Bacheloroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for medisin og helsevitenskap
Institutt for sirkulasjon og bildediagnostikk

Elise Kraft Holmen
Ingrid Øfsti Brandsæter
Sunniva Flø

Convolutional Neural Networks i mammografi

Hvor godt presterer Convolutional Neural Network-modeller til klassifisering av mammogrammer, og hvilke faktorer påvirker klassifiseringsevnen?

Bacheloroppgave i Radiografi
Veileder: Beathe Sitter
Mai 2019

NTNU
Kunnskap for en bedre verden

Elise Kraft Holmen
Ingrid Øfsti Brandsæter
Sunniva Flø

Convolutional Neural Networks i mammografi

Hvor godt presterer Convolutional Neural Network-modeller til klassifisering av mammogrammer, og hvilke faktorer påvirker klassifiseringsevnen?

Bacheloroppgave i Radiografi
Veileder: Beathe Sitter
Mai 2019

Norges teknisk-naturvitenskapelige universitet
Fakultet for medisin og helsevitenskap
Institutt for sirkulasjon og bildediagnostikk

NTNU
Kunnskap for en bedre verden

Sammendrag

Hensikt: I Norge blir kvinner mellom 50 og 69 år invitert til mammografiscreening annethvert år. En av ulempene ved screening er falskt positive resultater. I dag er det to radiologer som tyder hvert mammogram og i fremtiden kan kanskje den ene radiologen erstattes med nevrale nett. Innenfor bildediagnostikken er det Convolutional Neural Network (CNN) som er best egnet til å tyde mammogrammer. I vår oppgave har vi sammenlignet elleve studier med ulike CNN-modeller. Vi har tatt utgangspunkt i Area Under the Curve (AUC)-verdien, da det gir en verdi på hvor godt modellen klassifiserer lesjoner. Vi har også diskutert hva CNN-modeller kan bidra med i mammografi.

Metode: Vi har utført en litteraturstudie for å svare på problemstillingen. PubMed, Scopus og Oria ble brukt til å finne fagfellevurderte artikler. Inklusjons- og eksklusjonskriterier har resultert i de elleve mest representative artiklene for vår problemstilling. Vi leste og analyserte dataene og trakk ut elementer som kan ha påvirket AUC-verdiene til CNN-modellene.

Resultat: AUC-verdiene varierte mellom 0.77 og 0.99. Det ser ut at det som har mest påvirkning på AUC-verdien er antall bilder og antall lag. Antall bilder som var inkludert i studiene varierte mellom 301 og 28 294, og vi ser at de med flest bilder har høyest AUC-verdi. Det er også forskjell på hvor mange lag de ulike modellene har. De med flest lag har generelt fått høyere AUC-verdier. Én studie har ikke brukt transfer learning og en annen har ikke validert. Disse studiene har relativt lav AUC-verdi.

Konklusjon: CNN-modellene presterer bedre enn radiologene. Vi tror at CNN-modeller kan erstatte den ene radiologen i fremtiden og bidra med å redusere antall tilbakekallinger. Det kan føre til reduserte kostnader, mindre arbeidsmengde for radiologer, og mindre psykisk belastning for pasienter.

Abstract

Purpose: All women aged 50 to 69 are invited to participate in the Norwegian Breast Cancer Screening Program biannually. One of the disadvantages with screening is high recall rates. Today, there are two radiologists who interpret each mammogram, and one of them might be replaced by a neural network in the future. Convolutional Neural Network (CNN) is the best model suited to interpret mammograms in digital imaging. In our study, we compared eleven studies with different CNN-models. We based our study on the Area Under the Curve (AUC) value, which evaluates how well the model classifies lesions. We also discussed how CNN-models can assist in mammography.

Methods: We decided to do a literature study to answer our thesis. PubMed, Scopus and Oria were used to find articles which were peer-reviewed. We set several criterias to include and exclude articles and therefore, we ended up with eleven studies representable for our thesis. We read and analysed the data, extracting the key elements which may have affected the AUC value to the CNN-models.

Results: The AUC values varies between 0.77 and 0.99. It appears to be the number of pictures and layers that has the biggest impact on the AUC value. The number of pictures used ranged from 301 to 28 294, and the studies using the biggest dataset got the highest AUC-value. Furthermore, this was also the case regarding layers. The studies with many layers, got higher AUC-values. One of the studies did not use transfer learning, and a second study did not validate. Both of these got low AUC-values.

Conclusion: CNN-models performed better then radiologists. We believe that CNN-models might replace one of the radiologists in the future, and contribute to reduce recall rates. A result of this is reduced medical cost, less workload for radiologists and less psychological stress to patients.

Innholdsfortegnelse

Innledning

Mammografiscreening

Brystkreft er den kreftsykdommen som rammer flest kvinner i Norge. I 2017 var det 3 589 nye tilfeller av brystkreft blant norske kvinner (Kreftregisteret, 2018b). Mammografiprogrammet i Norge ble innført som et pilotprosjekt i 1995 og ble nasjonalt i 2005 (Hofvind, 2017, s.7). Det ble utviklet med mål om å redusere dødeligheten av brystkreft. Alle kvinner mellom 50 og 69 år blir invitert til mammografiscreening annethvert år (Kreftregisteret, 2019b). En studie utført mellom 2007 og 2014 viste at gjennomsnittlig oppmøte per screeningrunde var 75% av inviterte kvinner (Sebuødegård, Sagstad, og Hofvind, 2016, s. 1449). Dødeligheten av brystkreft har gått ned etter oppstarten av mammografiprogrammet. I 1995 var det 789 kvinner som døde av brystkreft (Kreftregisteret, u.å.), kontra 623 kvinner i 2016 (Kreftregisteret, 2018a). Det er vanskelig å si om dette skyldes screening eller andre årsaker, som for eksempel bedre behandling (Kreftregisteret, u.å.).

En screeningradiolog må tyde 4000 mammografiundersøkelser av kvinner hvert år for å opprettholde kompetansen sin (Kreftregisteret, 2019a, s. 5). To radiologer vurderer mammogrammene uavhengig av hverandre (Kreftregisteret, 2017). Bildene har lite kontrast og dette gjør bildene vanskelige å tyde (Xi, Shu, og Goubran, 2018, s. 1). Det blir sett etter asymmetri, uregelmessige områder med økt tetthet, forkalkning og hudfortykkelser (Breastcancer.org, 2016). Bildene blir også sammenlignet med gamle bilder for å se etter endringer. Har kvinnene merket forandring selv, blir det notert. Bildene blir klassifisert med en score fra 1 til 5, hvor 1 er normal/benign og 5 er malign. Alle bildene som får en score ≥ 2 blir vurdert på et møte med begge radiologene og det blir bestemt om kvinnene skal bli tilbakekalt til etterundersøkelse (Kreftregisteret, 2019a, s. 9).

De største ulempene ved screening er overdiagnostisering, og falskt positive og falskt negative mammogrammer (Forskningsrådet, 2015, s. 33). Overdiagnostikk er når det behandles for brystkreft uten at kvinnen har en kreftsykdom som ville ført til plager, symptomer eller død (Hofmann, 2017). Falskt positive screeningresultat oppstår når kvinner blir tilbakekalt til etterundersøkelser som viser seg å være negative (Hofvind *et al.*, 2017, s.

26). Roman *et al*. (2013, s. 3952) fant at den kumulative risikoen var 20% for at kvinnene i mammografiprogrammet vil bli tilbakekalt på grunn av falskt positivt resultat minst en gang i løpet av ti screeningrunder. Når pasienten har kreft, men det ikke blir oppdaget på mammogrammet, er dette et falskt negativt resultat (Norsk forskningsråd, 2015, s. 31).

Kunstig intelligens

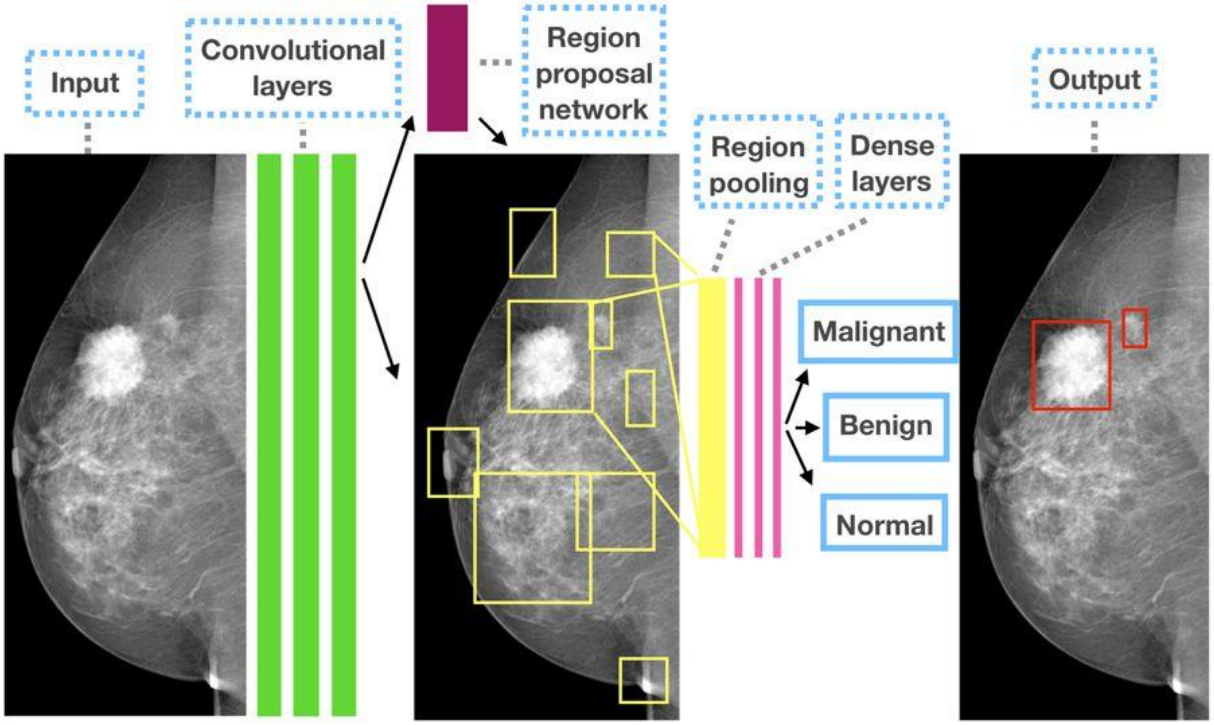
Kunstig intelligens (artificial intelligence, AI) er et vidt begrep som beskriver teknikker hvor man bruker en datamaskin til å utføre oppgaver med minimal menneskelig innblanding. Innenfor medisin er AI delt i to hovedgrupper; fysisk og virtuell. Den første gruppen inkluderer bruk av fysiske objekter, medisinsk utstyr og roboter. Den virtuelle gruppen, maskinlæring, omhandler blant annet nevrale nettverk. Dette er algoritmer som lærer av erfaring (Hamet og Tremblay, 2017, s. 37).

Computer Aided Detection (CAD) er en maskinlæringsteknikk som gjenkjenner mønstre og markerer mistenkelige strukturer på bilder. Det ble godkjent til klinisk bruk innenfor mammografi i USA i 1998 og blir brukt som hjelpemiddel innenfor røntgen, mammografi, CT, MR, ultralyd og PET (Song *et al*., 2015, s.1). Systemet utvikles ved at algoritmer trenes opp på et bildesett. Treningsprosessen inkluderer preprosessering, segmentering av anatomiske områder, egenskapsuttrekking og klassifisering. Det siste steget, klassifisering, er under utvikling og er ennå ikke i klinisk bruk (Katzen og Dodelzon, 2018, s. 305).

CAD ble utviklet for å redusere falskt negative funn og etter innføringen av systemet har deteksjon av kreft økt. Systemet kan erstatte radiolog nummer to, slik at kun én radiolog tolker bildet (Fathy og Ghoneim, 2019, s.175). Ulempen ved bruk av CAD er at den kan gi mange falskt positive markeringer som fører til flere tilbakekallinger (Katzen og Dodelzon, 2018, s. 307). Dette kan være en av grunnene til at CAD ikke har blitt innført i mange europeiske land, blant annet Norge. Her brukes det dobbelttyding istedenfor (Hofvind *et al.*, 2017, s.61).

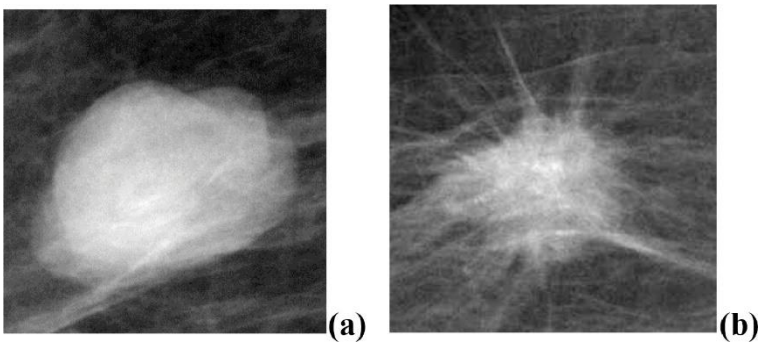
En undergruppe av maskinlæring er nevrale nettverk. Hovedstyrken til det nevrale nettverket er at det kan lære å organisere karaktertrekk fra et stort datasett, uten menneskelig påvirkning (Aboutalib *et al*., 2018, s. 5902). Nevrale nettverk er inspirert av

arkitekturen til menneskehjernen og består av flere lag; et input-lag, et eller flere skjulte lag og et output-lag (Figur 1). Det er forskjellige typer skjulte lag og de består av forskjellige funksjoner. Hvilken type funksjon laget har, avhenger av hvilken oppgave som skal løses og hvilken type data som blir brukt. (Nigam, 2018). Et nevralt nettverk som har mer enn ett skjult lag kalles dypt nevralt nettverk. Det finnes flere typer dype nevrale nettverk, for eksempel Recurrent Neural Network (RNN) og Convolutional Neural Network (CNN) (Nigam, 2018). (Aboutalib *et al*., 2018, s. 5902). En typisk oppbygging av en CNN-modell består av flere skjulte lag, som for eksempel konvolusjons-, pooling-, fully-connected- og normaliseringsfunksjoner (Nigam, 2018). Gjennom disse lagene trekkes det ut egenskaper fra et input-bilde og det produseres et output-bilde (Hosny *et al*., 2018, s. 501) hvor det avgis et resultat med et diagnoseforslag (Abildgaard *et al*., 2018, s.2). Det er altså CNN og ikke RNN som brukes når egenskaper skal trekkes ut fra bilder. RNN er laget for å se tilbake på tidligere beregninger og gjøre nye beregninger ut fra dette (Nigam, 2018). RNN tar hensyn til de tidligere lagene og er derfor ikke trent til å hente ut forskjellig informasjon ved de forskjellige lagene, slik CNN gjør. RNN er mest egnet for tekstdata, språkdata og klassifisering av forutsigbare problemer, og den er ikke egnet for tabelldata og bildedata. (Brownlee, 2018).



Figur 1: Eksempel på en oppbygning til en CNN-modell. Første mammogrammet er input-bildet. Deretter brukes konvolusjonslag, som består av ulike filtre. Neste mammogram viser lokalisasjon av objekter på bildet funnet av Region Proposal Network. Etter Region Pooling og Dense layers, produseres et output-bilde. (Ribli et al*, 2018).*

For at vi skal kunne bruke CNN-modeller til å kategorisere mammogrammer, må de trenes opp for å kjenne igjen strukturene i benigne og maligne lesjoner (Figur 2). Treningsprosessen for en CNN-modell krever store mengder med data, ofte flere tusen bilder, til trening, validering og testing (Carneiro, Nascimento og Bradly, 2015, s. 653). Konsekvensen av et for lite datasett kan være overfitting. Problemet med overfitting er at modellen blir for lik treningssettet, slik at modellens generaliserbarhet blir dårligere (Kim *et al.*, 2017, s.1274). Én måte å kompensere for et lite datasett på, er transfer learning. Dette er en metode hvor en modell som har blitt trent til én oppgave, brukes til å løse en annen lignende oppgave (Brownlee, 2017).

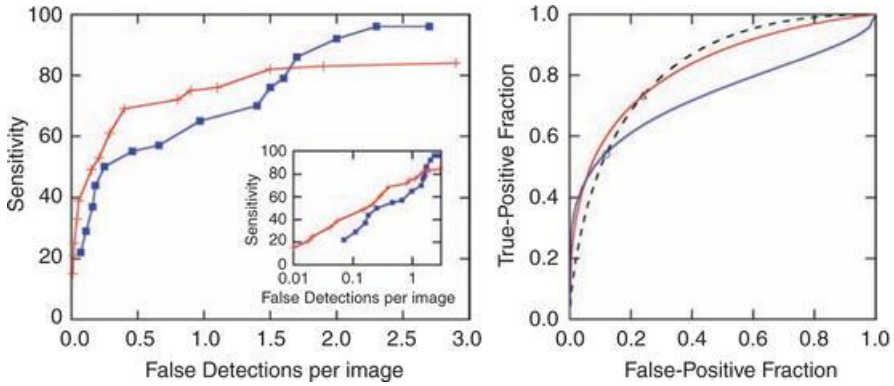


Figur 2: Et eksempel på en benign (a) og en malign (b) lesjon (Chougrad, Zouaki og Alheyane, 2018).

FROC- og ROC- kurver

Free Response Operating Characteristic (FROC) -kurver og Receiver Operating Characteristic (ROC) -kurver blir ofte brukt til å evaluere modeller (Figur 3). ROC-kurven representerer sammenhengen mellom sann-positiv-fraksjon/sensitivitet langs y-aksen og falskt-positiv-fraksjon langs x-aksen (Fathy og Ghoneim, 2019, s. 180). I en FROC-kurve plottes antall falskt positive bilder langs x-aksen. Ved en ROC-kurve får man to utfall, benign eller malign, mens ved en FROC-kurve viser flere utfall, fra ingen lesjon detektert til muligens flere lesjoner. En FROC-kurve gir dermed mer detaljer om mammogrammet (Nishikawa, 2010, s. 93).

Brattheten på kurven er viktig, da man ideelt sett ønsker å ha høyest mulig sann positiv rate og en minst mulig falskt positiv rate (Narkhede, 2018). Area Under the Curve (AUC) gir en verdi på hvor godt modellen klassifiserer. AUC-verdien varierer fra 0 til 1; jo høyere AUC-verdi, jo bedre er modellen på å skille mellom klassene (Ragab *et al.,* 2019, s. 10).



Figur 3: Eksempler på henholdsvis en FROC- og en ROC-kurve. X-aksen på ROC-kurven går fra 0-1 mens ved FROC-kurven opererer man med et annet intervall (Nishikawa, 2010).

Problemstilling

Hensikten med denne oppgaven er å se hvor godt Convolutional Neural Network-modeller presterer til klassifisering av mammogrammer, og hvilke faktorer som påvirker klassifiseringsevnen.

Metode

For å besvare problemstillingen vår har vi utført en litteraturstudie. Det har blitt gjort mye forskning på dette området fra før, noe som gjorde at vi kunne samle inn mye litteratur på kort tid. Mange studier gir forskjellige perspektiver, man kan sammenligne flere resultater og dermed se det i en større sammenheng.

For å finne fagfellevurderte artikler brukte vi tre databaser; PubMed, Scopus og Oria. Søkeordene vi brukte var: "Convolutional neural networks", "ROC curve" og "mammography" (Tabell 1). For å sikre relevansen på artiklene våre ble det satt inklusjonskriterier; artiklene skulle være fagfellevurdert, publisert i 2015 eller senere og måtte presentere en AUC-verdi.

Ved færre enn 50 treff leste vi gjennom tittel og sammendrag på alle treffene. Totalt ble dette 56 treff, som utgjorde 42 ulike artikler. Totalt fikk vi 29 potensielt relevante treff, hvorav 5 artikler ble funnet i to eller tre databaser. 21 ulike artikler ble lest i sin helhet.

Tretten artikler ble utelukket på grunn av at de omhandlet andre bildemodaliteter. Vi ekskluderte også ni artikler som omhandlet vevsforandringer og mikrokalsifikasjoner. Det var ni artikler som omhandlet risikopasienter, lokalisasjon av tumor eller klassifisering av brysttetthet som alle ble utelukket. I tillegg ekskluderte vi artikler som omhandlet kjemoterapi og andre nevrale nettverk og annen AI. Artikler som hadde annet fokus enn CNN-modeller, for eksempel utvidelse av datasett, bildeprosessering eller transfer learning ble også ekskludert. Vi ekskluderte én metodeartikkel. Én potensielt relevant artikkel ble utelukket da den ikke var tilgjengelig for NTNU og den var for dyr til å kjøpes. Totalt inkluderte vi elleve ulike artikler.

Tabell 1: Søkeskjema. Oversikt over databaser, søkeord og antall treff. Gir også oversikt over hvor mange artikler vi anså som potensielt relevante og de inkluderte artiklene, både fra hvert søk og totalt.

Søk	Søkeord	Treff	Potensielt relevant	Inkludert
Scopus				
1	"Convolutional neural networks" and mammography	162	n/a	n/a
2	"Convolutional neural networks" and "ROC curve" and mammography	24	11	8
PubMed				
1	"Convolutional neural networks" and mammography	22	10	5
2	"Convolutional neural networks" and "ROC curve" and mammography	5	4	2
Oria				
1	"Convolutional neural networks" and mammography	52	n/a	n/a
2	"Convolutional neural networks" and "ROC curve" and mammography	5	4	2
Totalt			**21**	**11**

På grunn av inklusjons- og eksklusjonskriteriene våre har vi funnet de elleve mest representative artiklene for vår problemstilling. Vi har brukt kriterier utarbeidet av Whittaker og Williamson (2011, s. 32-33) for å kvalitetssikre valg av artikler. Artiklene har et godt strukturert oppsett og begrunner valg av metode. Et tydelig metodekapittel gjør studiene reproduserbare og sammenlignbare, noe som styrker deres validitet. Flere nevner også styrker og svakheter med egen studie, noe som styrker deres reliabilitet. Alle studiene er retrospektive og det er derfor flere etiske aspekter som ikke trengs å ta hensyn til. For eksempel har ikke studiene trengt samtykke for å bruke mammogrammene (Aboutalib *et al.*, 2018, s. 5903).

I oppgaven vår har vi sammenlignet metode og resultat i elleve studier. Vi tok utgangspunkt i AUC-verdien for å vurdere modellene. Da vi analyserte dataene trakk vi ut elementer som kan ha påvirket AUC-verdiene til CNN-modellene. Det er ikke hensiktsmessig i denne oppgaven å gjøre statistiske analyser fordi resultatene fra artiklene kan ikke sammenlignes.

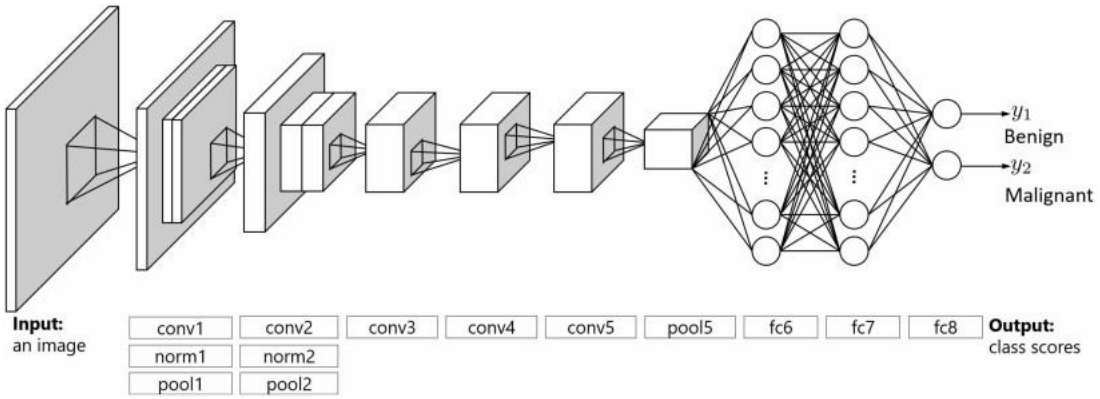
Resultat

Vi har valgt ut elleve artikler som bruker ulike CNN-modeller (tabell 2). AUC-verdiene i studiene varierer fra 0.77 til 0.99. Noen av studiene presenterte flere AUC-verdier og vi valgte da å bruke den høyeste verdien.

Tabell 2: Resultattabell. Oversikt over de elleve studiene angitt med forfatter og publikasjonsår, med hvilken CNN-modell som er brukt, antall lag, totalt antall bilder (N), transfer learning (ja/nei), validering - eventuelt kryssvalidering og AUC-verdi fra ROC-kurve. Studiene er listet fra høy til lav AUC-verdi. Én av studiene oppgir AUC-verdi fra FROC-kurve, denne er merket med stjerne.

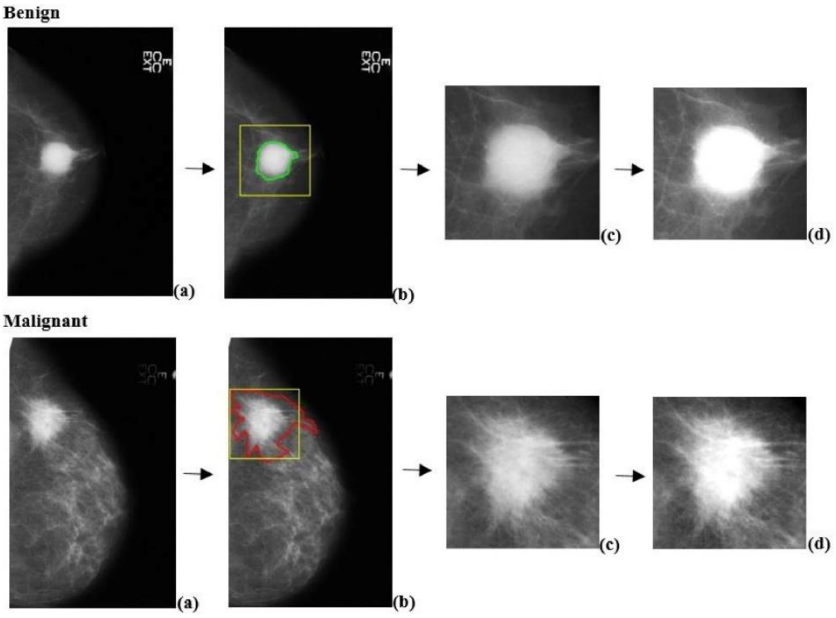
Forfatter og år	Modell	Lag	N	Transfer Learning	Validering	AUC
Chougrad (18)	Inception v3	221	6 229	Ja	Ja- kryss	0.99
Ribli (18)	VGG16	16	ca. 3 030	Ja	Ja	0.95
Ragab (19)	AlexNet	8	5 272	Ja	Ja- kryss	0.94
Hagos (18)	VGG-lignende	10	28 294	Ja	Ja	0.933*
Aboutalib (18)	AlexNet	8	9 648	Ja	Ja- kryss	0.77-0.96
Huynh (16)	AlexNet	8	607	Ja	Ja- kryss	0.86
Yemini (18)	Inception v3	Ikke oppgitt	410	Ja	Ja- kryss	0.86
Arevalo (16)	Egen modell	3	736	Nei	Ja	0.82
Wang (18)	AlexNet	8	301	Ja	Ja- kryss	0.813
Kooi (17)	VGG-lignende	9	1 804	Ja	Ja- kryss	0.8
Zhang (17)	AlexNet	8	2 263	Ja	Ikke oppgitt	ca 0.8

Fem av studiene brukte CNN-modellen AlexNet (Figur 4) og vi ser at disse generelt har lavere AUC-verdier. Det ser ut til at modellene med flest lag har høyere AUC-verdi. En studie har bygget sin egen modell og har ikke brukt transfer learning. Sju av studiene har brukt kryssvalidering.



Figur 4: Eksempel på arkitekturen til en CNN-modell. Zhang et al. (2017) har laget sin egen versjon av AlexNet (Zhang et al., 2017).

Studiene har hentet mammogrammer fra ulike databaser. Det er brukt ulikt antall bilder fra databasene, da studiene har ulike inklusjonskriterier. Antall bilder varierer fra 301 til 28 294. Sju av elleve studier har utvidet datasettet for å få flere bilder ved å lage kunstig data. Dette har blitt gjort på forskjellige måter, for eksempel ved å rotere bildene (Ragab *et al.*, 2019, s. 11), legge på støy (Chougrad, Zouaki og Alheyane, 2018, s. 21) eller trekke ut Region Of interest (ROI) (Arevalo *et al.*, 2016, s.251). De tre studiene av Wang *et al*. (2018), Zhang *et al.* (2017) og Aboutalib *et al.,* (2018) har ikke utvidet datasettet.



Figur 5: Eksempel på preprosessering. Første raden er en benign lesjon og den andre raden er en malign lesjon. (a) Er det originale mammogrammet (b) avgrenset lesjon (c) utvalgt ROI (d) ROI etter normalisering. (Chougrad, Zouaki og Alheyane, 2018)

Bildene har blitt bearbeidet ulikt i artiklene. Chougrad, Zouaki og Alheyane (Figur 5) (2018, s. 21) og Arevalo *et al.,* (2016, s. 251-252) har brukt Global Contrast Normalization (GCN) for å få bildene mer sammenlignbare. Wang *et al.* (2018, s. 2) har kombinert tre variasjoner av bildene for å lage pseudo-bilder. Zhang *et al. (*2017) nevner ingenting om bearbeiding av bildene.

Diskusjon

Metodediskusjon

For å sikre at søket vårt var dekkende valgte vi søkeord basert på problemstillingen vår. Vi brukte tre ulike databaser slik at vi fikk et større utvalg av artikler og satte inklusjonskriterier for å avgrense søket. Vi leste tittel og sammendrag på alle søkene som ga under 50 treff. Eksklusjonskriteriene sørget for at vi fikk utelukket uaktuelle artikler, slik at vi satt igjen med elleve artikler som dekket vår problemstilling. En praktisk fordel ved en litteraturstudie er at man ikke trenger tilgang til pasienter og dermed er det flere etiske aspekter som man ikke trenger å tas hensyn til.

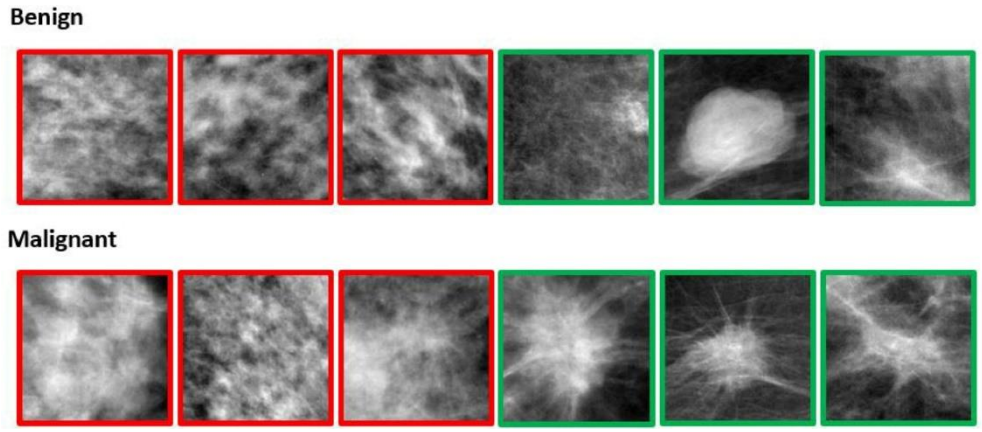
Vi har valgt å inkludere en studie som presenterer en AUC-verdi fra FROC-kurve, istedenfor fra ROC-kurve. Grunnen til at vi har inkludert denne artikkelen var at den har beskrevet hvordan den har bygget opp CNN-modellen sin, og dette var noe av det vi ønsket å undersøke i vår oppgave. Forskjellen på FROC- og ROC-kurver er at FROC-kurven representerer én eller flere lesjoner per bilde, mens ROC-kurven kun sier om det er sykdom på bildet eller ikke. Disse kurvene blir som regel brukt til å svare på forskjellige spørsmål. Noen ganger er ROC-kurver mest egnet mens andre ganger er det FROC-kurver (Hillis, Chakraborty og Orton, 2017, s. 1603). Selv om denne AUC-verdien kommer fra to ulike kurver, har vi valgt å behandle de likeverdig, og sammenlignet de.

Studiene hadde ikke samme tilnærming til utvikling av CNN-modeller. Modellene hadde ulik preprosessering, ble trent og testet på forskjellig sammensetning av bilder, hadde ulik oppbygning og forskjellig antall lag. Ulempen med disse forskjellene er at det ikke er mulig å gjennomføre en statistisk analyse og vi kan ikke si hvilken modell som er best. På en annen side kan forskjellene være en styrke for oppgaven vår. Grunnen til dette er at man får presentert et bredt spekter av CNN-modeller innenfor samme bildemodalitet, og man viser forskjellige måter å utvikle CNN-modeller på. Flere studier med samme metode og CNN-modell hadde styrket oppgaven, men dette er enda ikke mulig innenfor mammografi.

Diskusjon av resultat

AUC-verdiene i resultattabellen varierte mellom 0.77 og 0.99. Antall bilder inkludert i studiene varierte mye, og vi ser at de med flest bilder har høyest AUC-verdi. Det er også forskjell på hvor mange lag de ulike modellene har. De med flest lag har generelt fått høyere AUC-verdier.

For å oppnå en optimal modell må modellen trenes på et bredt utvalg av bilder. Utvalget bør være representativt og inneholde både tette, fettrike, store og små bryst. Det er vanskelig å tolke mammogrammer av tette bryst da disse ofte er utydelige. Brystmasse og tumorer er hvite eller lysegrå på mammogrammer, mens fettvev er svart eller mørkegrå (Figur 6). Utfordringen med tette bryst er at de har lite fettvev, det er derfor vanskelig å skille tumorer fra brystmasse (Kreftregisteret, 2019c). Det er kun Chougrad, Zouaki og Alheyane (2018, s. 29) som har diskutert utvalget sitt. De mener at de kunne valgt mer utfordrende bilder, for eksempel mammogrammer med tett brystvev. Kooi *et al.* (2017, s.1018) bruker bilder av kvinner som har blitt tilbakekalt til å teste modellen sin på. Ribli *et al.* (2018, s. 2) og Aboutalib *et al.* (2018, s. 5903) bruker også mammogrammer som har blitt tilbakekalt, til både trening og testing. Det kan være flere årsaker til at kvinner blir tilbakekalt, alt fra dårlig kvalitet, uerfaren radiolog til tette bryst, som gjør mammogrammene vanskeligere å tyde. Uansett årsak kan det være en fordel å bruke slike typer bilder, da modellen i praksis også vil bli brukt på utydelige bilder. Resultatene fra en modell som kun har blitt trent på tydelige bilder vil lettere oppnå en høy AUC-verdi, enn modeller trent på utydelige bilder. AUC-verdien vil dermed ikke være representativ for modellens ytelsesevne.



Figur 6: Øverste rad er benigne lesjoner og den andre raden er maligne lesjoner. Bildene med rød ramme er feilklassifisert og de med grønn ramme er riktig klassifisert. De som er feilklassifisert er bilder av tette bryst. (Chougrad, Zouaki og Alheyane, 2018)

Ytelsesevnen til dype nevrale nett er direkte proporsjonal med mengden data som er tilgjengelig for trening (Shokri og Shmatikov, 2015, s. 1310). Det er stor variasjon i hvor mange bilder som har blitt anvendt; fra 301 til 28 294. Dette er fordi de har brukt ulike databaser og inklusjonskriterier, i tillegg har de bearbeidet bildene forskjellig. Ved å forstørre datasettet kan man få høyere AUC-verdi (Zhang *et al.*, 2017, s. 796). Et større datasett gjør at modellen får mer data å trene på, slik at generaliserbarheten øker og modellen blir mer nøyaktig. Arfan Jaffar (2017, s. 287-288) utførte en studie hvor han utvidet datasettet ved å rotere og flippe bildene. Han økte dermed datasettet fra 2 122 bilder til 19 000 bilder. Huynh, Li og Giger (2016, s. 2) trakk ut ROIer fra mammogrammene for å få et større datasett. De som har forstørret datasettet har endt opp med flere bilder enn det som står i resultattabellen (Tabell 2). Det gjør det vanskelig å sammenligne AUC-verdiene basert på antall bilder. Aboutalib *et al.* (2018, s. 5905) fikk den høyeste AUC-verdien da de testet på det største datasettet og mener det kan være en sammenheng.

En av de største utfordringene ved trening av CNN-modeller er at det trengs store datasett (Wang *et al.*, 2018, s.2). Store datasett med mammogrammer er ikke tilgjengelig, og det brukes derfor transfer learning. Ti av elleve studier har brukt transfer learning, hvorav åtte har fortrent modellene på ImageNet, som er en database med ikke-medisinske/naturlige bilder (Aboutalib *et al.*, 2018, s. 5903). Kooi *et al.* (2017, s. 1019) mener det er en ulempe å trene modeller på naturlige bilder da bildene er ulike fra mammogrammer og har fortrent modellen på screening-mammogrammer. Selv om modellen er fortrent på mammogrammer

har de fått en av de laveste AUC-verdiene. Det er vanskelig å si om dette skyldes bildene modellen er fortrent på, antall mammogrammer eller om det er andre årsaker. Arevalo *et al.* (2016) som ikke har brukt transfer learning har fått en av de laveste AUC-verdiene.

Studiene har utført ulik preprosessering på datasettene. Preprosessering gjøres for å forbedre bildet ved å fremheve dets egenskaper, som gjør at modellen presterer bedre (Arevalo *et al*., 2016, s. 251). Ved å preprosessere tas det vekk små variasjoner, og man får frem de forskjellene man er interessert i. En type preprosessering som blir brukt i studiene er normalisering. Det finnes flere måter man kan utføre normalisering på. Både Chougrad, Zouaki og Alheyane (2018) og Arevalo *et al.* (2016) har brukt Global Contrast Normalization (GCN). Dette brukes for å fjerne variasjoner i bildene og gjøre de mer like. Arevalo *et al*. (2016, s. 251) bruker GCN for å rette opp i lyset på bildene, da pikselverdiene blir påvirket av digitaliseringsprosessen. Ved normalisering regnes det ut et gjennomsnitt av intensiteten for hvert bilde, før denne trekkes fra hver piksel i bildet. Vi ser at Chougrad, Zouaki og Alheyane (2018) har fått en høyere AUC-verdi enn Arevalo *et al*. (2016), selv om preprosesseringen er relativt lik. Arevalo *et al.* (2016, s.254) og Kooi *et al.* (2017, s.1022) viser at AUC-en blir dårligere uten normalisering.

Studiene har brukt ulike CNN-modeller. Fem av de har brukt AlexNet, som bruker åtte lag. Disse studiene har generelt lavest AUC-verdi. AlexNet har blitt brukt til tross for at det finnes andre mer avanserte modeller som er bedre for transfer learning. Dette er en enkel modell som det finnes mye litteratur om. Valg av en annen modell kunne derfor resultert i høyere AUC-verdi (Huynh, Li og Giger, 2016, s.4). Tre studier har brukt ulike varianter av VGG, hvorav VGG16 har flest lag og høyest AUC-verdi blant VGG-modellene. Chougrad, Zouaki og Alheyane (2018, s. 22) har brukt modellen Inception v3 med 221 lag, og har scoret høyest med en AUC-verdi på 0.99. Yemini, Zigel og Lederman (2018, s. 1) har også brukt Inception v3, men det er ikke oppgitt antall lag. Zhang *et al.* (2017, s. 296) tror at dypere nettverk kan gi høyere AUC-verdi. Dette stemmer med Chougrad, Zouaki og Alheyane (2018, s. 28) sitt resultat, hvor det dypeste nettverket fikk den høyeste AUC-verdien.

Ni av elleve studier har brukt fortrente modeller i sin forskning. Ifølge Huynh, Li og Giger (2016, s. 4) er dette en ulempe, da muligheten for å endre på arkitekturen på modellen er

begrenset. Som et resultat av dette, har studiene nedskalert bildet for å passe til modellens input. Ved å nedskalere bildet får man dårligere oppløsning, noe som kan gi dårligere resultat (Arevalo *et al.*, 2016, s. 255).

Datasettet deles i trening, validering og testing for å evaluere generaliserbarheten til modellen (Hagos, Mérida, og Teuwen, 2018, s. 92). Validering er en del av kvalitetssikringen av en modell (Wibetoe, 2018). Studiene har brukt forskjellige måter å validere modellene på. Sju studier har brukt kryssvalidering enten i form av k-fold Cross Validation eller Leave-One-Case-Out (LOCO) Validation. Ifølge Huynh *et al.* (2016, s. 3) er 5-fold Cross Validation foretrukket for CAD. Hagos, Mérida, og Teuwen (2018, s. 92) og Arevalo *et al.* (2016, s. 253) har brukt 10% av datasettet til validering. Zhang *et al.* (2017) er den eneste studien som ikke har nevnt validering og denne studien har fått lavest AUC-verdi.

Ni av studiene har brukt et eget datasett til testing, men ikke alle oppgir størrelsen på dette datasettet. Arevalo *et al.* (2016, s. 253) og Hagos, Mérida og Teuwen (2018, s. 92) har brukt 40% av datasettet til testing, Chougrad, Zouaki og Alheyane (2018, s. 26) har brukt 20% mens Aboutalib *et al.* (2018, s. 5903) har kun brukt 5%. Wang *et al.* (2018, s. 4) har ikke brukt et eget datasett, men laget et gjennomsnitt av resultatene etter valideringen til å sette en AUC-verdi. Testing på et eget datasett gjør modellen mer pålitelig da dette er nye bilder for modellen. Det er dette som vil fortelle hvor god den vil være i praksis.

Det er vanskelig å si hvilke av modellene som er best, da alle har utført studiene forskjellig. Av de elleve mener vi det er Chougrad, Zouaki og Alheyane (2018), Aboutalib *et al.* (2018) og Ragab *et al.* (2019) som har utviklet de beste modellene. Henholdsvis har de fått AUC-verdiene 0.99, 0.77-0.96 og 0.94. Studiene har store datasett og de har brukt separate datasett til trening og testing noe som gjør modellene mer robuste. Aboutalib *et al.* (2018, s. 5903) og Ragab *et al.* (2019 s. 1) har brukt AlexNet som består av 8 lag, mens Chougrad, Zouaki og Alheyane (2018, s. 22) har brukt Inception v3 med 221 lag. Det kan være en av disse faktorene som har ført til forskjell i AUC-verdi. En annen fordel med disse modellene er at alle har brukt kryssvalidering, som er den beste valideringsmetoden på CNN-modeller.

Veien videre

Mangel på nøyaktige data på radiologers prestasjon i litteraturen hindrer oss i å gjøre en god sammenligning mellom radiologer og CNN-modeller (Aboutalib *et al*., 2018, s. 5907). Det kan være stor variasjon i prestasjonen til radiologer, avhengig av erfaring og opplæring. Cole *et al.* (2014, s. 909) utførte en studie hvor de sammenlignet radiologers prestasjon med to ulike CAD systemer, hvorav radiologene fikk en AUC-verdi på 0.71 uten CAD og 0.72 med CAD. CNN-modellene vi har sammenlignet har fått en AUC-verdi på 0.77 eller høyere. Dette viser at modellene presterer bedre enn radiologer, når de trenes på akkurat de bildene som er brukt i studiene. Ved å videreutvikle modellene med for eksempel større datasett og mer representative bilder, kan man se for seg at CNN-modeller kan bli et hjelpemiddel for radiologer i fremtiden.

En av ulempene med screeningprogrammet er tilbakekallinger. Den kumulative falskt positive risikoen er 20% (Roman *et al.*, 2013, s. 3952). CAD har som intensjon å redusere falskt negativt funn, men dette har muligens økt antall falskt positive funn (Hofvind *et al.,* 2017, s. 60). I diagnostisk mammografi er det verre å overse kreft, enn å ta en biopsi av en lesjon som er benign (Nishikawa, 2010, s. 95). Høye falskt positive rater har vært hovedargumentet mot CAD, men man har sett en forbedring i løpet av de siste årene. Fazal *et al.* (2018, s. 248) presenterer tall på sensitivitet på ulike CAD-modeller som viser at sensitiviteten har økt fra 70% i 1999 til 85-90% i 2016. Høyere sensitivitet betyr mindre sannsynlighet for falskt positive resultat (Malt og Stoltenberg, 2017). Studiene som tallene refererer til ble gjort med CAD brukt på lunge-CT, men man kan forvente at den samme forbedringen er gjort innenfor mammografi. Ribli *et al*. (2018, s. 5) mener modellen de har utarbeidet vil kunne ta over for tradisjonell CAD. Ved å kombinere CNN-modeller med CAD, kan man forbedre resultater og dermed redusere antall falskt positive funn ytterligere. CNN og CAD jobber med å redusere falskt positive funn, men det er viktig at antallet falskt negative funn holdes lavt.

Obermeyer og Emanuel (2016, s. 1218) mener at kunstig intelligens vil ta over mye av radiologenes arbeid. Andre mener at selv om AI vil overgå menneskelig intelligens er det usannsynlig at det vil erstatte radiologene helt (Hosny *et al.*, 2018, s. 508). Fremtiden til CNN og CAD kan være å assistere radiologene, for eksempel ved å erstatte den ene radiologen.

Fazal *et al*., (2018, s. 249) presenterer to måter CAD kan assistere radiologer på i fremtiden. Ved den første metoden vil radiologen først tolke mammogrammet uten hjelp av CAD. Deretter brukes funnene til CAD og radiologen gir en diagnose basert på kombinasjonen av dette. Ved den andre metoden gir CAD funnene først, deretter tolker radiologen bildet og setter en diagnose.

Ved å bruke CNN-modeller kan sannsynligheten for at noe blir oversett, reduseres. Ulempene er at man setter for mye lit til systemet og blir mer avhengig av teknologi. Når CNN-modellen finner en diagnose først, kan det være lett å stole for mye på dette resultatet og man kan overse noe som burde vært oppdaget. Dette kan føre til at radiologenes ferdigheter reduseres (Cabitza, Rasoini og Gensini, 2017, s. 517).

CNN-modeller kan føre til at radiologene presterer bedre da de får hjelp til tolkingen. Dette kan bidra til redusering av falskt positive screeningresultater. Resultatet av dette er mindre arbeidsmengde for radiologene, reduserte kostnader og mindre psykisk belastning hos pasientene (Chougrad, Zouaki og Alheyane, 2018, s. 19). Arbeidsmengden til radiologene vil kunne gå ned ved hjelp av CNN-modeller. Når CNN-modeller erstatter den ene radiologen blir det mindre arbeidsoppgaver og i tillegg vil færre falskt positive resultat også frigi tid. CAD vil være dyrt å innføre, men har de siste årene vært i stor utvikling. Om man kombinerer CAD og CNN-modeller kan systemet kanskje bli så bra at det er verdt kostnadene. Mange tilbakekallinger gir høye medisinske kostnader (Aboutalib *et al*., 2018, s. 5902) og CNN-modeller kan redusere disse ved reduksjon av tilbakekallinger. Det vil også føre til at færre kvinner må gå gjennom den psykiske belastningen som tilbakekalling kan påføre.

CNN-modeller blir ofte betegnet som Black Box, da man ikke vet hva som skjer mellom input og output, altså i de skjulte lagene. Man kan ikke se hvordan maskinen har kommet til én spesifikk konklusjon, noe som gjør det vanskelig å vite hvor det skjer feil. Det er vanskelig å vite om maskinen er generaliserbar på tvers av maskinvare, protokoller og pasientgrupper (Hosny *et al.*, 2018, s. 507). Det hadde vært nyttig for radiologer å vite hvordan maskinen skiller mellom ulike grupper, for eksempel mellom benign og malign, slik at de kan bruke det i sitt arbeid. Med Black Box vet man ikke hva som skjer med personopplysninger under

trening av modellene. Innsamlede data kan ikke slettes og brukerne kan ikke kontrollere hvordan informasjonen blir brukt.

Siden man ikke kan se hvordan maskinen har kommet frem til en konklusjon er det vanskelig å si hvem som skal ha ansvaret for feil i diagnostiseringen. På den ene siden er det radiologene som har ansvaret for å tolke bildene og gi en riktig diagnose. Selv om de baserer vurderingene på CAD og CNN-modellen er det radiologene som tar den siste beslutningen. På den andre siden gir leverandørene garantier for at systemet fungerer slik det skal. I USA må alle CAD-systemer få klarering fra the Food and Drug Administration (FDA) før de kan brukes. De har blant annet ansvaret for å sikre at alt medisinsk utstyr er forsvarlig (The Food and Drug Administration, 2018). Det er en sikkerhet i at det må gis en godkjenning før systemene tas i bruk. Dersom CAD og CNN-modeller tas i bruk i Norge må det avklares hvem som skal stå ansvarlig dersom feil oppstår.

Konklusjon

I vår oppgave har vi sett på bruken av elleve ulike CNN-modeller innenfor mammografi. Vi tok utgangspunkt i AUC-verdiene og sammenlignet oppbygningen til modellene. Alle studiene fikk en høy AUC-verdi, 0.77-0.99, til tross for at de har brukt ulike fremgangsmåter. Det ble anvendt fire ulike modeller, og lagene varierte fra 3 til 221. Det var også store forskjeller på størrelsen på datasettene, fra 301 til 28 294 bilder. Disse to faktorene ser ut til å ha påvirket AUC-verdien mest. Utvalget av bilder påvirker resultatet; tydelige bilder vil kunne gi en høy AUC-verdi, men verdien vil ikke være representativ for utydelige bilder. Det er derfor viktig at det blir brukt representative bilder for at modellen skal bli robust.

Studiene har fått høye AUC-verdier på tross av ulike metoder, og dette tyder på at bruken av CNN-modeller til tolkning av mammogrammer er relativt robust. CNN-modellene i denne oppgaven presterer bedre enn radiologer og dette viser at bruken av CNN-modeller i mammografi er lovende. Ved videreutvikling av CNN-modeller kan de få stor innvirkning på radiologenes arbeidsliv. I Norge brukes dobbelttyding og vi tror at CNN-modeller kan erstatte den ene radiologen i fremtiden. CNN-modeller kan bidra med å redusere antall tilbakekallinger. Det kan føre til reduserte kostnader, mindre arbeidsmengde for radiologer og dermed bedre arbeidsflyt, og mindre psykisk belastning for pasienter.

Referanseliste

Abildgaard, A. *et al.* (2018) Vil radiologer bli erstattet av kunstig intelligens? *Tidsskrift Norsk Legeforening*, s. 1-5. http://doi.org/10.4045/tidsskr.18.0587

Aboutalib, S. S. *et al.* (2018) Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening. *Clinical Cancer Research*, *24*(23), s. 5902–5909. https://doi.org/10.1158/1078-0432.CCR-18-1115

Arevalo, J. *et al.* (2016) Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, *127*, s. 248–257. https://doi.org/10.1016/j.cmpb.2015.12.014

Arfan, M. (2017) Deep Learning based Computer Aided Diagnosis System for Breast Mammograms. *International Journal of Advanced Computer Science and Applications*, *8*(7), s. 286-290. https://doi.org/10.14569/IJACSA.2017.080738

Breastcancer.org. (2016) How doctors Interpret Mammograms. Tilgjengelig fra: https://www.breastcancer.org/symptoms/testing/types/mammograms/interpret (Hentet: 26. februar 2019)

Brownlee, J. (2017) *A Gentle Introduction to Transfer Learning for Deep Learning*. Tilgjengelig fra: https://machinelearningmastery.com/transfer-learning-for-deep-learning/ (Hentet: 29. April 29 2019)

Brownlee, J. (2018) *When to Use MLP, CNN, and RNN Neural Networks*. Tilgjengelig fra: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/ (Hentet: 21. februar 2019)

Cabitza, F., Rasoini, R. og Gensini, G. F. (2017) Unintended Consequences of Machine Learning in Medicine. *JAMA*, *318*(6), s. 517–518. https://doi.org/10.1001/jama.2017.7797

Carneiro G., Nascimento J. og Bradley A.P. (2015) Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models. In: Navab N., Hornegger J., Wells W., Frangi A. (eds) *Medical Image Computing and Computer-Assisted Intervention* – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Sted: Springer, s. 652-660.

Chougrad, H., Zouaki, H. og Alheyane, O. (2018) Deep Convolutional Neural Networks for breast cancer screening. *Computer Methods and Programs in Biomedicine*, 157, s. 19–30. https://doi.org/10.1016/j.cmpb.2018.01.011

Chougrad, H., Zouaki, H. og Alheyane, O. (2018) Fig. 1. Samples of mammography mass lesions [digitalisert bilde]. Tilgjengelig fra: https://doi.org/10.1016/j.cmpb.2018.01.011 (Hentet 15. mai 2019)

Chougrad, H., Zouaki, H. og Alheyane, O. (2018) Fig. 2. Pre-processing of the mammograms [digitalisert bilde]. Tilgjengelig fra: https://doi.org/10.1016/j.cmpb.2018.01.011 (Hentet 15. mai 2019)

Chougrad, H., Zouaki, H. og Alheyane, O. (2018) Fig. 9. Examples of regions of interest containing mass lesions [digitalisert bilde]. Tilgjengelig fra: https://doi.org/10.1016/j.cmpb.2018.01.011 (Hentet 15. mai 2019)

Cole, E. B. *et al.* (2014) Impact of Computer-Aided Detection Systems on Radiologist Accuracy With Digital Mammography, *American Journal of Roentgenology Diagnostic Imaging and Related Sciences*, 203 (4), s. 909-916. http://doi.org/10.2214/AJR.12.10187

Fathy, W. E. og Ghoneim, A. S. (2019) A Deep Learning Approach for Breast Cancer Mass Detection. *International Journal of Advanced Computer Science and Applications*, 10(1), s. 175-182. https://doi.org/10.14569/IJACSA.2019.0100123

Fazal, M.I. *et al.* (2018) The past, present and future role of artificial intelligence in imaging, *European Journal of Radiology*, *2018* (105), s. 246-250. https://doi.org/10.1016/j.ejrad.2018.06.020

Forskningsrådet (2015) *Research-based evaluation of the Norwegian Breast Cancer Screening Program*. Oslo: The Research Council of Norway. Tilgjengelig fra: https://www.regjeringen.no/contentassets/444d08daf15e48aca5321f2cefaac511/mammografirapport-til-web.pdf (Hentet: 25.februar 2019)

Hagos Y. B., Mérida A. G. og Teuwen J. (2018) Improving Breast Cancer Detection Using Symmetry Information with Deep Learning. In: Stoyanov D. *et al.*, (eds) *Image Analysis for Moving Organ, Breast, and Thoracic Images*. RAMBO 2018, BIA 2018, TIA 2018. Lecture Notes in Computer Science, vol 11040. Sted: Springer, s. 90-97

Hamet, P. og Tremblay, J. (2017) Artificial intelligence in medicine. *Metabolism*, *69*, s. 36–40. https://doi.org/10.1016/j.metabol.2017.01.011

Hillis, S. L., Chakraborty, D.P., og Orton, C.,G. (2017) ROC or FROC? It depends on the research question, Medical Physics, 44(5), s. 1603-06. https://doi.org/10.1002/mp.12151

Hofmann, B. (2017) Overdiagnostikk, *Store Norske Leksikon.* Tilgjengelig fra: http://sml.snl.no/overdiagnostikk (Hentet: 21. januar 2019)

Hofvind, S. *et al*. (2017) *The Norwegian Breast Cancer Screening Program, 1996-2016: Celebrating 20 years of organised mammographic screening* (Cancer in Norway 2016 - Special Issue) Oslo: Cancer Registry of Norway. Tilgjengelig fra https://www.kreftregisteret.no/globalassets/cancer-in-norway/2016/mammo_cin2016_special_issue_web.pdf (Hentet: 1.april 2019)

Hosny, A. *et al*. (2018) Artificial intelligence in radiology. *Nature Reviews. Cancer*, 18, s. 500–510. https://doi.org/10.1038/s41568-018-0016-5

Huynh, B. Q., Li, H. og Giger, M. L. (2016) Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3), s. 1-5. https://doi.org/10.1117/1.JMI.3.3.034501

Katzen, J. og Dodelzon, K. (2018) A review of computer aided detection in mammography. *Clinical Imaging*, 52, s. 305–309. https://doi.org/10.1016/j.clinimag.2018.08.014

Kim, Y. *et al*. (2017) Avoiding Overfitting in Deep Neural Networks for Clinical Opinions Generation from General Blood Test Results. *Precision Healthcare through Informatics*, (245), s. 1274 https://doi.org/10.3233/978-1-61499-830-3-1274

Kreftregisteret (u.å.) *Færre dør av brystkreft*. Tilgjengelig fra: https://www.kreftregisteret.no/Generelt/Nyheter/Farre-dor-av-brystkreft/ (Hentet: 21.mai 2019)

Kreftregisteret (2017) *Hva kan mammografi screening innebære?* Tilgjengelig fra: https://www.kreftregisteret.no/globalassets/mammografiprogrammet/informsjonsmateriell/2017-desember/faktaark_bokmal.pdf (Hentet: 26. februar 2019)

Kreftregisteret (2018a) *Brystkreft.* Tilgjengelig fra: https://www.kreftregisteret.no/Generelt/Fakta-om-kreft/Brystkreft-Alt2/ (Hentet: 21. mai 2019)

Kreftregisteret (2018b) *Cancer in Norway 2017 - Cancer incidence, mortality, survival and prevalence in Norway*. Oslo: Cancer Registry of Norway, 2018. Tilgjengelig fra https://www.kreftregisteret.no/globalassets/cancer-in-norway/2017/cin-2017.pdf (Hentet: 28. mars 2019)

Kreftregisteret (2019a) *Kvalitetsmanual i Mammografiprogrammet, Radiologi*. Tilgjengelig fra: https://www.kreftregisteret.no/globalassets/mammografiprogrammet/rapporter-og-publikasjoner/20190327_kvalitetsmanual-radiologer_hele.pdf (Hentet: 16. mai 2019)

Kreftregisteret (2019b) *Mammografiprogrammet*. Tilgjengelig fra: https://www.kreftregisteret.no/screening/Mammografiprogrammet/ (Hentet: 26. februar 2019)

Kreftregisteret (2019c), *Mammografisk tetthet.* Tilgjengelig fra: https://www.kreftregisteret.no/screening/Mammografiprogrammet/Tetthet/ (Hentet: 20. mai 2019)

Kooi, T. *et al.* (2017) Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network, *Medical Physics*, 44(3), s. 1017–1027. https://doi.org/10.1002/mp.12110

Malt, U. og Stoltenberg, C. (2017) Sensitivitet - test, *Store Norske Leksikon*. Tilgjengelig fra: https://snl.no/sensitivitet_-_test (Hentet 1. mai 2019)

Narkhede, S. (2018) *Understanding AUC - ROC Curve*. Tilgjengelig fra: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (Hentet: 23. februar 2019)

Nigam, V. (2018) *Understanding Neural Networks. From neuron to RNN, CNN, and Deep Learning.*)Tilgjengelig fra: https://towardsdatascience.com/understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90 (Hentet: 21. februar 2019)

Nishikawa, R. M. (2010) Computer-aided Detection and Diagnosis. In U. Bick & F. Diekmann (Eds.), *Digital Mammography*, Medical Radiology. Sted: Springer, s. 85-106. https://doi.org/10.1007/978-3-540-78450-0_6

Nishikawa, R. M. (2010) Fig. 6.5. A comparison of FROC and ROC curves [digitalisert bilde]. Tilgjengelig fra: https://link.springer.com/chapter/10.1007%2F978-3-540-78450-0_6 (Hentet: 15. mai 2019)

Norsk forskningsråd (2015) *Research-based evaluation of the Norwegian Breast Cancer Screening Program: Final report*. Tilgjengelig fra: https://www.kreftregisteret.no/globalassets/mammografiprogrammet/forskning/combinesiste-1.pdf (Hentet: 21. februar 2019)

Obermeyer, Z. og Emanuel, E. J (2016) Predicting the Future - Big Data, Machine Learning, and Clinical Medicine, *The New England Journal of Medicine,* 375, s. 1216-1219. https://doi.org/10.1056/NEJMp1606181

Ragab, D. A. *et al.* (2019) Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ, 7,* e6201, s. 1-23. https://doi.org/10.7717/peerj.6201

Ribli, D. *et al.* (2018) Detecting and classifying lesions in mammograms with Deep Learning. *Scientific Reports, 8(1),* s. 1-7. https://doi.org/10.1038/s41598-018-22437-z

Ribli, D. *et al.* (2018) Fig. 1: The outline of the Faster R-CNN model for CAD in mammography [digitalsert bilde]. Tilgjengelig fra: https://doi.org/10.1038/s41598-018-22437-z (Hentet 15. mai 2019)

Roman, M. *et al.* (2013) The cumulative risk of false-positive results in the Norwegian Breast Cancer Screening Program: Updated results. *Cancer,* 119(22), s. 3952–3958. https://doi.org/10.1002/cncr.28320

Sebuødegård, S., Sagstad, S. og Hofvind, S. (2016) Oppmøte i Mammografiprogrammet. *Tidsskrift Norsk Legeforening*, 136, s. 1448-51. https://doi.org/10.4045/tidsskr.15.1013

Shokri, R. og Shmatikov, V. (2015) Privacy-Preserving Deep Learning. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security,* Denver, Colorado, USA, 12.-16. Oktober 2015, New York: Association for Computing Machinery, s. 1310–1321. https://doi.org/10.1145/2810103.2813687

Song, S. E. *et al.* (2015) Computer-aided detection (CAD) system for breast MRI in assessment of local tumor extent, nodal status, and multifocality of invasive breast cancers: preliminiary study. *Cancer Imaging 2015,* 15(1), s. 1-9. https://doi.org/10.1186/s40644-015-0036-2

The Food and Drug Administration (2018), *What We Do.* Tilgjengelig fra: https://www.fda.gov/about-fda/what-we-do (Hentet: 21. mai 2019)

Wang, Y. *et al.* (2018) A hybrid deep learning approach to predict malignancy of breast lesions using mammograms, *SPIE Medical Imaging.* Houston, Texas, USA, 6. mars 2018. s. 1-6. https://doi.org/10.1117/12.2286555

Wibetoe, G. (2018) Validering, *Store Norske Leksikon*. Tilgjengelig fra : https://snl.no/validering (Hentet 1.april 2019)

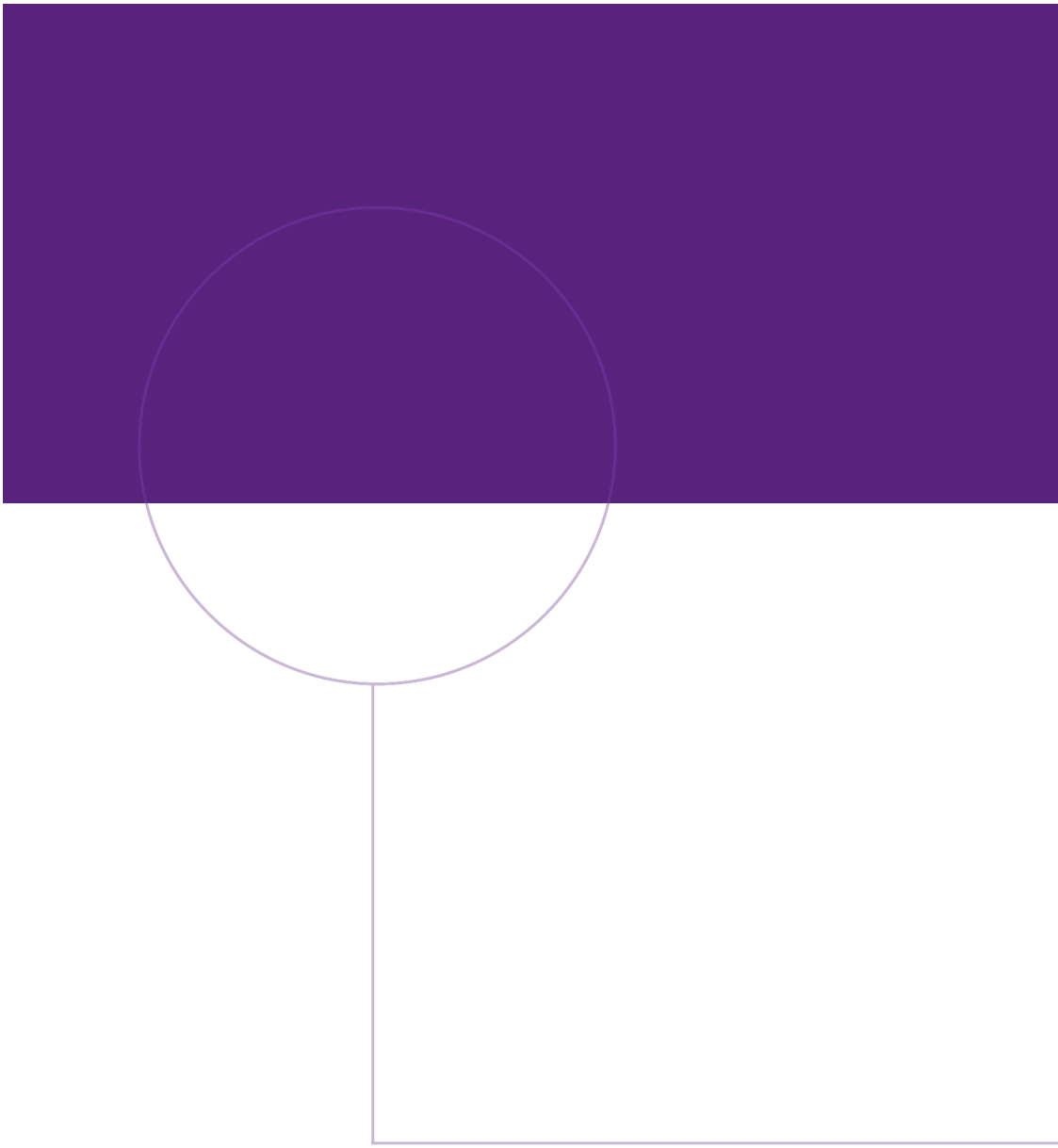
Whittaker, A. og Williamson G. R. (2011) *Succeeding in Research Project Plans and Literature Reviews for Nursing Students*. Exeter: Learning Matters

Xi, P., Shu, C. og Goubran, R. (2018) Abnormality Detection in Mammography using Deep Convolutional Neural Networks. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* Roma, Italia, 11.-13. Juni 2018, IEEE s.1–6 https://doi.org/10.1109/MeMeA.2018.8438639 .

Yemini, M., Zigel, Y. og Lederman, D. (2018) Detecting Masses in Mammograms using Convolutional Neural Networks and Transfer Learning, *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)* Israel, Desember 2018, IEEE, s. 1-4, https://doi.org/10.1109/ICSEE.2018.8646252

Zhang, X. *et al.* (2017) Classification of mammographic masses by deep learning. *56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)* Kanazawa, Japan, 19-22 September 2017, IEEE s. 793–796. https://doi.org/10.23919/SICE.2017.8105545

Zhang, X. *et al.* (2017) Fig. 2: Architecture of the alexnet [digitalisert bilde]. Tilgjengelig fra: https://ieeexplore.ieee.org/document/8105545 (Hentet: 15. mai 2019)



NTNU

Kunnskap for en bedre verden