertpraksis.no/4.kritisk-vurdering/4.1-sjekklister

Helsedirektoratet. (2021). *Tilrettelegging for bruk av kunstig intelligens i helsetjenesten*. Helsedirektoratet. https://www.helsedirektoratet.no/rapporter/tilrettelegging-for-bruk-av-kunstig-intelligens-i-helsesektoren-ny-01.10.2021/

Hernæs, K. H. & Skyrud, K. D. (2022). *Framtidens utfordringer for folkehelsen* (Temautgave 2022). Folkehelseinstituttet,. https://www.fhi.no/contentassets/1da364574c4d46649008cd300acb4602/folkehelserapporten---temautgave-2022.pdf

Jung, J., Dai, J., Liu, B. & Wu, Q. (2024). Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis. *PLOS Digit Health*, *3*(1), e0000438. https://doi.org/10.1371/journal.pdig.0000438

Malamateniou, C., Nkapp, K. M., Pergola, M., Woznitza, N. & Hardy, M. (2021). Artificial intelligence in radiography: Where are we now and what does the future hold? *Radiography*, *27*, S58-S62. https://doi.org/10.1016/j.radi.2021.07.015

Malterud, K. (2017). *Kvalitative metasyntese som forskningsmetode i medisin og helsefag*. Universitetsforlaget.

Matre, K. & Hole, R. M. (2015). *Brudd behandling* (2. utg.). Legeforlaget.

Meena, T. & Roy, S. (2022). Bone Fracture Detection Using Deep Supervised Learning from Radiological Images: A Paradigm Shift. *Diagnostics*, *12*(10). https://doi.org/10.3390/diagnostics12102420

Metodebok. (2023, 20.10.2023). Distale radiusfakturer. I *Metodebok* (3.1. utg.). Hentet 5.mai fra https://metodebok.no/index.php?action=topic&item=VcRjNDBp

Morgan, M., Tigges, S., Sciacca, F. & Knipe, H. (2015, 6. Januar, 2024). *Sensitivity and specificity*. Hentet 22.april fra https://radiopaedia.org/articles/sensitivity-and-specificity

OECD.AI. (u.å.). *Catalogue of Tools & Metrics for Trustworthy AI; Metrics: Accuracy*. Hentet 1.mai fra https://oecd.ai/en/catalogue/metrics/accuracy

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R. L., Langlotz, C., Shpanskaya, K., Lungren, M. P. & Ng, A. Y. (2018). MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. *arXiv:1712.06957 [physics.med-ph]*, (4). https://doi.org/10.48550/arXiv.1712.06957

Shreffler, J. & Huecker, M. R. (2023). *Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios*. StatPearls [internet]. Treasure Island (FL). https://www.ncbi.nlm.nih.gov/books/NBK557491/

Støren, I. (2010). *Bare søk! Praktisk veiledning i å systematisere kunnskap* (1. utg.). Cappelen Damm.

Thodberg, H. H., Thodberg, B., Ahlkvist, J. & Offiah, A. C. (2022). Autonomus artificial intelligence in pediatric radiology: the use and perception of BoneXpert for bone age assessment. *Pediatric radiology*, *52*(7). https://doi.org/10.1007%2Fs00247-022-05295-w

Tørresen, J. (2013). *Hva er kunstig intelligens*. Universitetsforlaget.

Vestre Viken. (2024). Kunstig intelligens. https://www.vestreviken.no/fag-og-forskning/bildediagnostikk/kunstig-intelligens/#visste-du-at

Aase, H. S., Nes, H. & Brøgger, H. M. (2022). Radiologi og framtid. *Overlegen 2*, *22*. https://overlegen.digital/fagbladet/2-2022/8/

Figurer

Interaction Design Foundation. (2016). What is artificial intelligence (AI)? I. Interaction Design Foundation. https://www.interaction-design.org/literature/topics/ai

Morgan, M. A. (2015). Derivations of sensitivity and specificity. I. Radiopaedia. https://radiopaedia.org/cases/receiver-operating-characteristic-roc-curve

Vedlegg

Vedlegg 1: Søkehistorikk

Søk	**Database**	**Søkeord**	**Dato**	**Antall treff**	**Artikler benyttet**	**Tittel**	**Forfattere**
1	Scopus	"Artificial intelligence" AND "wrist fracture" OR "distal radius fracture" AND "diagnostic imaging" OR "radiographic image interpretation" OR "image analysis" AND x-ray	07.03.2024	153	1	Commercially-available AI algorithm improves radiologists' sensitivity for wrist and hand fracture detection on X-ray, compared to a CT-based ground truth	Jacques *et al.* (2023)
2	PubMed	Fracture AND artificial intelligence AND X-ray AND diagnostic AND wrist	26.02.2024	60	5	Deep learning assisted diagnosis system: improving the diagnostic accuracy of distal radius fractures	Zhang *et al.* (2023)
						Detection and localization of distal radius fractures: Deep learning system versus radiologists	Blüthgen *et al.* (2020)
						Detecting distal radial fractures from wrist radiographs using a deep convolutional neural network with an accuracy comparable to hand orthopedic surgeons	Suzuki *et al.* (2022)

						Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs	Gan *et al.* (2019) Cohen *et al.* (2023)
3	PubMed	Artificial intelligence AND x-ray AND scaphoid	03.03.2024	20	2	Evaluation of a convolutional neural network to identify scaphoid fractures on radiographs Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography	Li *et al.* (2023) Ozkaya *et al.* (2022)
4	Scopus	"Artificial intelligence" AND "scaphoid fracture" AND "diagnostic imaging" OR "radiographic image interpretation" OR "image analysis" AND x-ray	22.04.2024	61	1	Musculoskeletal radiologist-level performance by using deep learning for detection of scaphoid fractures on conventional multi-view radiographs of hand and wrist	Hendrix *et al.* (2023)

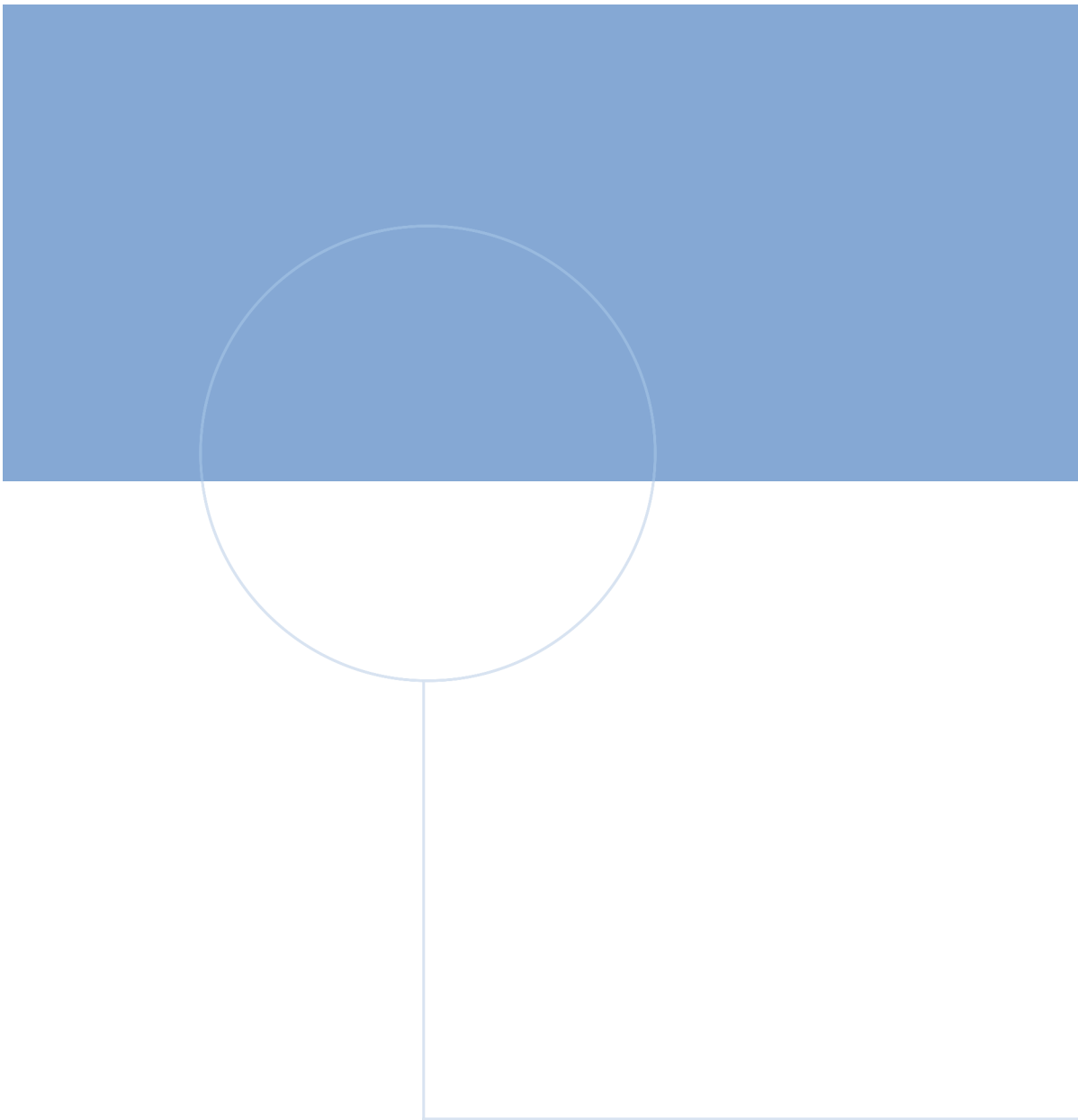
Vedlegg 2: Utregning av gjennomsnittet av Se og Sp til utvalgte studier

Studier	Beregnet hos	Sensitivitet/ spesifisitet	Tall i % fra studiene	Gjennomsnitt i %
Blüthgen *et al.* (2020)	KI	Sensitivitet	(81 + 80 + 90 + 82)/4	83.25%
		Spesifisitet	(100 + 86 + 97 + 78)/4	90.25%
	Radiolog	Sensitivitet	(86 + 98 + 86 + 98 + 90 + 98)/6	92.67%
		Spesifisitet	(86 + 94 + 97 + 90 + 62 + 76)/6	84.17%
Suzuki *et al.* (2022)	Ortoped	Sensitivitet	(96.0 + 96.0 + 96.0)/3	96%
		Spesifisitet	(98.7 + 93.3 + 97.3)/3	96.43%
Ozkaya *et al.* (2022)*	Ortoped	Sensitivitet	(72.0 + 86.0)/2	79%
		Spesifisitet	(92.0 + 98.0)/2	95%
Hendrix *et al.* (2023)	Radiolog	Sensitivitet	(75 + 75 + 83 + 80 + 83)/5	79.2%
		Spesifisitet	(94 + 84 + 88 + 91 + 81)/5	87.6%

*Verdier fra Ozkaya *et al.* (2022) er omgjort til prosent ved å gange verdiene med 100

Vedlegg 3: Utregning av gjennomsnitt for yrkeserfaring i Li et al. (2023)

Studie	Yrkeserfaring	Gjennomsnitt
Li *et al.* (2023)	(3 år + 3 år + 13 år + 14 år)/4	8.25 år



NTNU

Kunnskap for en bedre verden