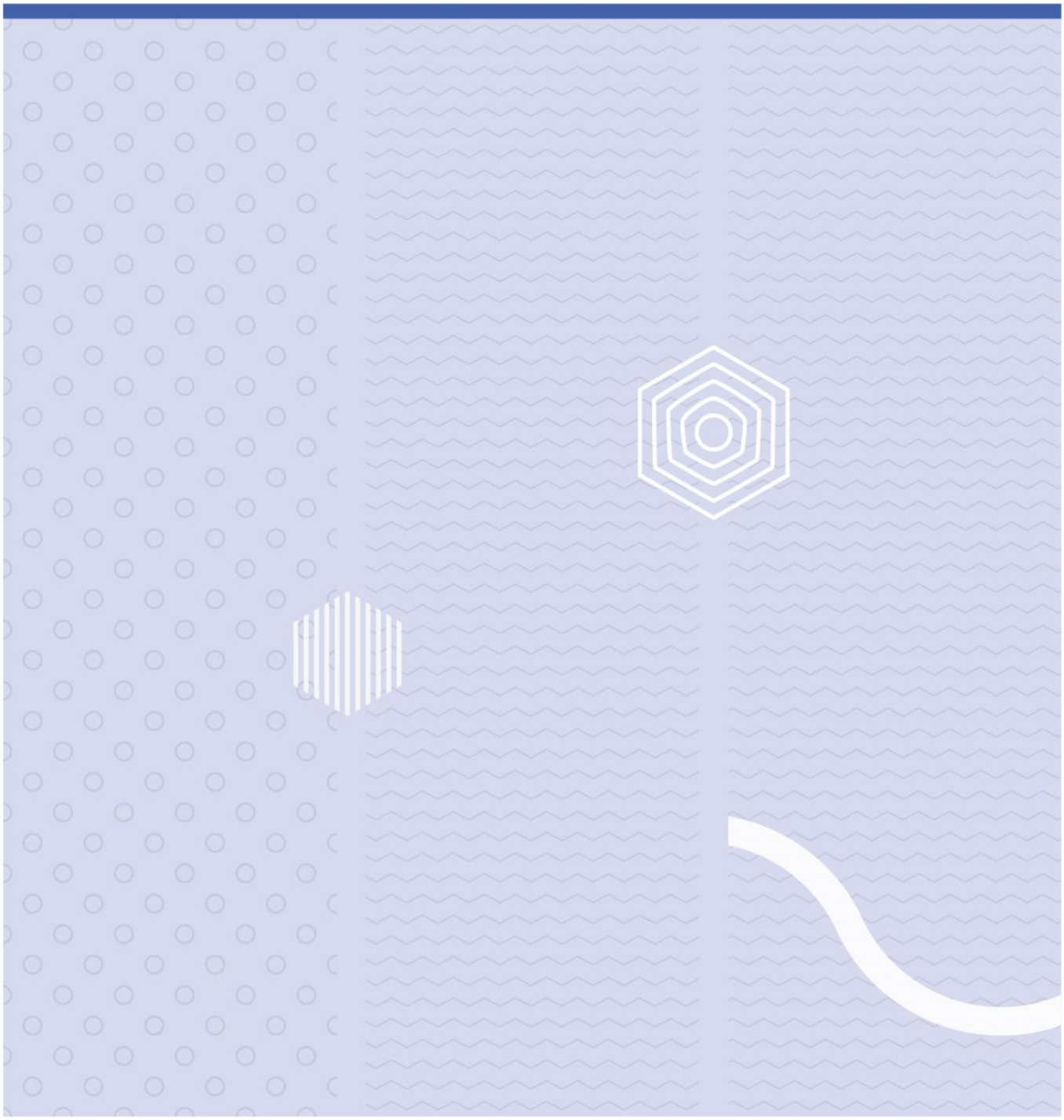
Universitetet i Sørøst-Norge
Fakultet for Helse – og sosialvitenskap
Institutt for Optometri, Radiografi og lysdesign
–

Bacheloroppgave
21/05/21
RADFOR610

Kandidatnummer: 6037, 6040, 6041 og 6049

«Hvordan kan dyp læring bistå radiologer i et screeningprogram for lungekreft med CT?»



Innholdsfortegnelse

Forord

Denne oppgaven er avsluttende for radiografi utdanningen ved Universitetet i Sørøst-Norge, avdeling Drammen. I denne bacheloroppgaven ønsker vi å belyse hvordan dyp læring kan bistå radiologer med hensyn til diagnostisk nøyaktighet, sensitivitet og spesifisitet ved screening av lungekreft. Vi vil også se på om dyp læring kan være med på å redusere antall falske positive og falske negative. Vi har tatt for oss et spennende tema som er i stadig utvikling. Det har vært veldig interessant å jobbe med kunstig intelligens som kommer til å få en større rolle innen radiologi i fremtiden.

Vi vil takke vår veileder Endre Grøvik for engasjement, konstruktive tilbakemeldinger og god veiledning. Vi vil også takke bibliotekarene på USN Drammen som har hjulpet oss med søk, kilder og kvalitetsvurdering i forbindelse med oppgaven.

Drammen, mai 2021

Kandidatnummer: 6037, 6041, 6040, 6049
RADFOR610, Universitet i Sørøst - Norge

Sammendrag

Mål

Målet med oppgaven er å undersøke om dyp læring kan bistå radiologer når det kommer til spesifisitet, sensitivitet og nøyaktighet ved et lungescreeningprogram med CT. Vi vil også se på om dyp læring kan redusere antall falske negative og falske positive.

Metode

Metoden vi benyttet for denne oppgaven var en systematisk oversikt. Søk ble gjort i MEDLINE via Ovid, Embase og usystematiske søk i Google Scholar og Nature. Søkeord ble identifisert ved bruk av PICO skjema. Søkeordene som ble benyttet til å søke i databasene var deep learning, lung screening, lung cancer og CT. Artiklene ble kvalitetsvurdert ved hjelp av en egen sjekkliste.

Resultat

Det ble tatt utgangspunkt i fem forskningsartikler. Resultatet er delt inn i tre deler: Sensitivitet og spesifisitet, nøyaktighet og falske positive og falske negative.

Konklusjon

I de inkluderte artiklene oppnår DL - modellene resultater for nøyaktighet, sensitivitet og spesifisitet på lik linje, eller bedre enn radiologer. DL - algoritmene har også vært med på å bistå til bedre resultater ved å redusere både falske negative og positive. Dette viser til at DL kan bidra til å gjøre noen av de tidkrevende og omfattende oppgavene i et screeningprogram for lungekreft. Det kan være med på å forhindre flere dødsfall uten å påføre radiologer en større belastning.

Nøkkelord

Dyp læring, Lungekreft, CT, Radiolog, Falske positive, Falske negative, Nøyaktighet, Sensitivitet og Spesifisitet.

Ordliste

DL	Deep learning / Dyp læring
CNN	Convolutional neural network
CT	Computer Tomography
LIDC IDRI	The lung image Database Consortium image collection
ROC	Receiver operating Characteristic curve
AUC	Area under curve; Total area under the ROC Curve
NLST	National Lung Screening Trial
CAD	Common computer aided diagnosis

1 Innledning

I året 2019 var det i overkant av 3200 nye tilfeller av lungekreft i Norge. Lungekreft er den tredje hyppigste kreftformen som rammer både kvinner og menn. Siden 1950-årene har antall tilfeller av lungekreft vært jevnt økende (Skjønsberg & Hofsli, 2020). Lungekreft er en kreftform som kan bruke lang tid på å utvikle seg, og rammer først og fremst personer som røyker eller de som har røyket tidligere (Kreftregisteret, 2021).

I en studie fra Belgia og Nederland deltok en gruppe mennesker som enten var røykere eller hadde røket tidligere. Noen av deltakerne ble tilfeldig valgt ut i et screeningprogram over 10 år. Dette førte til en reduksjon på 24% dødelighet av de som deltok i studiet. På bakgrunn av bl.a. denne studien arbeider Norsk lungekreftgruppe for å få et screeningprogram for lungekreft også til Norge. Dette gir et grunnlag til å tro at man kan redde over 500 liv dersom man utfører lavdose CT-undersøkelser av lungene til de i befolkningen av alle røykere og tidligere storrøykere (Skodvin, 2020).

I helsevesenet blir det stadig utviklet ny teknologi for å forbedre helsetjenestene, og særlig har kunstig intelligens tatt store fremskritt. Bruken av kunstig intelligens øker og vil få en større rolle i fremtiden innen radiologi, spesielt innenfor bildediagnostikk. Dyp læring (DL) er en type AI-system som tar i bruk kunstige nevrale nettverk for å analysere og evaluere store mengder data. Ved bruk av DL innen bildegjenkjenning og bildeanalyse vil dette kunne bidra til økt nøyaktighet og effektivitet innen bildediagnostikk og kan være et hjelpende verktøy for radiologer (Svoboda, 2020).

1.1 Problemstilling

Vi har valgt følgende problemstilling: *"Hvordan kan dyp læring bistå radiologer i et screeningprogram for lungekreft med CT?"*

For å avgrense denne problemstillingen har vi valgt følgende forskningsspørsmål:

- Hvordan kan DL bidra til diagnostisk nøyaktighet?
- Hvordan kan DL redusere falske positive og falske negative?

Målet for denne oppgaven er å få en oversikt over forskning innen DL, og hvordan ulike DL-teknikker kan bistå radiologer med å diagnostisere CT-bilder i et screeningprogram for lungekreft. I oppgaven vil vi fokusere på nøyaktighet, spesifisitet og sensitiviteten til de ulike DL-teknikkene, og se på hvordan DL bidrar til å redusere falske positive og falske negative. Det er kun tatt utgangspunkt i lavdose CT for å avgrense oppgaven.

2 Metode

I denne bacheloroppgaven gjennomførte vi en systematisk oversikt av temaet. En systematisk oversikt bruker både systematiske og avklarende metoder for å identifisere, velge ut og kritisk vurdere aktuell forskning. Dette utføres ved å innhente, sammenligne, analysere og gradere data i henhold til de studiene som er til stede i oversikten (Sykepleien, 2010). Systematisk oversikt er egnet for å svare på denne problemstillingen da det gir en faglig oppdatering på temaet som det allerede finnes flere studier og forskningsartikler om. Ved å ta utgangspunkt i problemstillingen og et PICO-skjema skal vi utføre relevante søk om temaet og gjøre en analyse av disse i oppgaven. Ved hjelp av god søkestrategi har dette gitt oss mulighet til å se på ulike studier som er gjort, for å se hvilke metoder innen temaet som har gitt best resultater og hvilke metoder som anbefales.

2.1 Identifisering av studier

Det ble gjort systematisk søk på MEDLINE via Ovid og Embase. I tillegg ble det gjort usystematisk søk i Google Scholar og Nature for å få et bredere utvalg av forskningsartikler. Søkene ble utført mellom 15 og 31 mars 2021. For å gjennomføre søkene lagde vi et PICO-skjema som tar for seg de viktige søkeordene i vår problemstilling. PICO-skjema blir benyttet for å definere søkeordene som ble brukt i søket (Helsebiblioteket, 2016). Vedlegg 1 viser søkeord på engelsk.

2.2 Søkestrategi i MEDLINE Ovid

Hovedkomponentene i søket på MEDLINE var deep learning, lung screening, lung cancer og CT. Ved å bruke “AND” i søkestrategien finner man overlapping mellom de ulike nøkkelordene man bruker i søket. Ved bruken av ‘’OR’’ samler man søkeord som er synonymer eller alternative begreper. Dette gir mulighet for utvidelse og sammenkobling av søkeord. Vi begrenset søket vårt til kun utgivelser fra 2015 til i dag da DL er forsket mye på de siste årene og derfor må søkene være nyere for å få en oppdatert systematisk oversikt.

Google Scholar og Nature er to av usystematiske søk som ble gjort for å finne forskningsartikler som omhandler tema om Lungekreft, DL, lunge screening og nøyaktighet. På nature søkte vi med søkeordene våre, men dette var ikke like effektivt som ved bruken av MEDLINE på å avgrense søket.

2.3 Utvelgelse av artikler

De inkluderende artiklene hadde krav om å inneholde hvordan DL kunne bistå radiologer, med fokus på nøyaktighet, spesifisitet og nøyaktighet. Resultatene fra studiene skulle vi bruke til å sammenligne ulike DL - modeller og radiologenes presentasjoner, for å se om DL kan hjelpe radiologer med å oppnå bedre resultater. Artiklene skulle også inneholde falske positive og falske negative. Artikler ble ekskludert hvis de inneholdt andre modaliteter enn CT og hvis studiene var eldre enn 2015.

2.4 Kvalitetsvurdering

For å kvalitets vurdere artiklene som er inkludert har vi laget vår egen sjekkliste i og med at våre artikler er av en teknisk karakter, og bruker andre studiedesign enn tradisjonelle medisinske artikler. Vi har brukt MMAT (Mixed Methods Appraisal Tool) som inspirasjon da dette verktøyet er beregnet for flere typer studiedesign enn Helsebiblioteket sine (Hong et al., 2018. s. 2). Spørsmålene 1 til 5 som vises nedenfor er de kriteriene vi går ut ifra i vurderingen. Det gis ett poeng for hvert kriterium som er fulgt.

1. Er formålet med studien klart formulert?
2. Er forskningsspørsmålene tydelige?
3. Går det klart frem av metoden hvordan studien er utført?
4. Er metoden egnet til å besvare forskningsspørsmålene?
5. Kommer resultatene klart frem?

2.5 Analyse

En grundig analyse av studiene er en nødvendighet når man skal lese og systematiserer dataen man henter. Samtidig er det viktig være kritisk til det man leser. Ved vår analyse av artiklene har vi brukt Kirsti Malterud sin metode som innebærer å dele analysen inn i fire ulike trinn (Malterud, 2011, s. 91):

1. Lese gjennom og få et oversiktlig bilde av studiet. Videre kan man drøfte ulike tema som kan være relevante til den valgte problemstillingen (Malterud, 2011, s. 98-99).
2. Organisere de relevante studiene. Her tok gruppen og systematiserte informasjonen merket med en farge, ved å ta i bruk denne teknikken kunne man kartlegge og gruppere ut informasjon relevant til studien vår (Malterud, 2011, s. 104).
3. Trekke ut og tolke relevant informasjon fra studiet (Malterud, 2011, s.104-107).
4. Legge sammen de ulike delene av teksten slik at de får frem hovedpunktene i studiet (Malterud, 2011. s. 108-109).

3 Teori

3.1 Lungekreft

Lungekreft er den kreftformen som tar flest liv hos menn og kvinner i Norge. Lungekreft inkluderer alle typer kreftformer i lungevevet og luftveiene. Man deler lungekreft inn i to hovedtyper: småcellet og ikke-småcellet. De fleste pasientene har kun hatt symptomer i få måneder før diagnosen er stilt. Selv om diagnosen blir oppdaget tidlig, er ofte sykdommen fremskreden og det er kun mulighet å fjerne svulsten ved operasjon ved 20-25% av tilfellene. Dette har ført til at sjansen for overlevelse er minimal og kun 8-15% er i live 5 år etter at diagnosen er konstatert. Lungekreft kan også utvikle seg til metastaser som spres til andre organer i kroppen (Felleskatalogen, 2020).

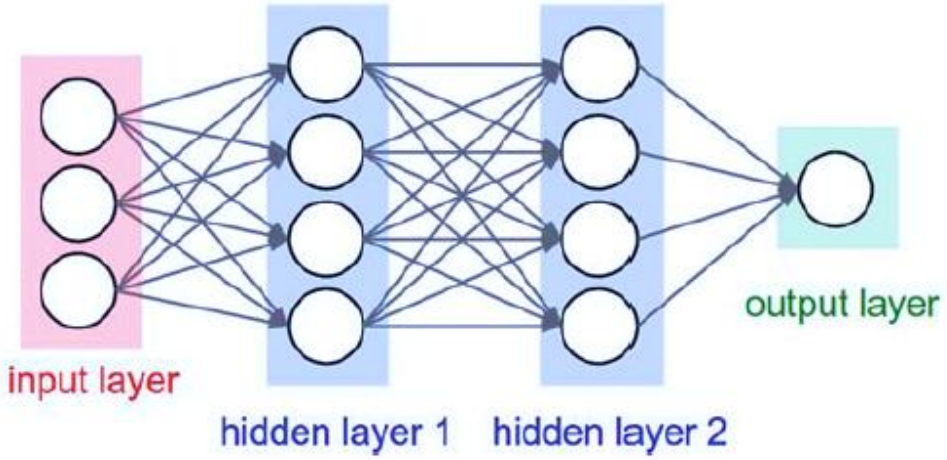
3.2 Lungescreening

En screeningundersøkelse er tester som blir utført for å finne sykdom før symptomene begynner. Formålet med screening er å oppdage sykdommen i en tidlig fase og på det mest behandlingsbare stadiet. Det viktigste målet med lunge screeningen er å redusere antall dødsfall fra en gitt sykdom. I lungescreening blir det utført tester av brystet med lav dose computertomografi (LDCT) av personer som har høy risiko for å utvikle lungekreft (RadiologyInfo, 2021).

3.3 Dyp læring

KI blir brukt innenfor de fleste felt i den moderne verden, spesielt er det formen maskinlæring som er den mest populære innen KI. I nyere tid er en form av maskinlæring kalt DL blitt mye tatt i bruk. Noen former av DL er basert på det samme prinsipp som det nevrale nettverket i hjernen. DL kan beskrives som et system hvor man behandler rådata og automatisk trekker ut de ønskede funksjonene. Den største forskjellen mellom maskin læring og DL er måten store mengder data kan bli håndtert på. Nettverket består av flere ulike lag, disse lagene er input, output og 2-3 skjulte lag (Kose & Alzubi, 2021, S. 136-137). Se figur 1.

Figur 1. Arkitekturen til et deep neural network (Kosa & Alzubi, 2021, s. 12).



3.4 CNN

CNN er en algoritme som er inspirert av den visuelle oppfatningen til mennesker. CNN er todimensjonale nettverk som ofte blir brukt innenfor medisinsk bildeanalyse, bilde klassifisering og objekt gjenkjenning. Bruken av CNN har gjort det mulig å ta for seg større og mer komplekse datasett enn det som var mulig tidligere (Kose & Alzubi, 2021, S. 137).

3.5 ROC og AUC

ROC (Receiver operating characteristic curve) er en sannsynlighetskurve. X-aksen indikerer en høyere antall falske positive enn sanne negative. Mens en høyere verdi langs Y-aksen indikerer et høyere antall sanne positive enn falske negativer (Narkhede, 2018).

AUC (Area under the curve) forteller oss hvor mye den modellen er i stand til å skille mellom ulike klasser. Jo høyere AUC, desto bedre er modellen til å forutsi. Modellen er utmerket når AUC er nær 1 (Narkhede, 2018).

3.6 Sensitivitet

Sensitivitet handler om evnen en test har for å kunne identifisere en sykdom riktig. Jo høyere sensitivitet en undersøkelse har, jo bedre sjansene er det for å få en sann positiv test. (Mokobi, F. (2020). Sensitiviteten er definert som:

$$\frac{Sann\ positive}{(Sann\ positive + falsk\ negative}$$

3.7 Spesifisitet

Spesifisitet er evnen en klinisk test har for å kunne identifisere at en diagnose ikke er korrekt (Mokobi, F. (2020). Spesifisiteten av en test er forholdet mellom en frisk pasient eller en pasient som er kjent for å ikke ha sykdommen, og teste negativ for den (Kose & Alzubi, 2021, s. 132). Spesifisitet er definert som:

$$\frac{Sann\ negative}{(sann\ negative + falsk\ positive)}$$

Spesifisitet er omvendt proporsjonal med sensitivitet, dette betyr at hvis sensitiviteten øker, vil spesifisiteten synke og omvendt.

Figur 2. Forvirring Matrise (Rajan, s. 2020).

	Har sykdom	Har ikke sykdom
Identifisert: Sykdom	Sann positiv (TP)	Falsk positiv (FP)
Identifisert: Ikke sykdom	Falsk negativ (FN)	Sann negativ (TN)
	Sensitivitet	**Spesifisitet**

3.8 Nøyaktighet

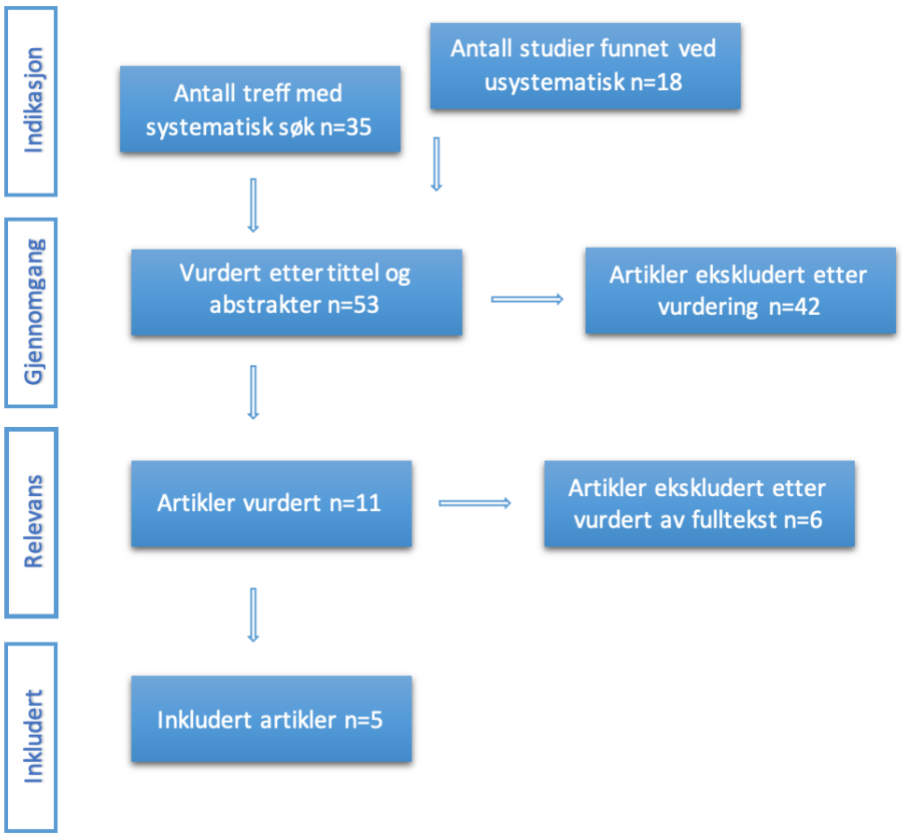
Nøyaktighet kan beskrives som den prosentandelen av antall positive resultater gitt det antallet av pasienter som deltar i en undersøkelse (Kose & Alzubi, 2021, s. 133). Formelen for nøyaktighet er definert som:

$$\frac{(sann\ positive + sann\ negative)}{(sann\ positive + sann\ negative + falsk\ negative + falske\ positive)}$$

4 Resultat

Ved systematisk søk ble 35 artikler funnet mens ved det usystematiske søket ble 18 artikler funnet og valgt til vurdering. De valgte artiklene fra systematisk og usystematisk søk ble gjennomgått av samtlige i gruppen for å vurdere om de var aktuelle. I prosessen videre ble de gjenværende 53 artiklene vurdert ut ifra sammendragene. Hvis sammendraget ikke omhandlet DL og CT lunger ble de ekskludert og ikke tatt med i videre vurdering. De 11 gjenværende artikler ble vurdert som aktuelle ble også gjennomgått av samtlige i gruppen. Videre i analysen ble gjort på bakgrunn av fulltekst. Av de 11 artikler ble seks ekskludert og vi satt igjen med fem valgte artikler. Artiklene som ble ekskludert etter full vurdering av teksten ble gjort på bakgrunn av at artiklene ikke handlet om nøyaktighet, sensitivitet, spesifisitet og om andre bildediagnostiske modaliteter.

Figur 3. Flytdiagram av utvelgingsprosessen.



Tabell 1. Oversikt over inkluderte artikler.

Forfatter og år	Navn på artikkel/problemstilling	Journal	Metode	Type data	Antall deltakere	Område/ Nasjonalitet	Kvalitetsvurdering med skår
Ardila, 2019	End-to-End Lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography	Nature Medicine	Studien foreslår en dyplæringsalgoritme som bruker pasientens nåværende og tidligere CT volum for å forutsi risikoen for lungekreft	Totalt 42,290 undersøkelser fra 14,851 pasienter	6 Radiologer	California, USA	3
Bhandary, 2019	Deep-learning framework to detect lung abnormality	Elsevier	Dette studie foreslår to ulike dyplærings metoder for å undersøke røntgen thorax og ct-lunger	Totalt 3500 røntgen og CT bilder	Ikke tilgjengelig	Karnataka, India	3
Riquelme & Akhloufi,	Deep learning for lung cancer nodules detection and classification in CT scans	MDPI	Studien gjennomgår nylige toppmoderne dyplæringsalgoritmer og arkitekturer foreslått som CAD systemer for påvisning av lungekreft.	244,527 bilder fra 1010 undersøkelser	4 Radiologer	Moncton, Canada	3
Song, 2017	Using deep learning for Classification of lung nodules on Computed tomography images	Hindawi	I denne artikkelen er det tre typer dype nevrale nettverk (CNN, DNN, og SAE) som er designet for lungekreft. Disse nettverkene brukes til CT- bilde klassifisering med noen modifikasjoner for godartede og ondartede lunge knuter.	244,527 bilder fra 1010 undersøkelser	4 Radiologer	Tianjin, China	3
Trajanovski, 2021	Towards radiologist-level cancer risk assessment in CT lung screening using deep learning	Computerized Medical Imaging and Graphics	Sammenligne ytelsen til dyplærings modell med moderne automatiserte algoritmer og radiologer som samt å vurdere algoritmens robusthet i heterogene datasett	Tre forskjellige databaser for CT-lungekreftscreening med lav dose	6 Radiologer	Cambridge, USA	5

4.1 Nøyaktighet, Sensitivitet og spesifisitet

Tabell 2 viser en oversikt over inkluderte artikler og hvordan de ulike DL teknikkene har scoret på nøyaktighet, sensitivitet og spesifisitet. 4 av 5 studier har tatt i bruk LIDC-IDRI, som er et bilde-databasesystem for screening av lungekreft på CT-bilder.

Tabell 2: Sammenligning av resultatene fra DL-modellene fra de valgte studiene.

Studie	DL - Teknikker	Nøyaktighet %	Sensitivitet %	Spesifisitet %
Ardila et al. 2019	CNN	-	81.5	89.3
Bhandary et al. 2019	MAN	97.27	98.09	95.63
Riquelme & Akhloufl. 2020	CNN	87.4	89.4	85.2
Song et al. 2017	CNN	84.15	83.96	84.32
Song et al. 2017	DNN	82.37	80.66	83.9
Song et al. 2017	SAE	82.59	83.96	81.35
Trajanovski et al. 2021	DNN	-	84	80

4.1.1 Sensitivitet og spesifisitet

Vi tatt for oss flere ulike DL-teknikker fra ulike studier som har foreslått deteksjon på knuter og vurdering av CT - lunge bilder. Alle DL-metodene viser høye verdier av nøyaktighet, sensitivitet og spesifisitet, men det er forskjeller fra studie til studie. Vi har samlet inn data fra de ulike studiene, Ardila et al. (2019), Bhandary et al. (2019), Riquelme & Akhloufl (2020), Trajanovski (2021) og Song et al. (2017). Fra tabell 2, kan vi se at DL-teknikken til Bhandary et al. (2019) har en sensitivitet og spesifisitet på henholdsvis 98.09 og 95.63% som er betydelig høyere.

4.1.2 Nøyaktighet

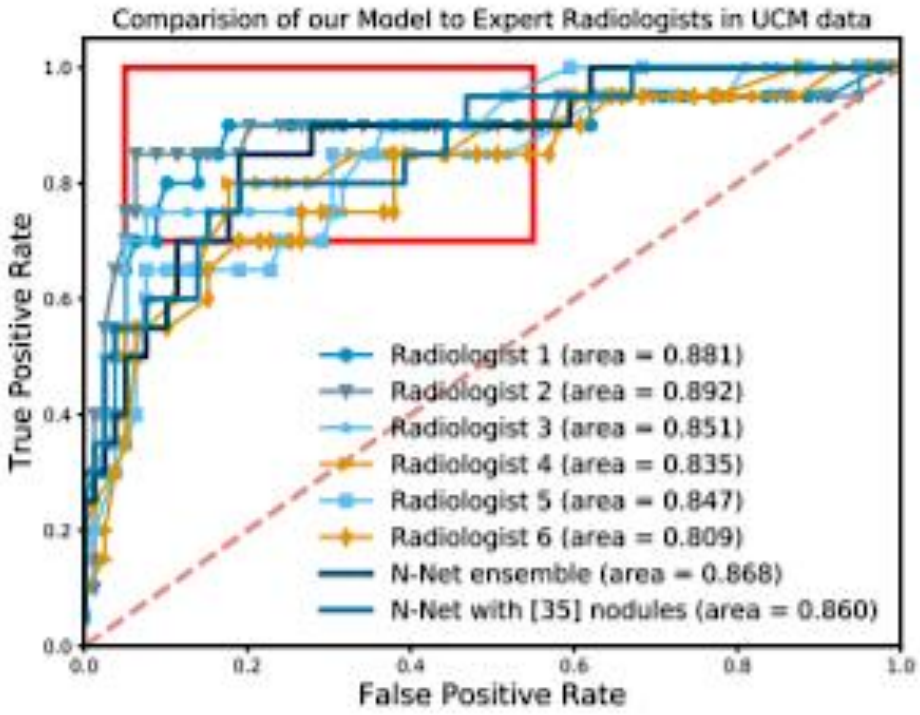
Kun tre av studiene evaluerte nøyaktigheten til DL-modellene. Studien til Bhandary et al. (2019) tar for seg klassifiseringen av lunge avvik som lungebetennelse og kreft ved hjelp av DL-modellen Modified Alex Net (MAN). MAN differensierte mellom maligne og benigne klasser på CT-lunge bilder, samt. ulike størrelser på lunge knuter. Studien indikerer at MAN-SVM bidrar til å oppnå forbedret klassifiseringsnøyaktighet (>97%) sammenlignet med andre DL-modeller (Bhandary et al., 2019, s. 2).

4.1.3 Falske positive og falske negative

Studien til Ardila et al. (2019) trente en CNN-algoritme og brukte longitudinelle pasientdata for å kunne forutsi risikoen for å utvikle lungekreft. DL-modellen i studiet oppnådde en reduksjon på 11% ved falske positive og 5 % ved falske negative (Ardila et al., 2019). Studien til Riquelme & Akhloufl (2020) vurderte bruken av topp moderne DL-modeller som Common computer aided diagnosis (CAD) for deteksjon av lungekreft. CAD systemene ble fordelt i to kategorier hvor den ene var et system for reduksjon av falske positive. Det høyeste resultatet oppnådd av en modell for klassifisering av falske positive og falske negative var på 0,996 AUC score (Riquelme & Akhloufl, 2020).

Studien til Trajanovski (2021) undersøker hvordan resultater en DL-modell som er designet for screening av lungekreft oppnår sammenlignet med seks radiologer. Figur 4 viser resultatet på AUC scoren mellom radiologene og DL-modellen. Resultatet viser at modellen oppnår en AUC score som er høyere enn fire av radiologene. De to siste radiologene oppnådde resultater som var høyere enn det vanlige gjennomsnittet (Trajanovski, 2021).

Figur 4. Sammenligning mellom resultatet til en DL-modell og seks radiologer. Figur med gitt tillatelse (Trajanovski, 2021).

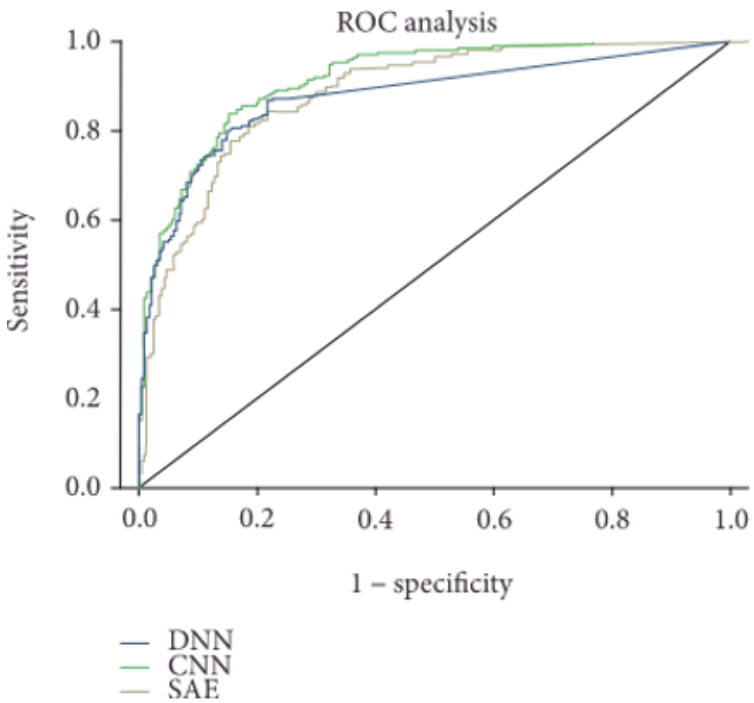


5 Diskusjon

5.1 Nøyaktighet, sensitivitet og spesifisitet

Studien til Song et al. (2017) tar for seg tre ulike DL-modeller som er designet for å oppdage lungepatologi: CNN, DNN og SAE (Song et al., 2017, s. 1). Basert på resultatene fra de tre DL-modellene hadde CNN den beste presisjonen, med nøyaktighet på 84,15%, sensitivitet på 83,96% og spesifisitet på 84,32%. Tabell 2 viser at DNN ikke scorer like godt på nøyaktighet og sensitivitet som SAE, men har en bedre effekt på spesifisitet på 83,9%. Når spesifisiteten er god betyr det at maligne lungeknuter hyppigere oppdages, noe som kan bidra til en tidlig diagnostisering av lungeknuter (Song et al., 2017, s. 6). For å sammenligne ytelsen til de ulike DL-modeller, blir det tatt i bruk ROC-kurve. Figur 5, sammenligner CNN og SAE og viser at CNN har en høyere nøyaktighet (AUC =0.916) enn SAE (AUC=0.877) (Song et al., 2017).

Figur 5. ROC kurve over forskjellig typer DL teknikker (Song et al., 2017, s. 5).



5.2 Falske positive og falske negative

Deteksjon av lungesvulster er en meget komplisert prosess og viktigheten av å oppdage svulstene er kritisk. Prosessen til en DL-modell benyttet av Riquelme & Akhloufl (2020) er delt inn i to steg. Det første steget tar for seg å finne alle sannsynlige lungesvulster, dette medfører at man får et høyt antall falske positive. Det andre steget går ut på å redusere det store antallet falske positive. Bruken av ulike DL-modeller som har andre oppbygninger fungerer på forskjeller måter og får derfor ulike resultater på reduksjon av falske positive og falske negative (Riquelme & Akhloufl, 2020, s. 61).

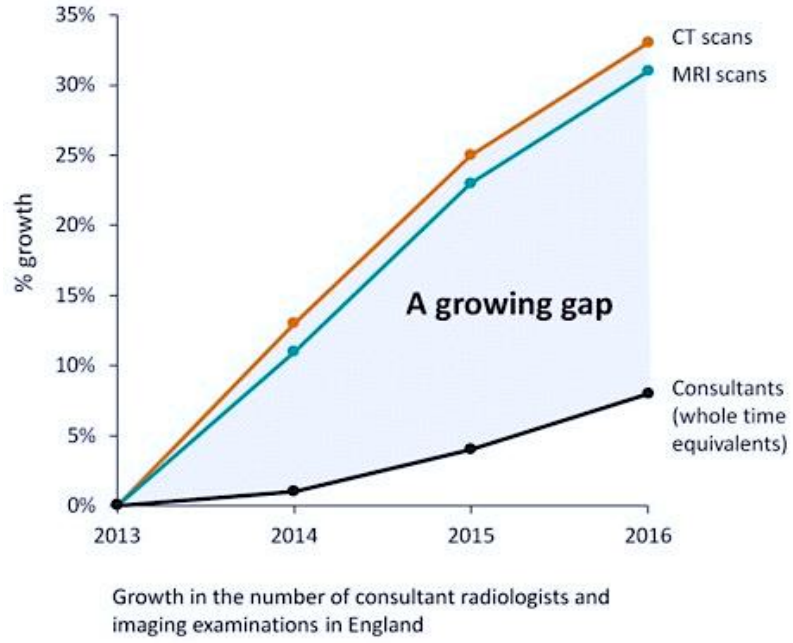
Både Ardila et al. (2019) og Trajanovski (2021) sammenlignet DL-modeller opp mot seks radiologer. Begge studiene oppnådde reduksjoner av falske positive på lik linje eller bedre enn radiologene. Ardila et al. (2019) var det eneste studiet som rapporterte en reduksjon av falske negative. DL-modellene oppnådde både like og bedre resultater enn det radiologene gjorde. Studiene fra Ardila et al. (2019) og Trajanovski (2021) konkluderer med at DL-modellene som ble benyttet kan både ta for seg klassifiseringen av svulster, eller fungere som et støtteverktøy for radiologer.

5.3 Utfordringer ved DL

Kunstig intelligens vil få en større rolle innen helse i fremtiden, men det er noen utfordringer som må forskes på før man kan realisere det helt ut (Nasjonalt senter for e-helseforskning, 2018).

CT bilder av lunge er svært tidkrevende å diagnostisere, og antall radiologer som vurderer disse undersøkelsene klarer ikke å dekke behovet av dagens økende etterspørsel. Dette kan ses i Figur 6. Denne arbeidsmengden kan føre til at erfarne radiologer gjøre feil. Menneskelig syn er også et faktum der radiologen kan overse små maligne svulster. 35% av lungeknuter blir oversett ved den første screeningen, derfor kan bruk av AI-systemer hjelpe med den store arbeidsmengden til radiologer og oppdage lungeflekker som ellers er vanskelig å oppdage (Svoboda, 2020).

Figur 6: Vekst i antall tolkende radiologer og bildeundersøkelser i England (Aumueller, J. 2019).



DL-modeller som blir brukt innen radiologi trenger tilgang til komplekse og store datamengder for trening, testing og validering av modellene, siden dagens problemer krever detaljerte analyser av større mengder pasientdata enn tidligere (Svoboda, 2020). Det oppstår flere etiske dilemmaer når det kommer til bruken av pasientdata til DL, bl.a. hvordan dataen blir utnyttet og hvordan personvernet til pasientene blir ivaretatt. Siden flere store IT-selskaper nå eier både sosiale medier og medisinske DL-modeller, oppstår det en risiko for at informasjonen kan utnyttes til pasientidentifikasjon. Dette kan føre til at informasjonen man ønsker å holde privat kan komme på avveie. Etisk bruk av pasientdata er derfor svært viktig, og det er satt som et krav at utviklere av modellene er klare over risikoen og gjøre alle nødvendige steg for å opprettholde personvern (Brady & Neri, 2020, s. 2-3).

En annen utfordring er bruken av algoritmer. Algoritmene som blir brukt kan gi dårlige eller feilaktige prediksjoner. Årsaker til dette kan være enten overtilpasning eller undertilpasning. Når det er en overtilpasning gjør algoritmen det bra på treningsdataene, men dårlig på nye prospektive data. Grunnen til dette er at algoritmen har lært spesifikke egenskaper i treningsdataene og ikke funnet en generell regel som forklarer variasjonen. Ved undertilpasning vil algoritmen fungere dårlig på treningsdata og

prospektive data. Disse to tilstandene er omvendt proporsjonale, det vil si at når den ene øker vil den andre reduseres. Man må derfor finne en balanse mellom disse, slik at prediksjonsfeilen er lavest mulig (Nasjonalt senter for e-helseforskning, 2018).

5.4 Metodekritikk

Artiklene er hentet fra ulike deler av verden: India, Kina, Canada og USA.
Relevant og viktig informasjon i denne oppgaven kan ha uteblitt, dette skyldes uthenting og prioritering av informasjon. Alle våre artikler er engelsk og skrevet med en faglig bakgrunn innen KI, dermed kan noe av informasjonen ha blitt oversatt eller tolket feil.
Vi har gjennomført kvalitetsvurdering av artiklene, men evnen til å vurdere artikler er begrenset. Kvalitetsvurderingen er gjort ut ifra sjekklisten som vi har måttet utvikle selv. KI er et relativt nytt fagfelt som stadig er i utvikling, noe som gjør at det finnes begrensninger på tilgjengelig forskning innenfor vår problemstilling.

6 Konklusjon

De valgte studiene viser at DL-modellene oppnår resultater for nøyaktighet, sensitivitet og spesifisitet på lik linje, eller høyere enn radiologer. DL kan bidra til å gjøre noen av de tidkrevende og omfattende oppgavene som gjør at screeningprogrammet for lungekreft kan forhindre flere dødsfall, uten å påføre radiologer en stor belasting. Ved å bistå radiologer med spesifisitet og sensitivitet ved screening av lungekreft vil man kunne redusere de høye kliniske og økonomiske kostnadene ved oversette diagnoser.

En av utfordringene som oppstår ved lungescreening er falske negative og falske positive. Ved bruken av DL-algoritmer unngår man å oppdage diagnoser i sene faser og redusere unødvendige biopsier som skyldes falske negative og falske positive (Ardila et al., 2019). DL har vært med på å bistå til bedre resultater ved reduksjon av både falske negative og falske positive.

7 Litteraturliste

Ardila, D., Kiraly, P.A., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D.P. & Shetty, S. (20. May. 2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine 25***,** 954-961. https://doi.org/10.1038/s41591-019-0447-x

Aumueller, J. (2019). How Artificial Intelligence can revolutionise healthcare. *Health Europa.* https://www.healtheuropa.eu/how-artificial-intelligence-can-revolutionise-healthcare/92824/

Bhandary, A., Prabhu, A., Rajinikant, V., Thanaraj, P.K., Satapathy, S.C., Robbins, D.E., Shasky, C., Zhang, Y.D., Tavares, J.M.R.S. & Raja, S.M. (2020). Deep-learning framework to detect lung abnormality - A study with chest X-ray and lung CT scan images. *Elsevier, Vol. 129:* 271-278. https://doi.org/10.1016/j.patrec.2019.11.013

Brady, A. & Neri, E. (2020). Artificial Intelligence in Radiology—Ethical Considerations. *Diagnostics*, *10*(4), 231. https://doi.org/10.3390/diagnostics10040231

Felleskatalogen (2020, 01.12) Lungekreft. https://www.felleskatalogen.no/medisin/sykdom/lungekreft

Helsebiblioteket (2016) PICO. https://www.helsebiblioteket.no/kunnskapsbasert-praksis/sporsmalsformulering/pio

Hong, Q. N., Pluye, P., Fabregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.P., Griffiths, F., Nicolau, B., O´Cathain, A., Rousseu, M.C. & Vedel, I. (2018). Mixed methods appraisal tool (MMAT), version 2018. *IC Canadian Intellectual Property Office, Canada.*

Kose, U. & Alzubi, J. (2021). *Deep Learning for Cancer Diagnosis*. Springer.

Kreftregisteret (2021, 21.01) Lungekreft. https://www.kreftregisteret.no/Temasider/kreftformer/Lungekreft/

Malterud, K. (2011). *Kvalitative metoder i medisinsk forskning.* (3.utg.) Universitetsforlag.

Mokobi, F. (2020). *What is sensitivity, Specificity, False positive, false negative?* Microbe Notes.https://microbenotes.com/sensitivity-specificity-false-positivefalsenegative/?fbclid=IwAR0oDKWYkpJvGIc4RfrOvjVGIpH7Oux9sXsBGycrzH9pS6DTwENzBytbtJo

Narkhede, S. (2018) Understanding AUC - ROC Curve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Nasjonalt senter for e-helseforskning (2018) Kunstig intelligens og maskinlæring i helsesektoren. https://ehealthresearch.no/faktaark/kunstig-intelligens-og-maskinlaering-i-helsesektoren

Radiologyinfo. (2021). *Lung Cancer Screening.* https://www.radiologyinfo.org/en/info/screening-lung

Rajan. S. (2020). *Analyzing the performance of classification models in machine learning.* Towards data science. https://towardsdatascience.com/analyzing-the-performance-of-the-classificationmodels-in-machine-learning-ad8fb962e857

Riquelme, D & Akhloufl, M.A. (2020). Deep learning for lung cancer nodules detection and classification in CT scans. *AI, 1*(1), 28-67. http://dx.doi.org/10.3390/ai1010003

Skjønsberg, O.H & Hofsli, E. (2020). Lungekreft i *Store medisinske leksikon*. https://sml.snl.no/lungekreft

Skodvin, T. Ø. (2020). CT-screening av lunger redder liv. *Tidsskriftet Den norske legeforeningen*. https://doi.org/10.4045/tidsskr.20.0175

Song, Q., Zhao, K., Luo, X & Dou, X. (2017). Using Deep Learning for Classification of Lung Nodules on Computed Tomography images. *Journal of Healthcare Engineering,* Vol. 2017. Artikkel 8314740. https://doi.org/10.1155/2017/8314740

Svoboda, E. (2020). Artificial intelligence is improving the detection of lung cancer. *Nature*, 5, 87, S20-S22. https://doi.org/10.1038/d41586-020-03157-9

Sykepleien (2010). Hvordan skrive en systematisk oversikt. https://doi.org/10.4220/sykepleienf.2010.0121

Trajanovski, S., Mavroeidis, D., Swisher, C.L., Gebre, G.B., Veeling, B.S., Wiemker, R., Kliunder, T., Tahmasebi, A., Regis, S.M., Wald, C., McKee, B.J., Flacke, S., MacMahon, H. & Pien, H. (2021). Towards Radiologist - level cancer risk assessment in CT lung screening using deep learning. *Computerized Medical Imaging and Graphics, 90,* Artikkel 101883. https://doi.org/10.1016/j.compmedimag.2021.101883

8 Vedlegg

Vedlegg 1: PICO-Skjema benyttet til søk.

Patient (pasient)	**Intervention (Intervensjon)**	**Comparison (Sammenligning)**	**Outcome (utfall)**
Lung cancer	Deep Learning, CT (lavdose ct) Lung Screening		Accuracy