Høgskulen på Vestlandet

BRA330 -
Bacheloroppgave

BRA330

Predefinert informasjon

Startdato:	24-04-2020 09:00	**Termin:**	2020 VÅR
Sluttdato:	18-05-2020 14:00	**Vurderingsform:**	Norsk 6-trinns skala (A-F)
Eksamensform:	Bacheloroppgave	**Studiepoeng:**	15
SIS-kode:	203 BRA330 1 O 2020 VÅR Bergen		
Intern sensor:	(Anonymisert)		

Deltaker

Kandidatnr.: 333

Informasjon fra deltaker

Antall ord *: 9163

Egenerklæring *: Ja

Inneholder besvarelsen konfedensielt materiale?: Nei

Jeg bekrefter at jeg har registrert oppgavetittelen på norsk og engelsk i StudentWeb og vet at denne vil stå på vitnemålet mitt *: Ja

Gruppe

Gruppenavn: (Anonymisert)

Gruppenummer: 2

Andre medlemmer i gruppen: 321

Jeg godkjenner avtalen om publisering av bacheloroppgaven min *

Ja

Er bacheloroppgaven skrevet som del av et større forskningsprosjekt ved HVL? *

Nei

Er bacheloroppgaven skrevet ved bedrift/virksomhet i næringsliv eller offentlig sektor? *

Nei

Høgskulen på Vestlandet

BACHELOROPPGAVE

Kunstig intelligens i medisinsk bildediagnostikk

Artificial intelligence in medical imaging

Kandidatnummer: 321 og 333

Radiografutdanning R17

Fakultet for helse- og sosialvitenskap FHS, Institutt for helse og funksjon, bachelor i Radiografi.

Veileder: Sundaran Kada

18/05/2020

Antall ord: 9163

Jeg bekrefter at arbeidet er selvstendig utarbeidet, og at referanser/kildehenvisninger til alle kilder som er brukt i arbeidet er oppgitt, *jf. Forskrift om studium og eksamen ved Høgskulen på Vestlandet, § 12-1.*

FORORD

Denne bacheloroppgaven er skrevet som avsluttende oppgave ved bachelorutdanningen i radiografi, ved fakultet for helse- og sosialvitenskap, under institutt for helse og funksjon på Høgskulen på Vestlandet.

Vi ønsker å takke vår veileder Sundaran Kada og sivilingeniørstudent Joakim Nyland for gode råd, veiledning og konstruktive tilbakemeldinger.

God Lesning.

BEGREPSFORKLARING/FORKORTELSER:

MRI/ MR – Magnetic Resonance Imaging/ Magnetisk Resonanstomografi

AI/KI - Artificial Intelligence/ Kunstig Intelligens

DL – Deep Learning/ Dyp Læring

ML – Machine learning/ Maskinlæring

ANN/ KNN - Artificial neural networks/ kunstig nevralt nettverk

CNN - Konvensjonelle nevrale nettverk

ROC – Reciver Operating Characteristic Curve

Tumor – En ansamling av en masse (kul, knute eller fortykkelse og kan skyldes mange ulike prosesser), kan være godartet og ondartet. Ved ondartet tumor, kalles det en svulst.

Caput – Medisinske (latin) termen for hode

Node – kontaktpunkt, mellom et kunstig nevron og synapse

Perceptron - En enkel modell for et biologisk nevron

“Image classifier” - Bildeklassifisering

“Swift” – Programmeringsspråk som benyttes av AppleInc.

TPR- True positive rate

FPR- False positive rate

TNR- True negative rate

FNR- False negative rate

SAMMENDRAG

Hensikt: Hensikten med denne oppgaven er å fordype seg innenfor emnet kunstig intelligens og deep learning. Med en fordypning innen temaet, ønsker vi å oppnå en forståelse for hvordan kunstig intelligens kan gjenkjenne en tumor i bildemateriale fra MR-caput undersøkelser.

Problemstilling: Hvordan kan en modell utvikles ved hjelp av kunstig intelligens, trenes og benyttes for gjenkjenning av tumor i MR-caput bildemateriale?

Metode: Ved bruk av metoden diagnostiske tester og programmeringsverktøyet Xcode, utvikles det to modeller. Disse modellene bygges på en deep learning (DL) algoritme, som er spesielt utviklet for bildegjenkjenning. Modellene trenes opp til å identifisere en tumor (sensitivitet og spesifisitet) i gitt bildemateriale.

Resultat: Modell-1 oppnår en sensitivitet på 1 (100%) og en spesifisitet på 0 (0%), mens modell-2 oppnår 0,8 (80%) i både sensitivitet og spesifisitet. Sensitiviteten og spesifisiteten som modellene har oppnådd ved gjenkjennelse av en eventuell tumor i bildematerialet brukes for å lage en ROC-kurve, hvor arealet under kurven sier noe om hvor pålitelig modellen er. Ved sammenligning av resultatene vises det at økt mengde bildemateriale gir modellen høyere pålitelighet.

Konklusjon: Når det trenes opp en modell basert på kunstig intelligens innen medisinsk bildediagnostikk, benytter en slik modell automatisert bildesegmentering og dataanalyse for å gjenkjenne attributter i bildematerialet som er gitt. Modellen detekterer og klassifiserer lesjoner i bildematerialet. Resultatene i denne oppgaven viser at det trengs store mengder bildemateriale, for å kunne utvikle en modell som kan benyttes innen medisinsk bildediagnostikk. Resultatene viser også hvordan en ROC-kurve kan benyttes for å finne høyest mulig pålitelighet mellom flere modeller.

SUMMARY

Purpose: The purpose of this bachelor`s thesis is to immerse into artificial intelligence and deep learning. With a deeper understanding of the subject, we want to gain an understanding of how artificial intelligence can recognize a tumour in imagery from MRI-caput examinations.

Problem statement: How can a model be developed using artificial intelligence, trained and used for tumour detection in MRI-caput imagery?

Methods: Using the diagnostic test method and a programming-tool, Xcode. Two models were developed. These models are "built" on a deep learning algorithm that is specific to image recognition, these models are trained to identify the tumour (sensitivity and specificity) in the given imagery.

Result: Model-1 achieves a sensitivity of 1 (100%) and a specificity of 0 (0%), while model-2 achieves 0.8 (80%) in both sensitivity and specificity. The sensitivity and specificity obtained by the models in recognizing a possible tumor in the imaging material are used to create a ROC-curve, where the area under the curve gives an indication of how reliable the model is. Comparing the results shows that an increased amount of images gives the model a higher reliability.

Conclusion: When training a model based on artificial intelligence in medical imaging, such as a model utilizes automated image segmentation and data analysis to recognize attributes in the imagery provided. The model detects and classifies lesions in the imagery. The results of this thesis shows that large amounts of imaging are needed to develop a model that can be used in medical imaging. The results also demonstrate how a ROC-curve can be used for finding the highest possible reliability between several models.

INNHOLDSFORTEGNELSE

Oversikt over Figurer

Oversikt over Tabeller

Oversikt over Formler

1 INNLEDNING

1.1 TEMA OG BAKGRUNN

Tema for oppgaven omhandler kunstig intelligens (KI) innen radiologi. KI er en datamaskin som har evnen til å "huske" og læres opp, til å gjenkjenne et materiale. KI defineres som "en type dataprogrammer med evne til å nå komplekse mål. Disse trekker som regel lærdom fra miljø" (Bjørkeng, 2019, s. 17).

Bakgrunnen for oppgaven er at KI er et relevant emne for helsepersonell, som innen medisinsk- avbildning benytter stråling (Murphy & Liszewsky, 2019, s.15). KI er et aktuelt og radiografrettet tema, som er "fremtiden" innenfor medisinsk bildeteknologi. Ifølge de yrkesetiske retningslinjene for radiografer, har radiografen flere ansvar. Blant annet skal radiografen nyttiggjøre seg av potensialet i ny teknologi og holde seg oppdatert og bidra i fagutvikling og forskning i sitt yrke (Radiografforbundet, 2015).

Deep learning (DL) er en teknikk innen KI der man har gjort store fremskritt og spesielt innenfor bildeanalyse. Dette kombinert med forbedret datakraft endrer både måten helsesektoren styres og oppfattes. Det produseres en enorm mengde avbildninger og tilhørende datamateriale ved en radiologisk avdeling. Dette har vekket interessen til noen av de største teknologiselskapene i verden. KI-bedrifter søker tilgang på pasientdata, med et formål om å trene dype læringsalgoritmer (DL) for å kunne automatisere oppgaver som bildeklassifisering og segmentering (Murphy & Liszewsky, 2019, s.15).

Per i dag er det relativt lite erfaring ved bruk av KI i pasientbehandling. Det kreves mye forskning på temaet for å kunne forstå hvordan KI fungerer og hvordan en algoritme skal benyttes på korrekt måte. Hvis et verktøy som KI skal benyttes i klinisk bruk, må det tas hensyn til viktige aspekt som menneskerettigheter, verdighet og personvern. Etisk bruk av KI legger fokus på å minimere feil og skader, samt ivareta pasientens personvern og rettigheter (Geis et.al, 2019, s. 2). Ifølge Bergsjø & Bergsjø (2019, s. 114) forskes det på ny teknologi innenfor KI og hvordan den nye teknologien kan revolusjonere kvalitet og diagnosesetting i helsesektoren. Et eksempel på dette er MIM-studien som gjennomføres av kreftregistret. Denne studien omfatter brystmammogrammer ved screeningundersøkelser som bruker læringsalgoritmer og en metode basert på KI. Målet er å utvikle en modell (programvare) som automatisk kan tyde bildemateriale i screeningmammogrammer (Kreftregistret, 2019).

KI vil i fremtiden kunne lette arbeidsmengden på mange områder innenfor radiologien. Fra oppgaver som omhandler alt fra henvisninger til bildeanalyser, men KI vil likevel kunne føre til feil/mangler via slike typer systemer. Derfor er det viktig at helsepersonell tar utviklingen på alvor og er med på utviklingen, fordi KI vil føre til en omstilling i arbeidsoppgaver og ansvar (Groote, 2018).

1.2 RADIOGRAFFAGLIG RELEVANS

I mai 2019 ble en felles arbeidsgruppe opprettet av European Federation of Radiographer Sociations (EFRS) og International Society of Radiographers and Radiologic Technologyists (ISRRT), for å utforske hvilken innvirkning KI kan ha for radiografyrket. Radiografer og radiologiske teknologer bør tilpasse bildebehandling og pasientbehandlingspraksis, for å sikre at den nye teknologien (KI) blir implementert, brukt og regulert korrekt. Endringer som gjøres i praksis må understøttetes av utdanning og opplæring, både for eksiterende og fremtidens studieplan. Radiografer og radiologiske teknologer spiller en aktiv rolle i planlegging, utvikling, implementering, bruk og validering av KI-applikasjoner i medisinsk avbildning og strålebehandling (Woznitza, 2020 s.93-94).

Teknologien vil være med å forenkle en hektisk arbeidshverdag for radiografer og stråleterapeuter, samt heve kvaliteten på tjenestene som blir utført på radiologisk avdeling. Radiografer har i løpet av de siste 20 årene vist en evne til å tilegne seg ny kunnskap og kompetanse som er viktig innenfor faget (Mikaelsen, u.å.).

1.3 HENSIKT OG PROBLEMSTILLING

Hensikten med oppgaven er en fordypning innenfor emnet KI, maskinlæring (ML) og DL. Med en fordypning innen temaet ønsker vi å oppnå en forståelse for hvordan KI og læringsalgoritmer fungerer, samt hvordan KI kan gjenkjenne en tumor i bildediagnostikken. I denne oppgaven blir det utviklet to modeller med ulik mengde treningsmateriale. Modell-1 trenes på en liten mengde bildemateriale, mens modell-2 trenes opp på en større mengde bildemateriale. Disse modellene baseres på en DL- algoritme, som er utviklet av selskapet Apple og trenes til å identifisere en tumor i magnetisk resonans (MR)-caput bildemateriale. Modellenes sensitivitet og spesifisitet vil bli analysert og sammenlignet. På bakgrunn av dette har vi utarbeidet en problemstilling som lyder;

"Hvordan kan en modell utvikles ved hjelp av kunstig intelligens, trenes og benyttes for gjenkjenning av tumor i MR-caput bildemateriale?"

1.4 AVGRENSNING

Oppgaven avgrenses primært til MR-caput bildemateriale i aksialt plan, med- og uten tumor. Det tas utgangspunkt i at leseren har kunnskaper om MR-bildeframstilling og patologiske funn, i form av tumor. Modellen som utarbeides avgrenses til å gjenkjenne en tumor i det diagnostiske bildematerialet, ved hjelp av en såkalt "ikke styrt læringsalgoritme".

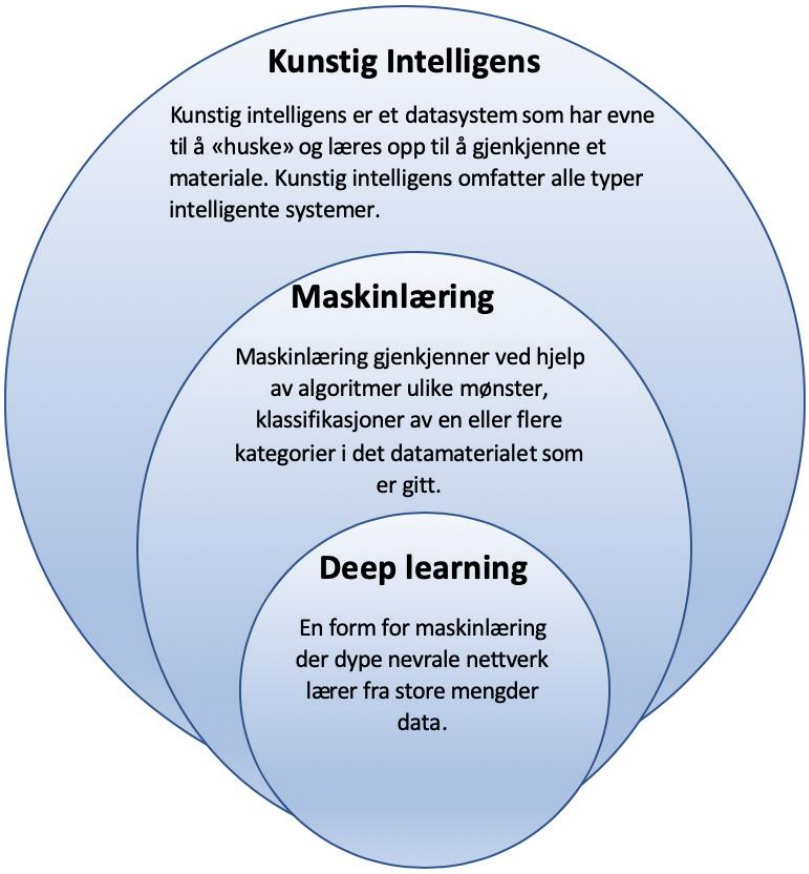
2 TEORI

2.1 KUNSTIG INTELLIGENS

KI er et begrep som oppsto på 50-tallet og ifølge Tørresen (2013, s. 15) kommer ordet intelligens fra det latinske ordet intellegentia, som betyr "å være forstandig". Det var under et seminar på Dartmouth College i 1956, at begrepet "kunstig intelligens" oppsto. Forskere som John McCarthy, Allen Newell, Marvin Minsky og Herbert Simon etablerte et samarbeid etter dette seminaret og ledet forskningen innenfor KI i flere tiår. Formålet med KI er å utvikle datasystemer som gir intelligent oppfattelse, resonnering og respons (Tørresen, 2013, s. 13-14).

KI omfatter alle typer intelligente systemer og i denne oppgaven vil det bli tatt for seg viktige underkategorier som ML, DL og kunstig nevralt nettverk (KNN).

ML omhandler systemer som lærer en maskin å gjøre bare det "mennesket" bestemmer at den skal gjøre. Dette gjøres via en algoritme som er en "oppskrift" på hvordan maskinen skal utføre en oppgave, i form av koder (Bergsjø &Bergsjø, 2019, s.56). Maskinen kan læres opp ved hjelp av forskjellige læringsalgoritmer, som blant annet styrt- og ikke styrt læring, slik at det kan utvikles en modell som kan løse ulike oppgaver (Bergsjø & Bergsjø, 2019, s. 55-57).



Figur 1. Sammenheng mellom kunstig intelligens, maskinlæring og deep learning

Ifølge Currie et. al (2019, s. 477) er KNN ryggraden i ML og DL ved medisinsk avbildning. KNN kan sammenlignes med menneskers nervesystem, der KNN er selve ryggraden i KI og DL. KI og DL er en avansert form for maskinlæring, som ved medisinsk avbildning er basert på hjernen vår, hvor kunstige nevroner er koblet sammen med synapser. Når DL brukes i medisinsk bildediagnostikk bygges det på et KNN, som er en type analysealgoritme sammensatt av lag med tilkoblede noder. En node kan sammenlignes med et kontaktpunkt mellom en nerve og synapse i det anatomiske nervesystemet, og blir omtalt som kunstige nevron (Currie et. al, 2019 s. 478).

Ved KI innenfor radiologi trengs det store mengder datamateriale for at en algoritme skal kunne læres opp til å detektere og klassifisere lesjoner, automatisere bildesegmentering og dataanalyse i gitt bildemateriale (Currie et.al, 2019, s. 478). Det er flere faktorer som kan påvirke nøyaktigheten til en algoritme ved DL. Antall ganger det sendes et datasett gjennom en algoritme, kalles for en iterasjon (en epoke). Dette har vist seg å øke test-nøyaktigheten, og innenfor medisinsk bildediagnostikk er det DL teknikken som er mest brukt (McBee et. al, 2018, s. 1475). Som oss mennesker kan også maskiner være forutinntatt og lite

gjennomskuelig. Ifølge Geis et. al (2019, s. 2) vil bruken av slike intelligente systemer innen radiologien øke risikoen for systematiske feil.

2.1.1 Algoritme

En algoritme kan defineres som "et sett med operasjoner som – hvis de følges- gis et bestemt resultat" (Apeland, 2015). En algoritme er en "oppskrift" på hvordan en kode skal bli riktig utført. Algoritmen trenger store mengder datamateriale for å kunne være pålitelig, og må instrueres via ord eller matematiske utregninger. Algoritmen vet ingenting når den startes, men målet er å få maskinen til å forbedre algoritmen selv. (Bergsjø & Bergsjø, 2019, s. 55-57). Det finnes en algoritme for alt en datamaskin potensielt klarer å utføre og mange av disse algoritmene er ikke engang oppdaget (Telle, 2017, s. 197).

2.2 MASKINLÆRING

Alan Turing prøvde å finne en definisjon på hva som må til for å kunne kalle en maskin intelligent. En moderne datamaskin som skal kunne gjøre oppgaver som mennesker, må ha mange egenskaper og er svært komplisert, men baseres på noen få grunnleggende ideer. Noen av de viktigste ideene ble utviklet for over 80 år siden av Alan Turing (Telle, 2017. S. 195-197).

> Mange algoritmer var kjent, men Turing var den første som ga en presis definisjon av hva en algoritme er. Han beskrev en maskin, eller hva vi kan kalle en modell for en datamaskin, og postulerte at denne «Turing-maskinen» kunne utføre enhver rekke av slike veldefinerte steg som ville falle innenfor det inntil da intuitive begrepet algoritme. Dette postulatet er i dag allment akseptert. En Turing-maskin kan utføre alle de beregninger som kan gjøres av en datamaskin, både i dag og i fremtiden. Man skulle dermed tro at den var meget komplisert, slik en datamaskin er, men det er den ikke. (Telle, 2017, s. 195)

Ifølge Telle (2017, s. 194) byr det ikke på de store problemene om man programmerer en datamaskin for å sjekke ut om regnestykker er utført korrekt. Oppfinnelsen av datamaskinen ble gjort av menneske slik at det skulle bli mulig og automatiserer oppgaver som har helt klare og presise definisjoner, som for eksempel regnestykker med multiplikasjon. Mens det å skille mellom forskjellige dyrearter, eksempelvis forskjellen på en hund og en katt, kan by på problemer ved den automatiserende prosessen som foregår ved ML. Allikevel har de siste årene vært revolusjonerende innenfor teknologiens verden. Ved hjelp av den nye ML har datamaskinen klart å gjennomføre og automatisere prosesser som vi selv utfører ubevisst.

Denne typen ML bruker store mengder med datamateriale som har et tilhørende resultat eller fasitsvar. Dette gjøres for å heve hypotesen for begrepet som skal læres. Hvis man har store mengder datamateriale med eksempelvis fasitsvar “hund”, læres maskinen opp slik at etter endt trening, vil maskin klare å anvende denne “hunde”- hypotesen på nye eksempler som dukker opp med innholdet “hund”. Deretter vil maskinen gi et fasitsvar på om innholdet på datamaterialet inneholder “hund”. Hvis maskinen ikke finner “hund” i innholdet, vil den gi beskjed om negativt resultat (Telle, 2017, S.194).

Tradisjonell ML handler i all hovedsak om å gjenkjenne mønstre, klassifikasjoner av en eller flere kategorier i datamaterialet som blir gitt, som for eksempel i medisinsk sammenheng (Dyrdal et. al, 2017, s. 7). Det trenes opp en klassifikator som skal skille mellom ulike kategorier, basert på egenskaper i det gitte datamaterialet. Maskinen bruker egenskaper fra gitt datamaterialet for å kunne klassifisere objekter. Dette blir gjort i en form av tallstørrelser, som er avledet fra objekt som er funnet i datamateriale. “Egenskapene som maskinen bruker til å klassifisere objektene, er tallstørrelser avledet fra objektene. For et objekt i et bilde kan f.eks. høyde og bredde være mulige egenskaper” (Dyrdal et. al, 2017, s. 7).

2.2.1 Forskjellige læringsalgoritmer

Når vi snakker om ML, dekker dette begrepet flere teknikker som vanligvis blir brukt, som for eksempel styrt-, ikke styrt læring og en forsterket læring (Kommunal- og moderniseringsdepartementet, s. 11).

> Når vi i dag hører om løsninger basert på kunstig intelligens, er det som regel løsninger som baserer seg på maskinlæring. Begrepet maskinlæring dekker en rekke ulike teknikker, der reglene utledes fra de dataene systemet trenes på, i motsetning til regelbaserte systemer der reglene er gitt av mennesker, ofte basert på eksperterfaring, forretningslogikk eller regelverk. (Kommunal- og moderniseringsdepartementet, s.11)

KI-systemer med ML, bygger på matematiske modeller som er basert på treningsdata. Disse modellene brukes så for å ta beslutninger. Ved styrt læring blir algoritmen trent med et datasett, hvor både "input”- og "output”-materiale er gitt av en operatør. Denne type algoritme bruker da både “input”- og “output”-materiale til å bygge en modell. Algoritmen som blir trent med styrt læring, får både «oppgaven» og «fasiten» oppgitt. Denne informasjonen blir benyttet slik at det kan bygges en modell. Dette resulterer i at modellen senere vil kunne ta en beslutning på basert datamateriale (Kommunal- og moderniseringsdepartementet, s.11).

Ikke styrt læring derimot fungerer ved at algoritmen får oppgitt kun ett datasett, det vil si kun "input"-materiale. Ikke styrt læring får kun "oppgaven" oppgitt. En ikke styrt læringsalgoritme må derfor selv, ut ifra mønstrene den finner i "input"-materiale ta egne beslutninger, når det blir tilført en nytt sett med datamateriale. Det er ved bruk av en ikke styrt læringsalgoritme det kan trenes opp og utvikles DL-algoritmer (Kommunal- og moderniseringsdepartementet, s.11).

Forsterkende læring baseres på en ikke styrt læringsalgoritme. Her brukes det også en operatør som gir tilbakemelding på om "output"-materialet er nøyaktige eller upresise. På denne måten blir algoritmen "matet" med tilbakemelding, som vil bidra til å forbedre modellen (Kommunal- og moderniseringsdepartementet, s.11).

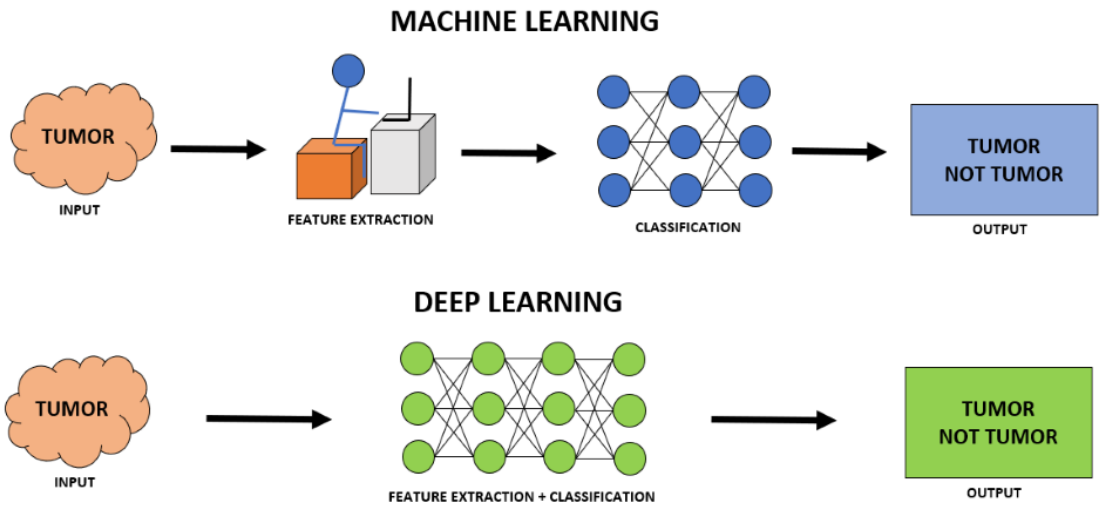
2.3 DEEP LEARNING OG KUNSTIG NEVRALT NETTVERK

DL en teknikk og ikke en spesifikk algoritme, og som tidligere nevnt er KNN basert på strukturen til biologiske nervesystemer som hos mennesket (McBee et. al, 2018, s. 1473). ML har siden 1950 årene hatt en rask utvikling. Utviklingen av KI-systemer har hatt som mål å etterligne menneskehjernens måte å gjøre beregninger på. I 1957 foreslo Frank Rosenblatt å bruke et såkalt "perceptron" for å gjøre beregninger. Et perceptron er en enkel modell for et biologisk nevron. Dette arbeidet ga opphavet til forskning på KNN (Dyrdal et. al, 2017, s 8).

KNN består av noder og "kunstige nevron". Ved DL kan disse være lagt opp i rekkefølge fra hundrede til millioner konfigurert rekker med lag. En node utgjør kontaktpunktene mellom disse lagene og de kunstige nevronene. Dette utgjør dybden i en DL-algoritme. DL og KNN som består av mange lag, blir generelt sett på som en mer sofistikert implementering av ML. KNN er i stand til å utføre en mer detaljert analyse og kombinere mer datamateriale. Hvert kunstige nevron, mottar informasjon fra andre kunstige nevron og utgangene fra disse nevronene er vektet. KNN tar sikte på å maksimere riktige svar sammenlignet med en referansetest. Dette gjøres ved å justere vekting på hvert kunstige nevron, basert på feilen som er beregnet på hver forplanting (Currie et. al, 2019, s.477). Informasjonsinnganger består av ulike lag av kunstige nevroner, hvor de ulike lagene består ulike nivåer av kunstige nevron. Hvert kunstige nevron i øvrige lag, kan kombinere kunstige nevroner fra lavere nivåer. Dette er for å kunne danne en ny og mer kompleks utgang. Når disse mellomlagene øker, vil også nøyaktigheten til utdataene i det høyeste laget øke. Enkel ML inkluderer som oftest bare et

lite antall av disse lagene. DL inneholder et høyere antall av disse lagene. Det vil si at flere lag, vil gi en mer nøyaktig modell (McBee et. al, 2018, s. 1473).

Gjennom hver iterasjon, konvergerer den matematiske løsningen til en mer nøyaktig løsning. Trenings-fasen oppnår best resultat med et stort sett av datamateriale. Store datasett innen medisinsk avbildning spiller en viktig rolle i å gi store, pålitelige treningsdata som ML- og DL-algoritmer kan lære av (Currie et. al, 2019, s. 477).



Figur 2. Maskinlæring vs. Deep learning

Figur 2 viser hvordan ML fungerer, hvor en operatør må spesifisere hvilke egenskaper i datamaterialet som skal fokuseres på og systematisk lærer maskinen opp ved å gi maskin en gitt "input" (inndata) og en gitt "output" (utdata).
Derimot så vil DL kjenne igjen egenskaper og attributter i datamaterialet ved nytt læringsmateriale som blir gitt ved "input", og på denne måten "lærer den seg selv". Output ved DL vil være en mer kompleks og avansert form for maskinlæring (Bergsjø & Bergsjø, 2019, s. 55-57).

DL sitt mål er visuell gjenkjennelse ved bruk av KI i radiologi og å produsere færre feiltolkninger, enn en manuell tolkning av en operatør (radiolog) (Currie et. al, 2019, s. 477). DL i medisinsk bildediagnostikk skiller seg fra andre teknikker, hvor bildematerialet fra diagnostiske bilder kan bestå av store mengder rådata, eksempelvis fra en CT eller MR-undersøkelse. Dette gjør at algoritmen må ha komplekse beregninger som er meget nøyaktige. I tillegg er bildematerialet innenfor radiologien veldig variert. Forskjellige type patologi kan føre til store ulikheter i bildematerialet, som brukes til å trene opp en modell basert på KI. En annen innvirkning kan være at ulike personer ser anatomisk ulike ut innvendig. Dette gjør at

det blir komplisert når en modell og en algoritme skal trenes opp (McBee et. al, 2018, s. 1476).

KNN består som tidligere forklart av flere lag, det første laget kalles "Input layer". Hvis det for eksempel benyttes diagnostisk bildemateriale, vil det første laget med kunstige nevroner kode for verdiene på inngangspikslene i det gitte bildemateriale. Disse kunstige nevronene utgjør utgangspunktet for det maskinens modell skal lære. For at modellen skal kunne trenes opp, trengs det matematiske transformasjoner og analyser av datamaterialet i "input layer". Dette kalles "hidden layer". I DL er det ofte flere "hidden layers", og disse lagene utfører mange beregninger og analyser som gjør at algoritmen får et bedre grunnlag for utgangsdataene. Disse utgangsdataene er et resultat av både Input- og hidden layer (McBee et. al, 2018, s. 1473)

2.3.1 Konvensjonelle nevrale nettverk

Konvensjonelle nevrale nettverk (CNN) er en annen form for KNN som er brukt innen radiologi. Denne formen DL kombinerer to typer "hidden layers" og produserer en "output layer" i form av klassifikasjon (Currie et. al, 2019, s. 478). CNN er et godt verktøy for å identifisere og trekke ut sine egne radiomiske funksjoner fra "input layer" og deretter koble det sammen med "output layer". Dette gjøres for å oppnå best mulig resultat (Currie et. al, 2019, 479).

2.4 FORDELER OG ULEMPER VED KUNSTIG INTELLIGENS

KI har et stort potensial som kan øke effektiviteten og nøyaktigheten innenfor radiologien. Den "nye" teknologien kan være et godt hjelpemiddel som bidrar til å lette arbeidsoppgaver i en hektisk hverdag. Likevel er det flere element som kan senke utviklingen, eller som kan gjøre det vanskelig å innføre en slik teknologi, der en viktig faktor er personvern. Det trengs store mengder med datamateriale om KI skal tas i bruk innen radiologi. Samt administrering og anonymisering av datamaterialet der informert samtykke og objektivitet må inkluderes, hvor målet er å ivareta personvern. Dette er et ressurskrevende arbeid som vil foregå over lengre tidsperioder (Geis et al, 2019, s. 2).

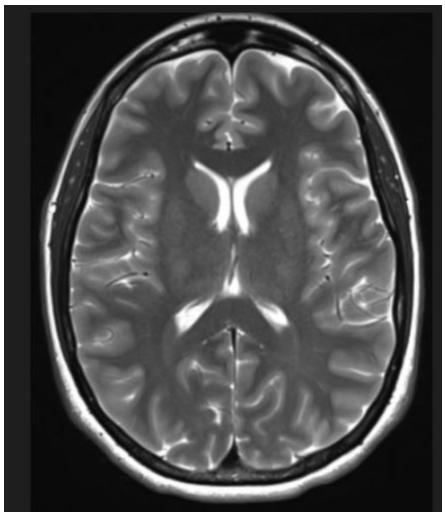
Det er lite erfaring ved bruk av KI i pasientbehandling og systematiske feil kan gi store konsekvenser. Det kreves mye forskning innenfor emnet for å forstå hvordan KI vil påvirke

radiologien. Den som benytter KI må forstå risikoen, sikre bruken og ta i bruk KI med fokus på å gjøre det beste for pasientene. Selv om de fleste endringer vil være positive, vil KI føre til sosiale og økonomiske endringer. Dette vil de mest utsatte samfunnene merke godt. Derfor må det sikres for at negative konsekvenser blir minimale, ved å fokusere på personvern og pasientbehandling (Geis et al., 2019, s. 2).

2.4.1 Mønstergjenkjenning i bildediagnostikken

Når objekter i bilder klassifiseres, er det mange egenskaper ved bildemateriale maskinen legger vekt på. Tallstørrelser i form av høyde, bredde, omkrets og areal er viktige opplysninger for maskinen. Samt farger, lysstyrke og tekstur. Disse egenskapene gjør at maskinen kan skille mellom ulike objekter ved hjelp av trening. Å trene opp maskinen til å kunne skille mellom for eksempel, bildemateriale med- og uten tumor fra hverandre, kan derfor være mer komplisert. Medisinske diagnostiske bilder er i utgangspunktet i svart/hvitt, noe som gjør at maskinen "mister" en klassifikasjonsegenskap, i form av farger (Kommunal- og moderniseringsdepartementet, 2020, s. 12).

Hjernevevet består av grå og hvit substans, der grå substans består av nerveceller og hvit substans består av nervetråder. Dette gir ulike gråtoner i eksempelvis MR-bilder. Der vil man se at hvit substans, har en lysere gråtone enn grå substans. Vevet er relativt homogent, der gråtonene "glir litt over i hverandre" (Elster, 2020).



Figur 3. MR-bilde, T2 vektet bildesekvens

Figur 3 viser hvordan normalt hjernevev ser ut på et MR-bilde med T_2 vektet bildesekvens, her ser man at grå og hvit substans har to forskjellige gråtoner.

3 METODE

3.1 VALG AV METODE

Metoden som ble benyttet i denne oppgaven er, utviklingen av en modell for diagnostiske tester. “Diagnostiske tester har som formål å identifisere og utelukke diagnoser, for eksempel syk/frisk, feber/ikke feber, risiko/ikke risiko” (Helsedirektoratet, 2016).

Med fokus på KI, ML og DL, var vi avhengige av hjelp fra fagpersoner for å komme i gang med å utvikle en bildegjenkjennings-modell, som kunne gjenkjenne patologi (tumor) i et medisinsk bildemateriale. Vi fikk hjelp av en sivilingeniørstudent som anbefalte å utvikle en modell via programmet Xcode, hvor det allerede finnes godt etablerte og testede algoritmer (AppleInc., u.å.).

Med Xcode og dens læringsalgoritmer, ble det utviklet to modeller. Disse modellene skulle identifisere hvilke diagnostiske bilder som inneholdt tumor og hvilke som ikke inneholdt tumor (normalt hjernevev). Utvalgt bildemateriale som ble benyttet i de to modellene ble hentet fra Radiopedia.org og består av MR-caput undersøkelser, i aksial plan. Modellene ble trent med ulik mengde treningsmateriale, slik at resultatene til hver modell kunne analyseres og sammenlignes, basert på mengde gitt treningsmateriale.

Det ble tilstrebet å få et resultat som kunne vise en forskjell i modellenes nøyaktighet, samt analysere betydningen av mengde tegningsmateriale. Det ble fokusert på forskjellen i sensitivitet og spesifisitet ved de to modellene. “Sensitiviteten er sannsynligheten for at en syk pasient får riktig svar, det vil si positiv test. Spesifisiteten er sannsynligheten for at en frisk pasient får riktig svar, det vil si negativ test” (Lydersen, 2017).

3.1.1 Referansetest

For å vurdere modellenes sensitivitet og spesifisitet, ble modellenes resultater målt opp mot en referansestandard. En referansestandard klassifiserer og sammenligner antall korrekte beslutninger som blir tatt med antall korrekte beslutninger som det faktisk finnes i det gitte test-materialet (Helsedirektoratet, 2016). Test-materialet i denne oppgaven består av 10 diagnostiske bilder, hvorav fem bilder med tumor og fem bilder uten tumor (friskt hjernevev). Hvis modellene oppnår 100% sensitivitet og spesifisitet, må det tas korrekt beslutning på alle de diagnostiske test-bildene. Det må også tas høyde for hvilke som er med- og uten tumor.

Referansetesten til hver av modellene ble sammenlignet, noe som ble gjort for å kunne analysere utfallet av hver modells referansetest, basert på gitt mengde treningsmateriale. På denne måten oppnås det en indikasjon på hva slags betydning mengden med treningsmateriale har, når det utvikles en tumorgjenkjennelses modell ved hjelp av KI. For å utforme en referansestandard i denne oppgaven ble det brukt en tabell som er hentet fra helsedirektoratet (2016).

Tabell 1. Oppsett referansestandard

Tumorgjenkjenningsmodell x: Iterasjoner: x Treningsbilder: x		**Referansestandard**		
		Patologi	**Ikke patologi**	**Total**
Test	**Positiv**	a	b	a+b
	Negativ	c	d	c+d
	Total	a+c	b+d	a+b+c+d

Sensitivitet	**Spesifisitet**
a/(a+c)	d/(b+d)

Tabell 1 er utarbeidet fra helsedirektoratet sin tabell fra "sjekkliste for vurdering av en studie som tester en ny diagnostisk test" (vedlegg 8.1) (Helsedirektoratet, 2016). Tabellen viser hvordan man kan regne ut de forskjellige verdiene for falske og positive svar ved oppsett av en referansetest.

Referansestandarden i denne oppgaven tilsvarer full "score" på alle bilder. Det vil si, at høyeste "score" og det mest optimale resultatet referansetesten kan oppnå, vil være en sensitivitet som er lik 1(100%) og en spesifisitet som er lik 1 (100%). Når modellene skulle utarbeides ble det ikke tatt utgangspunkt i en spesifikk diagnose, men å lage en modell som klarer å skille mellom bildemateriale med- eller uten tumor. Referansetesten ga grunnlaget for formelen som vises i tabell 3.1 og hvordan man beregner sensitiviteten og spesifisiteten til referansetesten.

3.1.2 Receiver Operating Characteristic Curve (ROC)

En ROC-kurve gjør det mulig å lage en fullstendig sensitivitets- og spesifisitetsrapport. ROC-kurven er et grunnleggende verktøy for evaluering av diagnostisk test. En slik kurve setter "true positive rate" (TPR) opp mot "false positive rate" (FPR). Området under ROC-kurven omtales som "arealet under kurven" (AUC) og er et mål på hvor godt en parameter kan skille

mellom to diagnostiske grupper (syk/frisk). Den diagnostiske ytelsen til en test, eller nøyaktigheten av en test for å diagnostisere syke tilfeller fra normale tilfeller, evalueres ved å bruke ROC kurveanalyse. ROC-kurver kan også brukes til å sammenligne den diagnostiske ytelsen til to eller flere laboratorie- eller diagnostiske tester. Når resultatene av en bestemt test vurderes, deles de i to populasjoner. Der en populasjon, er med en sykdom og den andre populasjonen, er uten sykdom (Sardanelli & Di Leo, 2008). I denne oppgaven er det en populasjon med tumor og en populasjon uten tumor.

En slik binær-klassifisering, der resultatene enten er positive eller negative, kan resultatene ha fire ulike utfall. "True positive" (TP), "false positive" (FP), "false negative" (FN), "true negative" (TN) (Sardanelli & Di Leo, 2008). En ROC-kurve benytter både "TPR", som tilsvarer sensitiviteten- og "FPR" som tilsvarer "1-spesifisiteten" til modellen. Ved å benytte TPR og FPR kan utfallet til hver av de to modellene beregnes.

3.1.3 Etisk aspekt

Radioepdia.org deler caser og erfaringer gjennom forskjellige pasientsaker. Hver enkelt sak som blir publisert på Radiopedia tilhører et medlem, som blir veiledet av dedikerte redaktører for å oppnå kvalitetsstandard og personvern (Radiopedia, u.å.). Det vil si, at alle personvernsopplysninger på de forskjellige undersøkelsene er fjernet/anonymisert. Det vil derfor ikke være mulig å identifisere personen på de forskjellige undersøkelsene. Det ble hentet ut bildemateriale fra forskjellige MR-undersøkelser. Det ble gjort uten å opprette medlemskap eller lisens hos Radiopedia.

Radiopedia har en egen "Creative Commons-license", som baserer seg på at man ikke trenger å be om tillatelse ved bruk av bilder. Så lenge den medvirkende brukeren blir attribuert, bruken er ikke- kommersiell eller at man ikke tar copyright på materialet. Hvis bruken av materialet faller utenfor "Creative Commons-license" må det søkes om tillatelse for bruk av materialet som eventuelt skal benyttes (Radiopedia, u.å.).

På bakgrunn av Radiopedia sine vilkår blir det ikke stilt et forskningsetisk vurderingsspørsmål i denne oppgaven. Dette er på grunnlag av at valgt bildemateriale som brukes i oppgaven, holdes innenfor rammen "Creative Commons-license" til Radiopedia.

3.2. BILDEMATERIALE

Det var viktig at bildemateriale som ble valgt, kom fra sikre og troverdige kilder. Nettsiden radiopedia.org ble benyttet for å finne diagnostisk bildemateriale, som kunne brukes til å lære opp begge modellene. Denne nettsiden har blitt anbefalt av faglærere som relevant innen radiografifaget. Radiopedia.org gir grundig beskrivelse av patologi i de ulike anatomiske områdene, i de casene som ligger på nettsiden. Ved å benytte søkeord som "brain tumour, MRI" fant vi relevante MR-caput undersøkelser med ønsket bildemateriale, hvor det var mulighet for å "bla" i snittene ved de ulike bildesekvensene i hver enkelt undersøkelse. Utvalgt bildemateriale som benyttes i denne oppgaven består av skjermbilder hentet fra forskjellige MR-caput undersøkelser, via radiopedia.org.

Når bildemateriale uten tumor skulle velges, ble radiopedia.org også benyttet. Det var problematisk å finne nok ønskede undersøkelser som besto av friskt hjernevev, altså bildemateriale uten patologi. Derfor valgte vi å benytte bildemateriale med underliggende patologi i andre «snitt». Det betyr at det kan finnes patologi i andre snitt enn de utvalgte snittene som ble benyttet som bildemateriale "uten tumor" i modellene.

3.2.1 Kriterier for bildemateriale

Inklusjonskriterier for utvalgt bildemateriale besto av bildemateriale hentet fra MR-caput undersøkelser, med t_2 vektet bildesekvens i aksialt plan. Dette bildematerialet ble avgrenset til å bestå av diagnostiserte tumorer og ikke annen patologi, som for eksempel en hjerneblødning eller hjerneaneurisme.

Det utvalgte bildematerialet som ble brukt for å trene opp en modellene via Xcode, tilstrebes så langt det lar seg gjøre, å være fra det samme området og "snittet" i hver enkelt MR-undersøkelse. På grunn av utvalgt mengde med bildemateriale i denne oppgaven, kan det ikke forventes at modellene som utvikles oppnår perfekte resultater.

For å trene opp modellene ble det brukt bildemateriale med- og uten tumor. Grunnet begrenset ressurser ble det valgt ut 20 diagnostiske bilder som treningsmateriale (se vedlegg 8.3) og 10 diagnostiske bilder som test-materiellet (se vedlegg 8.4).
Modell-1 besto av treningsmateriale som var basert på to diagnostiske bilder, mens modell-2 besto av treningsmateriale som var basert på 20 diagnostiske bilder. Treningsmaterialet besto

av 50 % med tumor og 50% uten tumor ved begge modellene.

3.3 KODINGSVERKTØY: XCODE OG PROGRAMMERING

Xcode er utviklet av selskapet Apple, og er et kodingsverktøy som er laget for programmering og utvikling av forskjellige applikasjoner. Xcode benytter et programmeringsspråk som kalles "swift". Dette programmeringsspråket er unikt for Xcode og Apple, da det er selskapet Apple som selv har utviklet programmeringsspråket "swift".

Xcode er utviklet for å enkelt kunne bruke programmering og koding. Det spesielle med denne typen programmering er at det ikke trengs store teknologiske ferdigheter for å kunne kode. Kodespråket "swift" er basert på egne, enkle kommandoer og funksjoner, som gjør det mulig å kode via programmet Xcode (AppleInc., u.å.).

Xcode inneholder en "image classifier" kode, som er en DL-kode. Denne er spesialutviklet for å klassifisere attributter i gitt bildemateriale (AppleInc., u.å.). Ved å laste ned ferdiglagde bibliotek som allerede inneholder en ferdigutviklet DL-kode "image classifier", er det mulig å utvikle og trene sin egen bildegjenkjennings modell (AppleInc., u.å.).

Disse bibliotekene inneholder en forhåndsskrevet kode som er enkel å implementere. Slike bibliotek gjør det mulig å utvikle kompliserte programmer og koder på relativt kort tid (Apple Inc, u.å.). Grunnen til at Xcode med kodespråket "swift" ble valgt som metode i denne oppgaven, er at Xcode allerede inneholder DL- koden "image classifier", som er utviklet for å klassifisere attributter i bildemateriale. Å utvikle tumorgjenkjennelses-modell som bruker en DL-kode fra grunn kan ta mange år, og det trengs mange tusen linjer med koder som skal være nøyaktige (Currie et. al, 2019).

3.3.1 Iterasjoner

Antall ganger modellens treningsmateriale blir sendt igjennom modellens algoritme, kalles iterasjoner. Det vil si, antall ganger modellen trener på hvert enkelt diagnostisk bilde-materiale (Currie et. al, 2019).

Hvor mange ganger et treningsmateriale blir sendt gjennom en algoritme vil påvirke utfallet (AppleInc, u.å.). Xcode stopper iterasjonene når modellen har oppnådd optimal respons (AppleInc, u.å.). Treningsmaterialet som ble benyttet i denne oppgaven, ble sendt igjennom

modellens algoritme to ganger. Iterasjonene ble da stoppet, fordi modellene var ferdig trent. Ved større mengder bildemateriale ville modellen gjennomført flere iterasjoner, fordi modellene ville brukt mer tid på å gjenkjenne og klassifisere egenskaper i treningsmateriale som ble gitt som input (AppleInc, u.å.).

Figur 4. Utsnitt fra Xcode som viser antall iterasjoner

Figur 4 viser hvordan DL-algoritmen har trent seg selv. Ved to iterasjoner er modell-2 trent 100% og er klar for testing (AppleInc., u.å.).

3.3.2 Treshold

Treshold er grensen eller terskelen som bestemmer hvor høy sannsynlighet det må være for positivt eller negativt resultat i prosent, ved å endre treshold kan sensitiviteten til modellen justeres (Warwick, 2012, s. 157-158)

En sensitivitets- og spesifisitetsutregning vil kunne evaluere modellens forutsigbarhet, samt effekten av å endre klassifiseringsgrensen og hvilke resultat dette gir. Ved å sette treshold på 0%, vil modellen ha en lav terskel og la alle bildene gå gjennom klassifiseringen. Når treshold settes på 100%, vil terskelen for klassifiseringen være høy, og modellen vil ikke godta bildematerialet. Når det utarbeides en ROC-kurve, vil den vise effekten av endringer i treshold (Warwick, 2012, s. 157-158).

3.4 UTARBEIDELSE AV MODELL VIA XCODE

For å sette i gang med programmet Xcode, er det viktig å ha klart adskilt hva som er treningsmateriale og test-materiale. Hvis treningsmateriale hadde blitt brukt som test-materiale, ville modellenes resultat blitt misvisende. I Xcode-biblioteket må bildematerialet ligge sortert i to ulike mapper. Når dette er gjort, vil treningsmaterialet implementereters i programmet og treningen av modellene starter.

For å utvikle en modell som gjenkjenner en tumor i diagnostisk bildemateriale, med hjelp av ML og DL, må bildematerialet importeres til et eksternt bibliotek. I denne oppgaven benyttes programmeringsverktøyet Xcode som inneholder en "image classifier" -kode og utvalgt bildemateriale ble importert til programmets bibliotek.

Det ble utviklet to modeller via Xcode med tilhørende tester. Treningsmaterialet som benyttes, ble plassert i to forskjellige filer. Noe som kan sammenlignes med to "mapper", mappene fikk egne navn "tumor" og "uten tumor". Treningsmaterialet som inneholdt tumor ble plassert i mappen "tumor" og treningsmaterialet som besto av friskt hjernevev ble plassert i mappen "uten tumor". Deretter ble disse mappene og datasettene sendt gjennom modellens DL-kode som trening. Begge modellene benyttet kun to iterasjoner, før Xcode ga beskjed om at modellene var ferdig trent (se Figur 4).

Bildematerialet som benyttes for å trene opp de to ulike modellene gir grunnlaget for utfallet og påliteligheten til hver av modellene. Etter gjennomført trening av modellene ble hver enkelt modell testet, via Xcode. For å kunne teste modellens nøyaktighet via Xcode, opprettes det to nye "testing" mapper som inneholder nytt bildemateriale, med- og uten tumor. Dette besto av 10 nye diagnostiske bilder, som ikke ble benyttet under trenings-fasen av modellene. Test-mappene må ha likt mappesystem som treningsmaterialet. På denne måten beregner hver modell selv prosentvis hvor pålitelig den er. Det ble benyttet det samme test-materialet på begge modellene. Det vil si, at den samme testen ble utført på modell-1 og på modell-2. Dette diagnostiske test-materialet vil da gi en indikasjon for hvor pålitelig og nøyaktig gjenkjennelse de to modellene har, av en eventuell tumor i gitt test-materialet. Modellenes gjenkjennelses egenskaper, baseres på gitt mengde treningsmateriale.

Test-materialet er en indikasjon på hvor nøyaktig og pålitelig utfallet til hver av de to modellene er. Samt, sensitiviteten og spesifisiteten som oppnås ved gjenkjennelse av hvilke bilde-materiale, som består av tumor og hvilke som ikke består av tumor.

3.4.1 Xcode, modell-1

Modell-1, ble basert på et mindre sett treningsmateriale, et diagnostisk bilde med tumor og et uten tumor. Dette bildemateriale ble plassert i to forskjellige mapper “tumor” og “uten tumor”. Gitt treningsmaterial ble kjørt gjennom med to iterasjoner ved trening. Deretter ble modellens referansetest kjørt, test-1.

3.4.2 Xcode, modell-2

Modell-2, ble basert på et større sett med treningsmateriale, 10 diagnostiske bilder med tumor og 10 diagnostiske bilder uten tumor. Disse bildene ble plassert i to forskjellige mapper “tumor” og “uten tumor”, slik som i modell-1. Gitt treningsmateriale ble kjørt gjennom med to iterasjoner ved trening. Deretter ble modellens referansetest kjørt, test-2.

4 RESULTAT

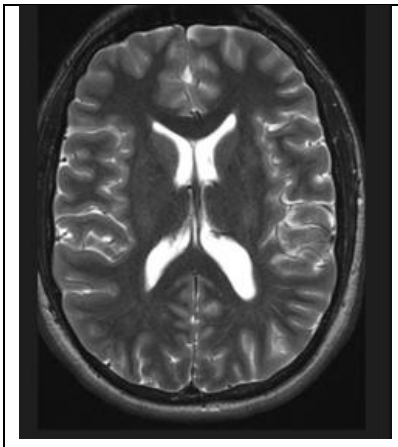
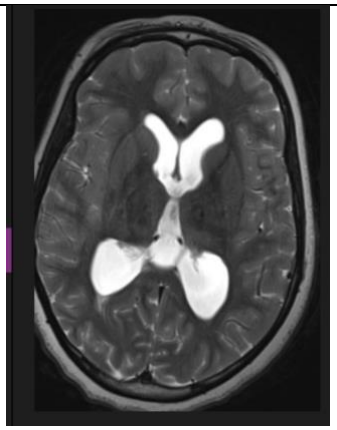
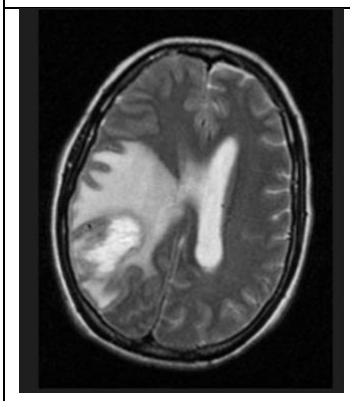
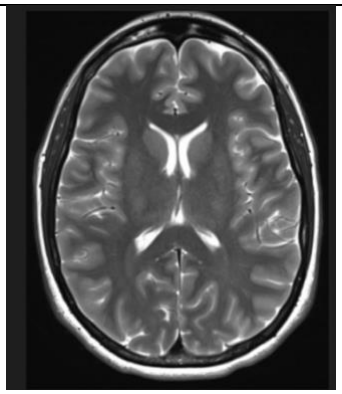
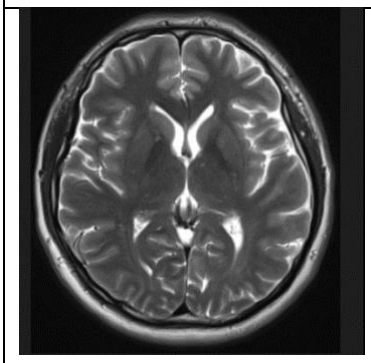
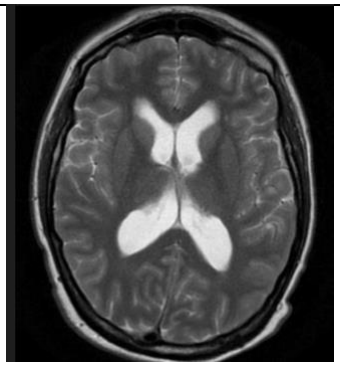
4.1 XCODE RESULTATER

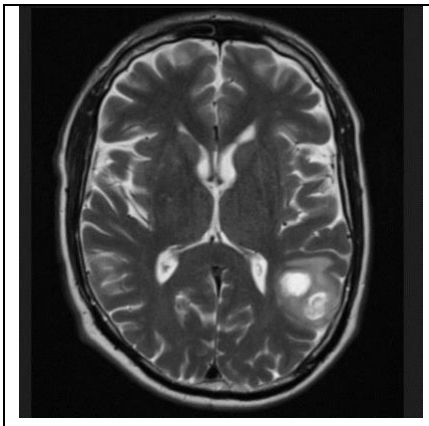
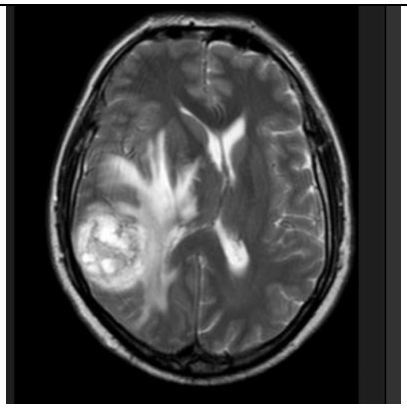
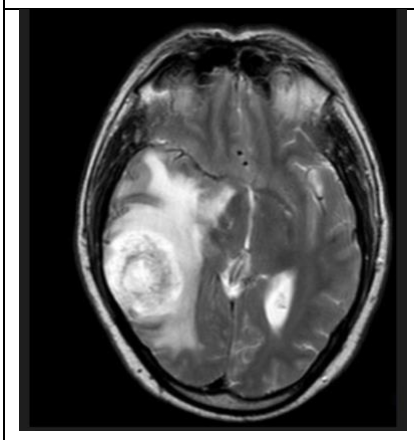
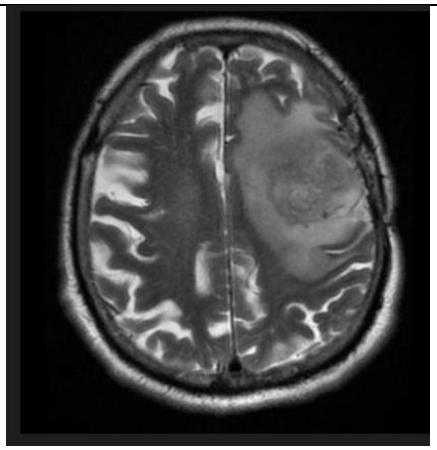
Testene viser påliteligheten og nøyaktigheten til de forskjellige modellene, som er basert på gitt mengde treningsmateriale til hver modell. Modell-1 trenes med et lite sett treningsmateriale, mens modell-2 trenes med et større sett treningsmateriale. På denne måten kan resultatene sammenlignes, basert på forskjellig mengde treningsmateriale. Både modell-1 og modell-2 ble sendt igjennom to iterasjoner under opptrening av modellene.

Når modellenes tester gjennomføres, gis det et nytt sett med test-materiale. Modell-1 og modell-2 testes med det samme test-materialet. Slik at resultatene som hver av modellene oppnår ved “test-runden” kan analyseres og sammenlignes.

4.1.1 Resultat Xcode-test, modell-1 og modell-2

Tabell 2. Testing av modell-1 og modell-2

 Test-bilde 1: Uten tumor Modell-1: Tumor (100% confidence) Modell-2: Uten tumor (100% confidence)	 Test-bilde 2: Uten tumor Modell-1: Tumor (100% confidence) Modell-2: Uten tumor (100% confidence)
 Test-bilde 3: Tumor Modell-1: Tumor (100% confidence) Modell-2: Tumor (100%confidence)	 Test-bilde 4: Uten tumor Modell-1: Tumor (100% confidence) Modell-2: Uten tumor (100% confidence)
 Test-bilde 5: Uten tumor Modell-1: Tumor (100% confidence) Modell-2: Uten tumor (100% confidence)	 Test-bilde 6: Uten tumor Modell-1: Tumor (100% confidence) Modell-2: Tumor (100% confidence)

 Test-bilde 7: Tumor Modell-1: Tumor (100% confidence) Modell-2: Tumor (76% confidence)	 Test-bilde 8: Tumor Modell-1: Tumor (100% confidence) Modell-2: Tumor (100% confidence)
 Test-bilde 9: Tumor Modell-1: Tumor (100% confidence) Modell-2: Tumor (100% confidence)	 Test-bilde 10: Tumor Modell-1: Tumor (100% confidence) Modell-2: Uten tumor (100% confidence)

Tabell 2 viser resultatene til modell-1 og modell-2. Resultatene til modell-1, viser at modellen har beregnet en sannsynlighet på 100% "tumor" på alle bildene. Som vil si, at modell-1 har "funnet" tumor i alle de 10 diagnostiske bildene som ble benyttet som test-materiale. Test-materialet, besto av fem diagnostiske bilder med- og fem diagnostiske bilder uten funn av tumor. I og med at 50% av bildene er med tumor, får modellen et resultat som tilsvarer 5/10 korrekte beslutninger.

Resultatene til modell-2, viser rett avgjørelse på 8/10 tilfeller. Selv om test-bilde 7 bare har 76% sannsynlighet for "funn" av tumor, beregner modellen den som en korrekt avgjørelse. Sammenlignes resten av test-bildenes resultater med fasit, vises det at modell-2 tar feil

beslutning i to av de diagnostiske undersøkelsene. Ved test-bilde 6 beregner modellen funn av "tumor" med 100% sannsynlighet, men sammenlignet med fasitsvar er dette feil beslutning.

I test-bilde 10 beregner modellen seg frem til "uten tumor" med 100% sannsynlighet, men her finnes det et stort patologisk felt med "tumor" i den høyre hjernehalvdelen av caput som modellen ikke har oppfattet.

4.2 ANALYSE

Mengden treningsmateriale som blir gitt, gir utfall på "test-runden" som gjennomføres av modellene i test-1 og test-2. I test-1, gir modellen et resultat hvor det finnes "tumor" i alle testbildene, som resulterer i en nøyaktighet på 50%. Test-2, som inneholder et større sett med treningsmateriale, har en høyere pålitelighet da den har et resultat som tilsier en nøyaktighet på 80%.

4.2.1. Sensitivitet- og spesifisitetsanalyse

Ved en referansestandard er det sensitivitet og spesifisitet som er i fokus. Høyeste "score" er 1/1, som vil si 100%. Det gjennomføres referansetester på hver modell, hvor modellens sensitivitet og spesifisitet beregnes. Referansetesten baseres på modellenes resultat etter gjennomført "test-runde". Referansestandarden er det resultatet modellene skal strebe etter å oppnå (Helsedirektoratet, 2016).

Tabell 3. Sensitivitet og spesifisitet

	Sensitivitet	**Spesifisitet**
Modell-1	1 (100%)	0 (0%)
Modell-2	0,8 (80%)	0,8 (80%)

Tabell 3 viser en oversikt over resultatene til de to ulike modellene som ble utarbeidet. Her vises sensitiviteten og spesifisiteten etter utregning basert på referansestandard.

Tabell 4. Referansetest, modell-1

Tumorgjenkjenningsmodell 1: Iterasjoner: 2 Treningsbilder: 2		Referansestandard		
		Patologi	Ikke patologi	Total
Test	Positiv	5	5	10
	Negativ	0	0	0
	Total	5	5	10

Sensitivitet	Spesifisitet
1	0

Tabell 4 viser sensitivitet og spesifisitet i tumorgjenkjenningsmodell-1 basert på referansestandard. Sensitiviteten er 0 og spesifisiteten er 1.

Tabell 5. Referansetest, modell-2

Tumorgjenkjenningsmodell 2: Iterasjoner: 2 Treningsbilder: 20		Referansestandard		
		Patologi	Ikke patologi	Total
Test	Positiv	4	1	5
	Negativ	1	4	5
	Total	5	5	10

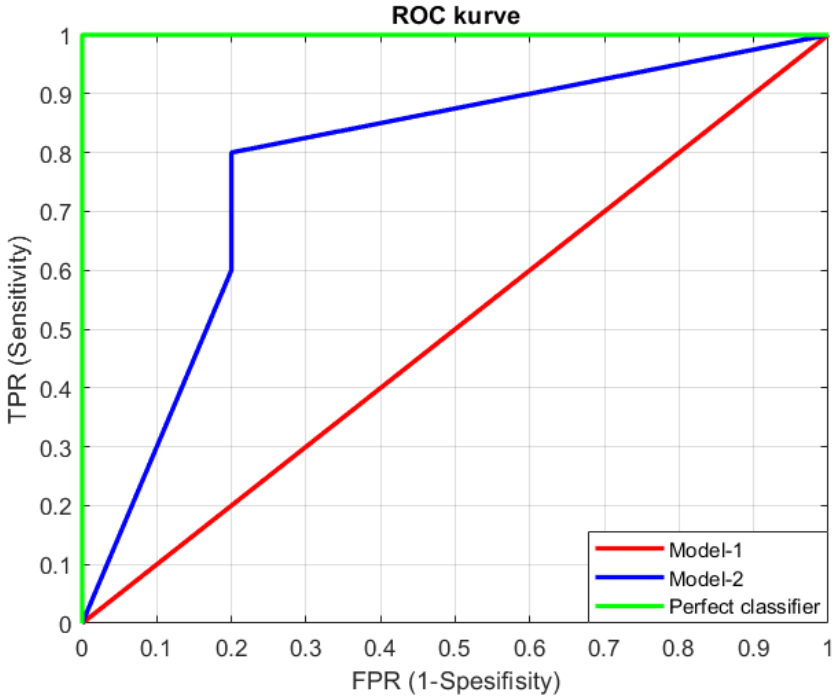
Sensitivitet	Spesifisitet
0,8	0,8

Tabell 5 viser sensitivitet og spesifisitet i tumorgjenkjenningsmodell-2, basert på referansestandard. Sensitiviteten har og spesifisiteten samme utfall 0,8.

4.2.2 ROC- KURVE

ROC-kurven baseres seg på referansetestens resultater, i form av TPR og FPR. Begge modellene blir inkludert i den samme kurven (Graf 4.1). Denne ROC-kurven er utarbeidet i MATLAB, der treshold er satt til 100%, 80%, 70% og 0% i alle kurvene. Endring av treshold gir lite utslag for modell-1 fordi den har en sannsynlighet på 100% for alle testbildene.

Grønn-kurve viser en "perfect classifier", som tilsvarer kriteriene til TPR og FPR, ved en optimal modell (perfekte resultater). Den blå kurven viser modell-2, hvor det har blitt gjort en endring i modellens treshold. TPR får en bratt kurve opp mot punkt (0,8) for så å "flate" ut i en diagonal linje. Den røde kurven viser modell-1, som oppnår lavere verdier i både TPR og FPR og dette vises som en lineær graf.



Figur 5. ROC kurve for modell-1 og modell-2

Grafen i Figur 5 viser sensitivitetsverdiene (TPR) og 1- spesifisitetsverdiene (FPR) ut ifra resultatene i modell-1 og modell-2.

Formel 1. Areal under grønn graf:

$$(1 \times 1) = 1$$

Formel 2. Areal under blå graf:

$$(0.2 \times 0.6) \times \frac{1}{2} + 0.8 \times (1 - 0.2) + (1 - 0.2) \times (1 - 0.8) \times \frac{1}{2} = 0.78$$

Formel 3. Areal under rød graf:

$$(1 \times 1) \times \frac{1}{2} = 0{,}5$$

AUC måler kvaliteten på modellens predikasjoner, uavhengig av klassifiseringsterskel. AUC går fra 0 til 1, der 1 er det optimale resultatet.

Ved en slik utregning vises det, at den grønne grafen “perfect classifier”, oppnår et resultat på 1(100%) som er det mest optimale resultatet modellene kan oppnå. Modell-2 (blå graf) har fått et resultat på 0,78 som betyr at modell-2 oppnår 78% av “perfect classifier” og modell-1

(rød graf) har fått et resultat på 0,5 som betyr at modell-1 oppnår 50% av “perfect classifier”.

5 DISKUSJON

5.1 DISKUSJON AV METODE

Hensikten med oppgaven er hvordan det utarbeides en bildegjenkjennings-modell som identifiserer tumor i bildemateriale. Metoden gir gode resultater. Ved å utforme en referansestandard, ble det mulig å utføre en referansetest som indikerer sensitiviteten og spesifisiteten til hver av de to modellene. Sammenligning av modellenes resultater viser at spesielt spesifisiteten øker, ved økt mengde bildemateriale. Derfor kan det antas at mengde treningsmateriale spiller inn på pålideligheten og nøyaktigheten, når en slik modell utvikles.

Metoden diagnostiske tester, som har blitt benyttet i denne oppgaven kan være utfordrende å diskutere. Grunnet at resultatene til de to modellene som blir benyttet gir svært spesifikke resultater, i form av tall. Både metoden og oppgaven er svært teknisk utarbeidet. Det gjør det vanskelig å diskutere så konkrete resultater, som består av tall.

5.1.1 Fordeler og Ulemper ved metoden

Utfordringer med en slik metode er, uansett hvor godt utarbeidet modellen er, så trengs det mye større mengde med bildemateriale for å kunne utvikle en pålitelig modell. Når det utvikles en modell innen KI som er tilpasset bildediagnostikken, burde det utvikles en algoritme fra bunn.

Fordelen er at metoden er brukervennlig og enkel å bruke. Derfor var det mulig for oss som radiografstudenter å lage en modell basert på oppbygging av KI-algoritmer. I tillegg kan en slik enkel modell gi indikasjoner på om dette er noe som er verdt å forske på, i en mer avansert “utforming”, ved bruk av andre mer spesifikke KI-metoder innenfor medisinsk bildediagnostikk.

Ulempen med en slik metode, er at kodingen og utarbeidelsen av algoritmen ikke er laget for medisinsk bildemateriale. Det biblioteket som Xcode består av, er i hovedsak laget for å utvikle enkle applikasjoner som spill (AppleInc., u.å.). Eksempelet som selskapet Apple viser til ved bruk av Xcode og deres algoritme, består av gjenkjennelse av ulike dyrearter. Derfor kan det antas at det er flere viktige klassifiseringskriterier som vil forsvinne, når det benyttes

medisinsk diagnostisk bildemateriale. Bildemateriale består kun av svart/hvit fargesammensetning og hjernevevets strukturer er relativt like. Derfor kan det tenkes at det er vanskelig for modellen å skille mellom ulike strukturer innad i bildemateriale. Det er kun små detaljer i bildene som vil kunne skille det diagnostiske bildematerialet fra hverandre (Elster, 2020). Derfor kan det antas at hvis en slik tumorgjenkjennelses-modell skal utvikles, slik at den kan benyttes i medisinsk bildediagnostikk, må man utvikle modellen med tilhørende algoritme helt fra bunn. Da vil det være mulig å kode modellens algoritme spesifikt etter ønskede egenskaper og strukturer. På denne måten vil bildemateriale bli attribuert og klassifisert på riktig måte. Dette ville gjort modellen bedre "utstyrt" for bruk innen medisinsk bildediagnostikk.

5.2 BILDEMATERIALE

Utvalgt bildemateriale ville optimalt sett hatt en høyere bildekvalitet ved tilgang på den originale undersøkelsen. Det ble tatt diagnostiske bilder fra en offentlig nettside for å samle inn datamaterialet. Det kan tenkes at dette ikke er den mest optimale måten å innhente diagnostisk bildemateriale på, som skal benyttes i en slik modell. Derfor kan antas at bildekvaliteten er redusert, som igjen kan føre til at modellen har mistet viktig informasjon i gitt bildemateriale. I tillegg vil mest sannsynlig dårlig bildekvalitet gjøre det vanskeligere for modellen å lære seg skillet mellom tumor og friskt hjernevev.

Når det utarbeides en slik bildegjenkjennings-modell vil det mest optimale vært direkte tilgang på store mengder bildemateriale, hvor bildemateriale er kvalitetssikret. Dette ville ikke vært mulig i denne oppgaven, da personvern og godkjenninger ville krevd mer tid og ressurser enn det som er disponibelt ved en bacheloroppgave.

I denne oppgaven er det kun blitt benyttet bildemateriale i aksialt plan. Det vil si, at det ikke er gjort målinger og klassifikasjoner i sagittal og koronalt plan, ved de ulike MR-caput undersøkelsene som benyttes. Derfor ble det ikke ble detektert patologi fra alle vinkler med modellene. Dette kan ha ført til manglende informasjon i det gitt bildemateriale.

Utvalgt bildemateriale som ble benyttet besto av hjernevev med- og uten tumor. Det bildematerialet som inneholdt tumor, besto av ulike "tumor"-diagnoser. Alt bildemateriale besto av T_2 vektet bildesekvens, i aksialt plan. Det kan tenkes at det ligger strukturer i hjernevevet som kan gjøre det vanskelig for modellene å gjenkjenne en eventuell tumor. Et

eksempel på slike strukturer kan være ventriklene. Det er store og væskefylte strukturer, som kan oppleves varierende i anatomisk utseende ved forskjellige pasienter. I utvalgt bildemateriale både ved trening og test, vises ventriklene i "forskjellige" størrelser på de diagnostiske bildene. Dette skyldes at snittene som er valgt ut, ikke befinner seg i eksakt samme området av caput ved de forskjellige MR-undersøkelsene. Derfor kan det tenkes at modell-2 i denne oppgaven, som oppnådde høyeste spesifisitet, beregnet en sannsynlighet på 100% "tumor" ved test-bilde 6. Ventriklene er veldig synlige, og derfor kan det tenkes at modellen har forvekslet dette med en tumor basert på gitt treningsmateriale.

Ved test-bilde 10, har modell-2 beregnet en sannsynlighet på 100% "uten tumor". Studeres alle test-bildene, vises det at tumoren som ligger på høyre hemisfære i test-bilde 10 har en "annen" gråtone enn de resterende test-bildene som inneholder en tumor. Gråtonene i test-bilde 10 "glir" mer over i hverandre og derfor kan det antas at dette er grunnen til at modell-2 kan ha hatt problemer med å skille tumoren fra det resterende hjernevev rundt tumor. Dette kan ha påvirket sensitiviteten og spesifisiteten i modell-2.

5.2.1 Feilkilder i bildematerialet

Bildematerialet ble hentet ut fra radiopedia.org, ved bruk av skjermbilder fra et utvalgt snitt. Dette bildematerialet består av ulike former for tumorer/patologi, noe som kan tenkes å ha påvirket resultatet, i form av at alle tumorene ser ulike ut. Derfor kan det tenkes at, det vil være vanskeligere for modellen å gjenkjenne tumoren. Dersom det hadde blitt benyttet samme type tumor, kunne modellen lettere gjenkjent mønster og klassifisert tumoren med en høyrere sensitivitet og spesifisitet. Det var vanskelig å finne tumorer i akkurat det "samme" snittet med likt utseende, i hver enkelt undersøkelse, noe som kan ha påvirket modellens pålitelighet.

Modellene er ikke basert på store mengder treningsmateriale og det utvalgte bildemateriale er tatt fra MR-undersøkelser i ulike "snitt" med ulik bildekvalitet. Det kan derfor tenkes å ha påvirket modellenes kapasitet til å analysere gitt bildematerialet.

Det kan stilles kritiske spørsmål til det utvalgte bildematerialet som besto av friskt hjernevev. Selv om snittet som ble utvalgt ikke inneholdt patologi, ble det brukt undersøkelser der det var patologi i andre snitt enn det utvalgte.

5.3 XCODE RESULTAT

I denne oppgaven ble det benyttet et eget program, Xcode med tilhørende eksternt bibliotek "image classifier"; en ferdig kode som gjør det enkelt å implementere og organisere de diagnostiske bildene (AppleInc, u.å.). Modellenes algoritme ble altså ikke bygd opp fra bunn, noe som betyr at den ikke er spesielt beregnet for svart/hvite bilder og for denne spesifikke oppgaven. Ved å bygge en algoritme selv fra bunn, kan man i en større grad styre hvilke kriterier algoritmen vektlegger, og vil da kunne øke nøyaktigheten til modellen (McBee et. al, 2018, s.1475). Xcode er ikke laget for å tolke medisinske bilder, men for å lage applikasjoner og spill. Det kan være utfordrende å lage en modell som skal tolke medisinske diagnostiske bilder uten farger, da fargebilder inneholder større kontrastforskjeller som kan gjøre det enklere for maskinen å tolke bildemateriale (AppleInc, u.å).

For at modellen skal bli så pålitelig som mulig, trengs det store mengder bildemateriale som kan benyttes som treningsmateriale og test-materiale (AppleInc., u.å.). Grunnet mangel på tid og ressurser, er benyttet bildemateriale hentet fra nettside og består ikke av store mengder. Derfor kan det tenkes at modellene som utarbeidet i denne oppgaven ville hatt en høyere pålitelighet, dersom det var tilgang til større mengder med bildemateriale, som for eksempel en egen database.

Datamaterialet fra en diagnostisk undersøkelse kan bestå av store mengder bildemateriale, som for eksempel en MR-undersøkelse. Originalt inneholder en MR-undersøkelse store mengder bildemateriale som skal analyseres, snitt for snitt. Dette gjør at en DL-modell må ha komplekse beregningsalgoritmer, som er meget nøyaktige ved analysering av medisinsk diagnostisk datamateriale (Currie et. al, 2019, s. 481). Datamaterialet som finnes innenfor radiologien er veldig variert. Ulike former for patologi kan føre til at gitt datamaterialet er forskjellig. Ulike personer kan se anatomisk ulike ut innvendig. Dette gjør at det kan bli komplisert å trene opp en pålitelig og nøyaktig modell (McBee et. al, 2018, s. 1476).

I to av de diagnostiske undersøkelsene, test-bilde 6 og test-bilde 10, mener modell-2 med 100% sannsynlighet at bildematerialet inneholder det modellen sier, men likevel er avgjørelsen feil. Dette indikerer at de to test-bildene oppfyller kriteriene til den trente modellen, som igjen betyr at modellen ikke er godt nok trent. Dette kommer av det begrenset mengde bildemateriale som ble benyttet til trening.

Det valgte bildemateriale i denne oppgaven er aksialt plan. Det ble strebet etter å oppnå valgt "snitt" fra samme område. Det kan tenkes at på grunn av at det ikke har blitt valgt ut "snitt"

fra nøyaktig det samme området av caput, kan dette ha påvirket modellenes utfall. Modellenes evne til å klassifisere og skille ulike deler av hjernevevet kan ha blitt påvirket, da det det kan oppleves at det valgte bildemateriale inneholder ulike strukturer. Dette kan ha forstyrret opptreningen av modellene. For eksempel, kan sees at ventriklene ha noe variert størrelse i det valgte bildematerialet. Det kan derfor antas at ventriklene kan ha påvirket modellenes utfall ved å "gjette" tumor på bildemateriale uten tumor. Dette på bakgrunn av at ventriklene er store strukturer i hjernen, som kan minne om store tumorer når en slik modell utvikles. Dette baseres på gitt bildemateriale som modellene er opptrent på.

5.3.1 Iterasjoner

Både modell-1 og modell-2 benytter to iterasjoner. Det er Xcode selv som stopper antall iterasjoner, og ifølge Apple er dette fordi Xcode mener den har oppnådd maksimal respons ut ifra det gitte settet med treningsmateriale (AppleInc, u.å.). Det kan tenkes at ved en større mengde bildemateriale, ville algoritmen som Xcode benytter kjørt treningsmaterialet gjennom flere iterasjoner. Modellene ville brukt mer tid på å gjenkjenne, klassifisere egenskaper og strukturer i det gitte bildebildematerialet. Som for eksempel, modell-1 ble innstilt med 20 iterasjoner, men stoppet automatisk etter to. Dette er på grunn av at modellen er innebygd med en algoritme som klassifiserer egenskaper, strukturer og gjenkjenner mønster i det gitte bildemateriale. Etter to iterasjoner, avsluttet modellen selv iterasjonene, fordi den har oppnådd optimal respons. På bakgrunn av dette blir treningsmaterialet kun kjørt gjennom modellens algoritme to ganger.

5.4 SENSITIVITET OG SPESIFISITET

Ideelt sett skulle det blitt brukt et bibliotek med flere tusen diagnostiske bilder i forskjellige plan (sagittale, koronale og aksiale). Dermed kan det antas at det ville ha styrket sensitiviteten og spesifisiteten i oppgaven hvis det hadde vært tilgang på en større mengde bildemateriale.

Det kan tenkes at modell-1, som blir trent på en liten mengde bildemateriale, oppnår en spesifisitet lik 0, fordi modellen indikerer at det finnes en tumor i samtlige test-bilder som blir gitt i test-materialet, og at det er derfor modell-1 oppnår en lav pålitelighet. I Modell-1 er 50% av resultatene "false positive", hvor modellen beregner en 100% sannsynlighet for funn av "tumor". Dette i motsetning til modell-2, som ble trent med en større mengde bildemateriale.

På bakgrunn av at modell-2 baseres på en større mengde bildemateriale enn modell-1, kan det antas at modell-2 oppnår en høyere spesifisitet enn modell-1.

Ved å se på det samlede resultatet har modell-1 en sensitivitet lik 1, mens modell-2 oppnår en sensitivitet på 0,8. Dermed har modell-1 en høyere sensitivitet enn modell-2, men ut ifra mengde bildemateriale har modell-1 null i spesifisitet. Det kan antas at modell-1 oppnår en høyere sensitivitet enn modell-2, fordi modell-1 beregner at det finnes "tumor" i alt bildemateriale med en sannsynlighet på 100%. Modell-2 beregner test-bilde 7 med 76% sannsynlighet for funn av "tumor", og derfor kan det tenkes at dette er grunnen til at modell-2 ikke oppnår like høy sensitivitet som modell-1.

Resultatet i referansetesten gikk ut ifra hvor mange "korrekte" og "ukorrekte" vurderinger modellenes test gir. Modell-1 vurderer alt bildemateriale med innhold av "tumor", derfor gir referansetesten også en høy sensitivitet. Ved sammenligning av modell-1 og modell-2, kan det tenkes at spesifisiteten til begge modellene trolig vil øke ved å benytte en større mengde bildemateriale under trenings-fasen.

Det kan antas at med økt mengde treningsmateriale, hvor modellene trenes til å gjenkjenne tumor i flere plan og snitt, ville det blitt gjennomført flere iterasjoner under trenings-fasen. Dermed kan det tenkes at det ville blitt oppnå en høyere sensitivitet og spesifisitet, noe som ville ha styrket resultatene i funnene. Det at det kun ble benyttet totalt 20 diagnostiske bildeundersøkelser i aksialt plan under trenings-fasen, kan ha påvirket modell-2 i en slik grad at resultatene i oppgaven ikke oppnår ønsket spesifisitet. Det kan derfor tenkes at sensitiviteten og spesifisiteten i modellen trolig økes ved å benytte en større mengde bildemateriale under trenings-fasen.

5.4.1 ROC-kurve

Når det ble utført en referansetest tilhørende hver modell, ble det mulig å beregne sensitiviteten og spesifisiteten, som baseres på de forutsetningene hver modell har, i form av mengde treningsmateriale. Ingen av modellene besto av store mengder treningsmateriale. Dette gjorde det utfordrende å opprette en ROC-kurve, da endringene i treshold ble for "små" til å gjøre store utslag i kurven. Det kan tenkes at den lave mengden med treningsmateriale ga utslag for at mange av test-bildene fikk resultater som tilsier "100% confidence". Dette gjelder både modell-1 og modell-2. Når det gjøres endringer i treshold kan det redusere eller øke sensitiviteten til en slik modell. Når treshold reduseres, vil en lavere sannsynlighet bli

"godtatt" av modellen. I test-bilde 7 ble det oppnådd en sannsynlighet på 76% ved modell-2, ved å justere treshold til 80%, blir modellen mer sensitiv og vil derfor ikke akseptere en sannsynlighet som er under 80%.

I ROC-kurven vises denne endringen i treshold i modell-1 og modell-2.
I Modell-1 hvor alt test-materiale oppnår en sannsynlighet på 100%, gir dette også en høy sensitivitet, men en lav spesifisitet. Dette gjenspeiles i ROC-kurven, fordi treshold på under 1 sender modellen til punkt (1,1). Det vises at modell-1 gir mange "false positive" resultat. Dette gjør at modell-1 oppnår et unøyaktig resultat. Modell-2 har en høyere spesifisitet og derfor vises det at kurven ligge høyere oppe på Y-aksen, og er nærmere en "perfect classifier". Ved å analysere AUC fremtrer det stor forskjell på modell-1 og modell-2 i form av pålitelighet. Det kan antas at dersom en slik modell skal bli brukt i medisinsk bildediagnostikk, burde AUC være så nærme 1 som mulig.

6 KONKLUSJON

I denne oppgaven ble det utviklet to modeller, modell-1 og modell-2 ved hjelp av KI og et kodingsprogram, Xcode med tilhørende tester. Dette programmet gjorde det mulig å implementere en ferdiglaget DL-algoritme. Modellene ble trent med et utvalgt bildemateriale, for så å bli testet med et annet utvalgt bildemateriale. Dette bildematerialet besto av MR-caput bildemateriale, med- og uten tumor, i aksialt plan. Sensitiviteten og spesifisiteten indikerer at modellene ble mer pålitelig ved større mengde bildemateriale. Modellene var basert på ulik mengde treningsmateriale. Ved å utvikle to modeller kunne vi analysere og sammenligne resultatene, hvor det hovedsakelig ble lagt vekt på forskjellen i sensitivitet og spesifisitet. Ved en slik sammenligning kunne vi besvare problemstillingen: *"Hvordan kan en modell utvikles ved hjelp av kunstig intelligens, trenes og benyttes for gjenkjenning av tumor i MR-caput bildemateriale?"*

De utførte testene viser påliteligheten og nøyaktigheten til de to forskjellige modellene, som er basert på gitt mengde treningsmateriale. Modell-1 trenes med et lite sett datamateriale, det gjenspeiles i et lite trolig resultat hvor sensitivitet er lik 0,5 og spesifisitet er lik 0. Modell-2 trenes med et større sett bildemateriale hvor resultatene er noe mer pålitelig, både sensitivitet og spesifisitet er lik 0,8. Resultatene indikerer hvor viktig det er med store mengder treningsdata når en slik type modell skal trenes opp til å gjenkjenne mønster og klassifisere

lesjoner i bildemateriale innenfor KI og DL. Gjennom hver iterasjon modellen gjennomfører, konvergerer den matematiske løsningen til en mer nøyaktig løsning. Treningsfasen oppnår best resultat med en større mengde bildemateriale. Hvis det skulle blitt utvikle en slik reel modell i bildediagnostikken, trengs det en modell som benytter et CNN, med tilhørende DL-algoritmer som er bygd opp fra bunnen. En slik type modell vil kreve mange år å utvikle, hvor modellens koder skal klassifiseres og spesialiseres ned til hver eneste minste detalj.

ROC-kurven som ble utviklet i denne oppgaven gir en indikasjon på hvor treshold bør settes for å få et best mulig resultat. AUC sier noe om hvor pålitelig modellen er, og resultatene i denne oppgaven viser at modell-2 har større AUC enn modell-1 og er derfor mer pålitelig. Å analysere en ROC-kurve kan være nyttig for å kunne utvikle en modell med høy pålitelighet.

Derfor kan man ved hjelp av KI, utvikle en modell som gjenkjenner tumorer. Basert på resultatene som oppnås i denne oppgaven, vises det at det trengs store mengder med treningsmateriale for at modellen eventuelt kan brukes innen medisinsk bildediagnostikk. Resultatene viser også hvordan en ROC-kurve kan benyttes for å utvikle en mest mulig pålitelig modell.

6.1 Videre forskning

KI er et fagfelt som det definitivt er verdt å forske på, spesielt innenfor bildediagnostikken. Det oppleves en økt pågang i antall undersøkelser og dermed trengs det mer ressurser innenfor fagfeltet.

Denne oppgaven viser at ved hjelp av enkel koding kan en modell klassifisere, strukturere og gjenkjenne mønstre i gitt bildemateriale. En slik type modell indikerer at dette er en metode det kan bygges videre på, hvis det utvikles en modell med tilhørende algoritmer som er spesiallaget for medisinsk bildediagnostikk. Det kan tenkes at det bør være tilgang på store mengder bildemateriale ved en slik modell-utvikling. Store mengder med bildemateriale kan for eksempel være tilgang på egne databaser med radiologisk bildemateriale.

Ved Oslo universitetssykehus er det allerede i gang et forskningsprosjekt innen hjernetumorer og KI. Dette prosjektet har fått 14 millioner kroner i støtte fra det europeiske forskningsrådet (ERC) for å forske på KI og hjernekreft. Denne forskningsgruppen skal finne ut hvordan KI kan hjelpe radiologene med å diagnostisere pasienter og følge med på hvordan sykdommen

utvikler seg. KI kan hjelpe radiologen og definere hva som er friskt og sykt vev i hjernen. Da kan det bli enklere å bestemme hva som er kreft på bildene (Vogt, 2019).

7 LITTERATURLISTE

Apeland. (2015). Hva er en algoritme? Hentet fra

https://www.apeland.no/hva-er-en-algoritme/

AppleInc. (u. å). Devveloper - Creating an Image Classifier Model. Hentet 1.april 2020 fra

https://developer.apple.com/documentation/createml/creating_an_image_classifier_model?fbclid=IwAR1m4nA3_7VaCPfLN_mKK9zW3Y6VWzg_TYZsscoT6gJpcHv-XnURgVjozXs

Bergsjø, L., O. & Bergsjø, H. (2019). *Digital etikk: Big data, algoritmer og kunstig intelligens.* Oslo: Universitetsforlaget.

Bjørkeng, P. (2019). *Kunstig intelligens, Den usynlige revolusjonen* (2.utg.). Oslo: Verga Forlag AS.

Currie, G., Hawk, E., K, Rohren, E., Vial, A., Klein, R. (2019). Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. *Journal of Medical Imaging and Radiation Sciences, 50,* 477-487. https://doi.org/10.1016/j.jmir.2019.09.005

Dyrdal, I., Aurdal, L., Løkken, K.H. & Engøy, T. (2017). *Teknologiske muligheter for tolletaten, mønstergjenkjenning og maskinlære* (FFI-RAPPORT 17/17026). Hentet fra https://publications.ffi.no/nb/item/asset/dspace:4230/17-17026.pdf

Elster, D., A. (2020). Question and answers in MRI.

Hentet 27. April 2020 fra https://mriquestions.com/index.html

Geis, R. J., Brady, A., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J., … Kohli, M. (2019). Ethics of artificial intelligence in radiology: summary of the joint European an North American multisociety statement. *Insights into imaging, 10* (101), 2-6. https://doi.org/10.1186/s13244-019-0785-8

Groote, I., R. (2018). Robotene kommer! *Tidsskriftet for Den Norske Legeforening.*

doi: 10.4045/tidsskr.18.0795

Helsedirektoratet. (2016). Diagnostiske tester. Hentet fra

https://www.helsebiblioteket.no/kunnskapsbasert-praksis/kritisk-vurdering/diagnostiske-tester

Kommunal- og moderniseringsdepartementet. (2020). *Nasjonal strategi for kunstig Intelligens* (plan/strategi (14.01.2020)). Hentet fra https://www.regjeringen.no/contentassets/1febbbb2c4fd4b7d92c67ddd353b6ae8/no/pdfs/ki-strategi.pdf

Kreftregistret. (2019, 18. november). MIM-studien: MASKINLÆRING I MAMMOGRAFIPROGRAMMET. Hentet fra https://www.kreftregisteret.no/Forskning/Prosjekter/maskinlaring-i-mammografiprogrammet/

Lydersen, S. (2017). Hva er sannsynligheten for riktig resultat av en diagnostisk test? *Tidsskriftet for Den Norske Legeforening*. doi: 10.4045/tidsskr.17.0409

McBee, M, P., Awan, A, T., Colucchi, A, T., Ghobadi, C, W., Kadom, N., Kansagra, A, P., … Auffermann, W, F. (2018). Deep Learning in Radiology. *Radiolgy research alliance, 25* (11), 1472- 1480. https://doi.org/10.1016/j.acra.2018.02.018

Mikaelsen, B. (u.å.). AI og fremtidens radiografer. Hentet 26.mars 2020 fra https://www.holdpusten.no/artikler/ai-og-fremtidens-radiografer/434127

Murphy, A. & Liszewsky, B. (2019). Artificial Intelligence and the Medical Radiation Profession: How Our Advocacy Must Inform Future Practice. *Journal of Medical Imaging and Radiation Sciences, 50,* 15-19. https://doi.org/10.1016/j.jmir.2019.09.001

Radiografforbundet. (2015, 23. juni). Yrkesetiske retningslinjer. Hentet fra https://www.radiograf.no/artikler/yrkesetiske-retningslinjer/436890

Radiopedia. (u.å.). Cases. Hentet 27.april fra https://radiopaedia.org/encyclopaedia/cases/all?lang=us

Sardanelli, F. & Di Leo, G. (2009). *Biostatistics for radiologists. Planning, performing and writing a radiologic study*. Trento, Italia; Springer-Verlag Mailand.

Telle, J., A. (2017). DEN NYE MASKINLÆRINGEN: KUNSTIG INTELLIGENS ELLER BARE GODE VERKTØY? *Nytt Norsk Tidsskrift, 34*, 192-204. doi:10.18261/issn.1504-3053-2017-02-08

Tørresen, J. (2013). *Hva er kunsting intelligens?* Oslo: Universitetsforlaget.

Vogt, Y. (2019, 8. november). Hjernekreft oppdages raskere med kunstig intelligens. *Forskningsmagasinet Apollon.* Hentet fra https://www.apollon.uio.no/artikler/2019/4_ai_hjernekreft.html

Warwick, Kevin. (2012). *Artificial intelligence: the basics*. New York: Routledge.

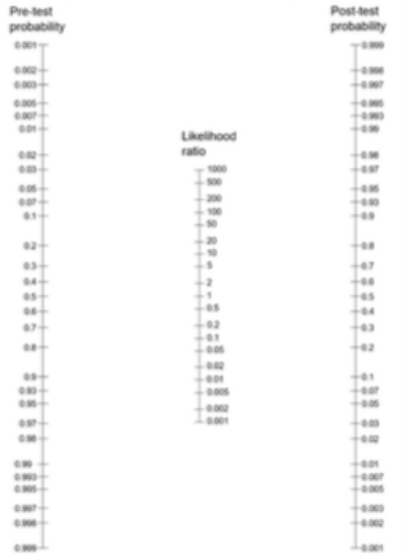
Woznitza, N. (2020). Artificial Intelligence and the Radiographer/Radiological Technologist Profession: A joint statement of the International Society of Radiographers and Radiological Technologists and the European Federation of Radiographer Societies. *Radiography, 26*, 93-95. Hentet fra https://www.radiographyonline.com/article/S1078-8174(20)30037-7/pdf?fbclid=IwAR0hXsDoM0xLsxCPY6C6uDMxsMwPKz-W8icCWd3_MnLsWU-8f_zbJOKtuT4

8 VEDLEGG

8.1 Vedlegg 1: Tabell; Referansestandard

Vedlegg: Definisjoner og formler for utregning

		Referansestandard		
		Syk	Frisk	Total
Test	Positive	a	b	a+b
	Negative	c	d	c+d
	Total	a+c	b+d	a+b+c+d



A Fagan nomogram (or Fagan's nomogram) for relations between Pre- and post-test probabilities and likelihood ratio. Wikimedia Commons [Nedlastet 09.05.2017]. Tilgjengelig fra https://commons.wikimedia.org/wiki/File:Fagan_nomogram.svg

Prevalens = (a+c)/(a+b+c+d) = pre-test sannsynlighet

Sensitivititet = a/(a+c)

Spesifisitet = d/(b+d)

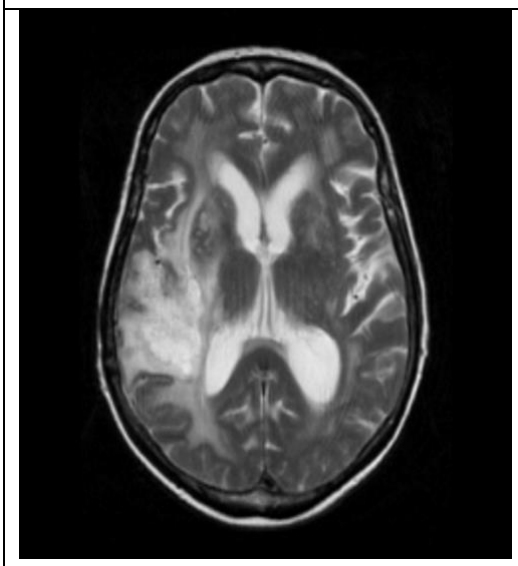
Likelihood-ratio for positivt testresultat:
LH+ = sensitivity /(1-specificity) =
(a/(a+c))/(b/(b+d))

Likelihood Ratio for negativt test resultat:
LH– = (1 – sensitivity)/specificity =
(c/a+c)/(d/(b+d))

Positive Predictive Value (PPV) = a/(a+b)

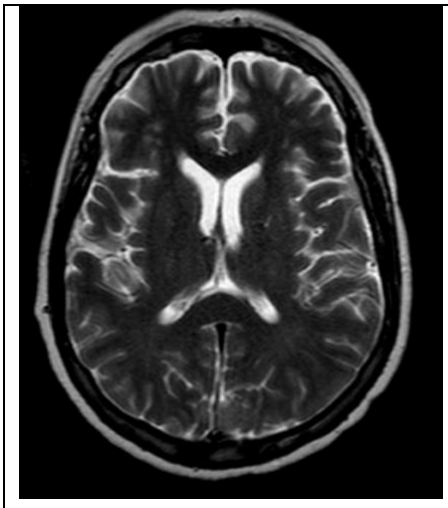
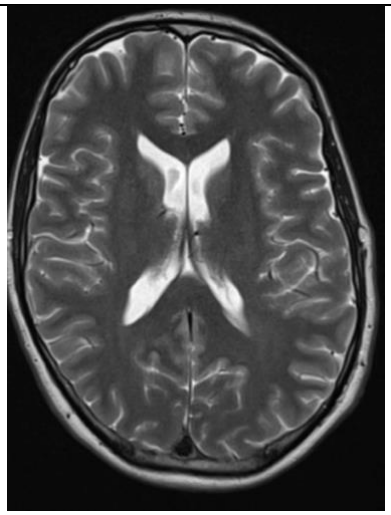
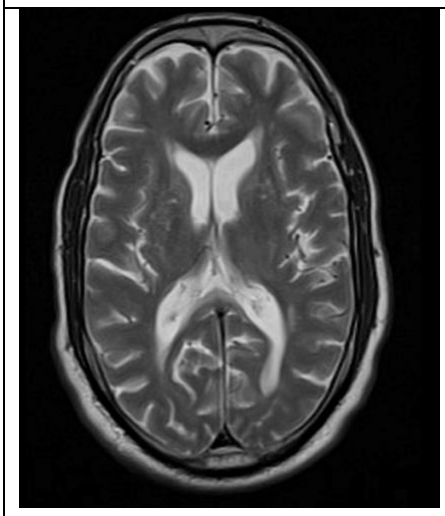
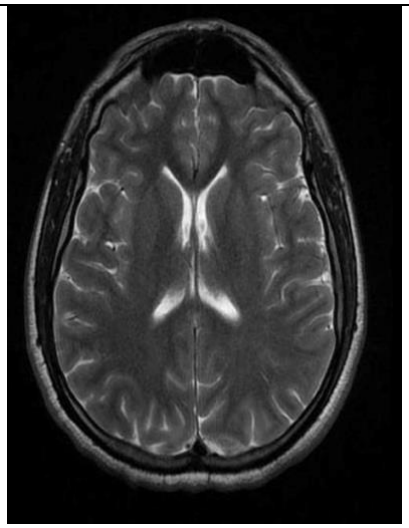
Negative Predictive Value (NPV) = d/(c+d)

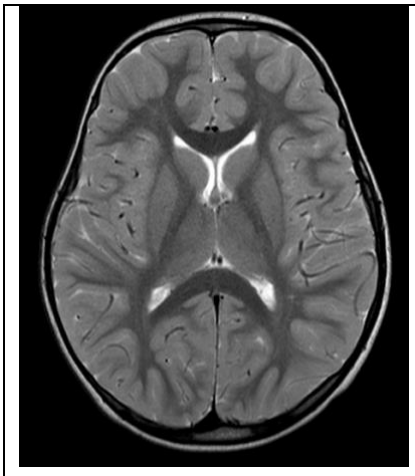
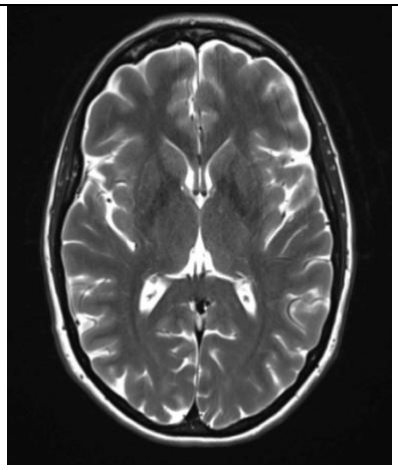
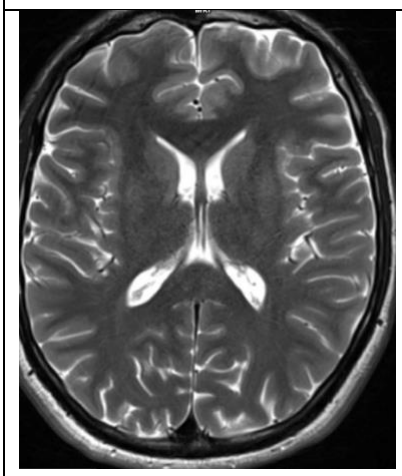
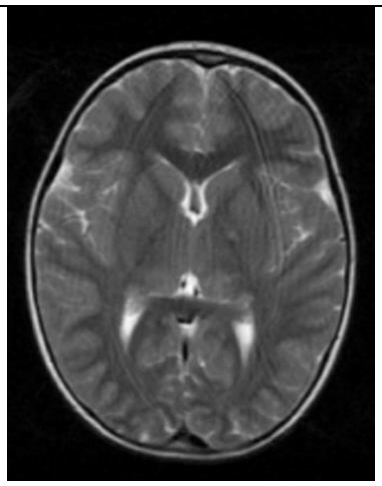
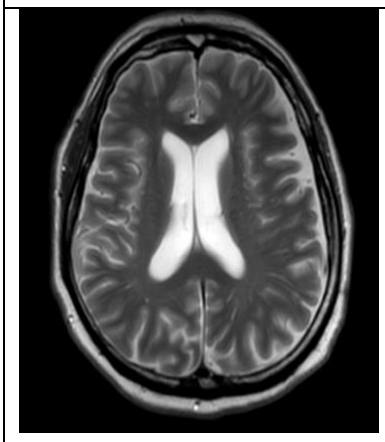
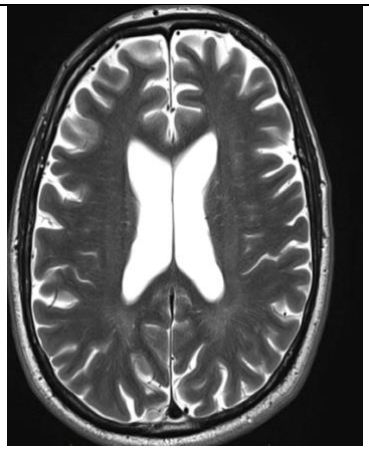
8.2 Vedlegg 1: Modell-1, Treningsmateriale

“Tumor”	“Uten tumor”
 Case courtesy of Dr Rodrigo Dias Duarte, Radiopaedia.org, rID: 60130	 Case courtesy of Dr Ian Bickle, Radiopaedia.org, rID: 52637

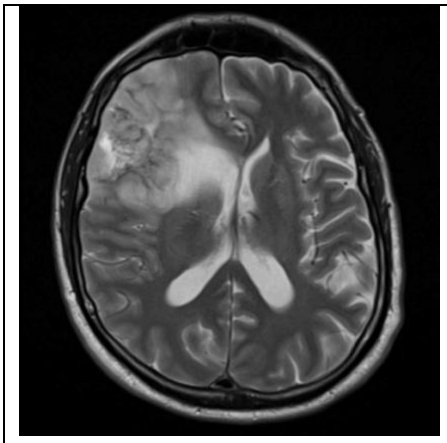
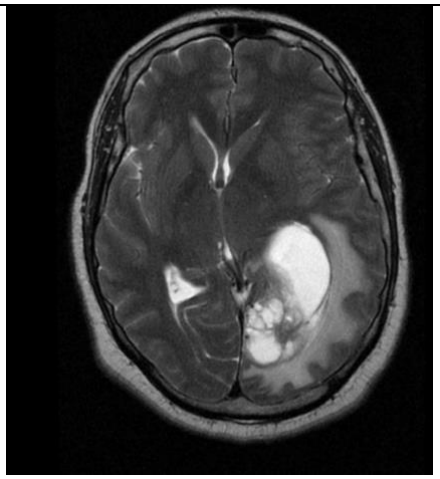
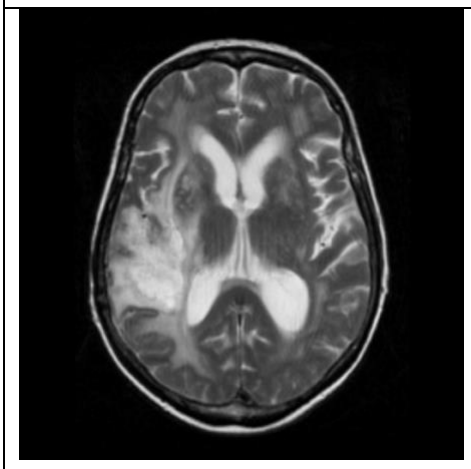
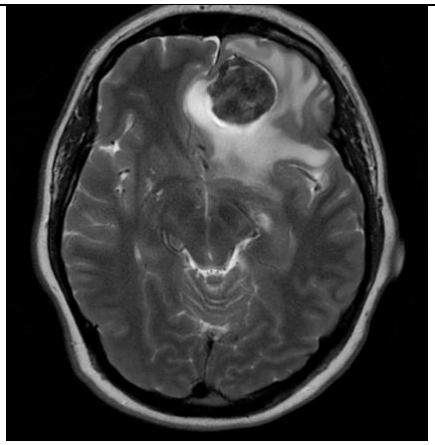
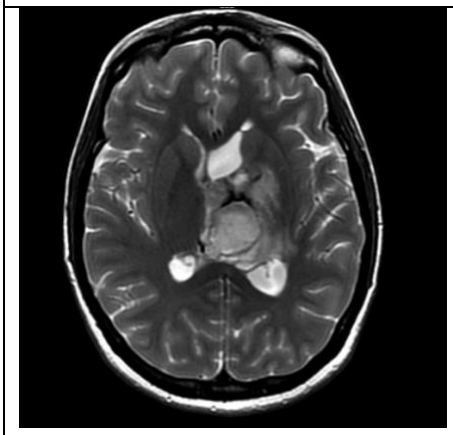
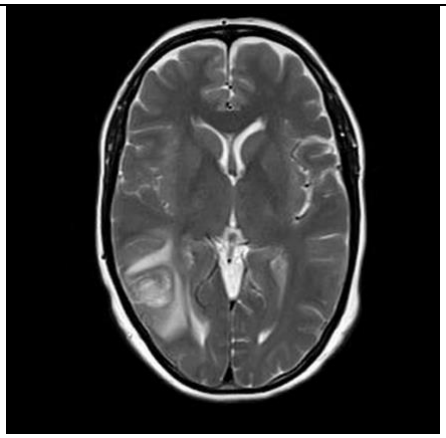
8.3 Vedlegg: Modell-2, Treningsmateriale

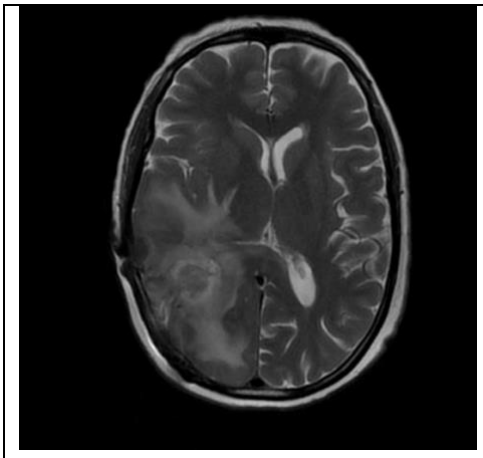
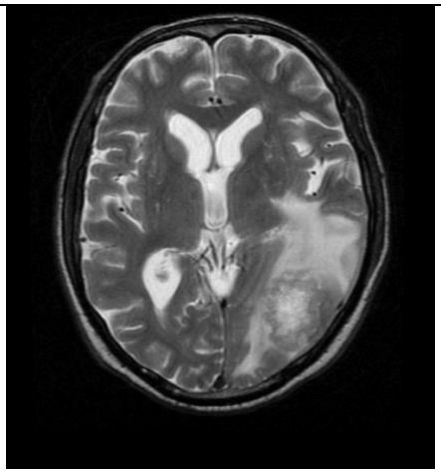
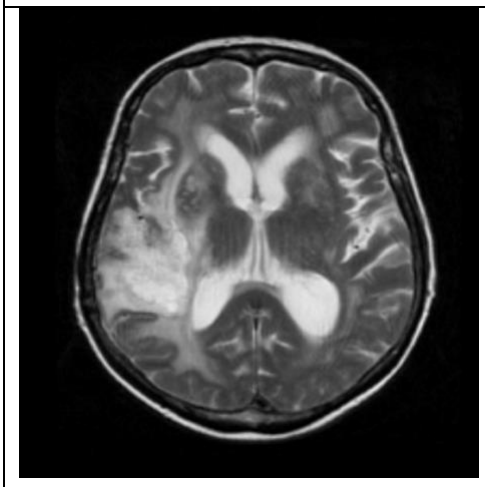
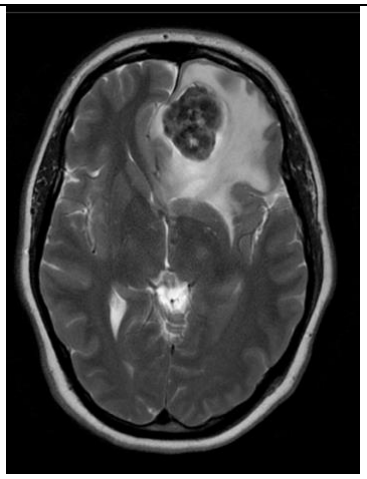
Treningsmateriale, uten tumor.

 Case courtesy of Dr Hani Salam, Radiopaedia.org, rID: 8713	 Case courtesy of Dr Bruno Di Muzio, Radiopaedia.org, rID: 39310
 Case courtesy of Dr Mahmoud Yacout Alabd, Radiopaedia.org, rID: 40092	 Case courtesy of Dr Ian Bickle, Radiopaedia.org, rID: 52637

 Case courtesy of Dr Mohammad A. ElBeialy, Radiopaedia.org, rID: 39578	 Case courtesy of Dr Bruno Di Muzio, Radiopaedia.org, rID: 55563
 Case courtesy of Dr Bruno Di Muzio, Radiopaedia.org, rID: 39310	 Case courtesy of Dr Vougiouklis Nikos, Radiopaedia.org, rID: 18352
 Case courtesy of Dr Anna Salwa Kaminska, Radiopaedia.org, rID: 73881	 Case courtesy of Dr Wen Jak Ma, Radiopaedia.org, rID: 65377

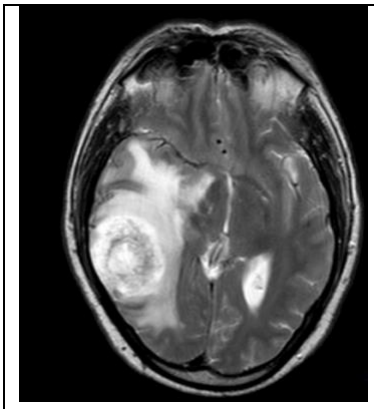
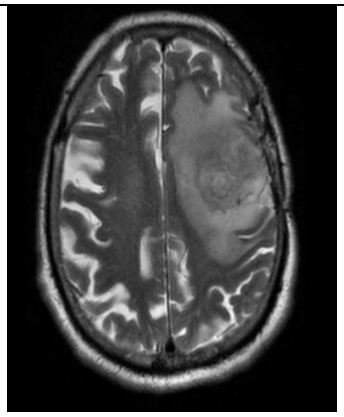
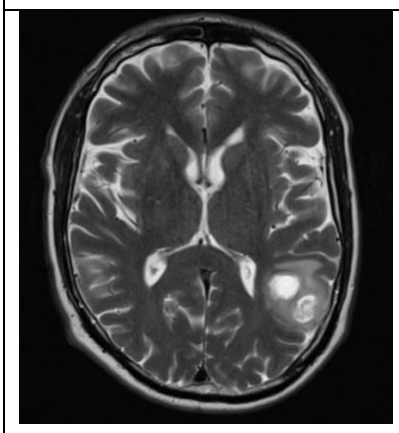
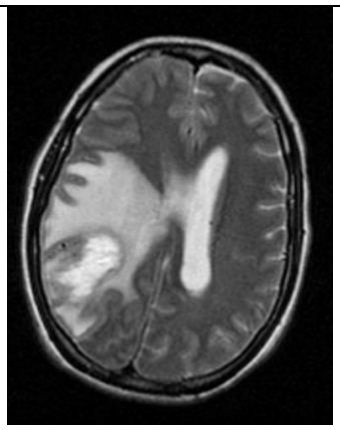
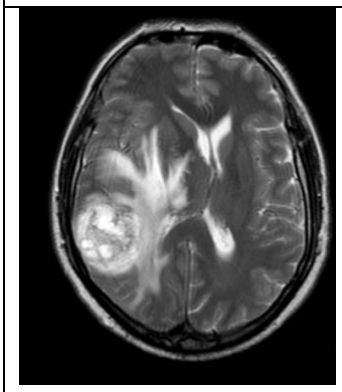
Treningsmateriale, tumor

 Case courtesy of Assoc Prof Frank Gaillard, Radiopaedia.org, rID: 55579	 Case courtesy of Dr Hani Salam, Radiopaedia.org, rID: 12349
 Case courtesy of Dr Rodrigo Dias Duarte, Radiopaedia.org, rID: 60130	 Case courtesy of Dr Andrew Lawson, Radiopaedia.org, rID: 26982
 Case courtesy of Assoc Prof Frank Gaillard, Radiopaedia.org, rID: 8771	 Case courtesy of Dr Bahman Rasuli, Radiopaedia.org, rID: 72618

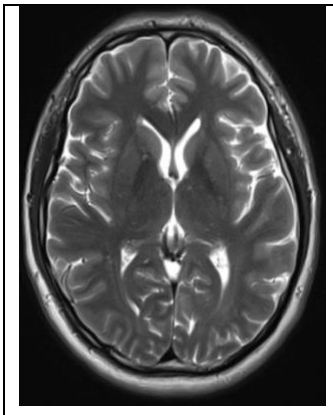
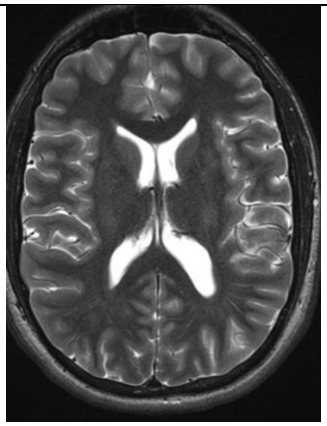
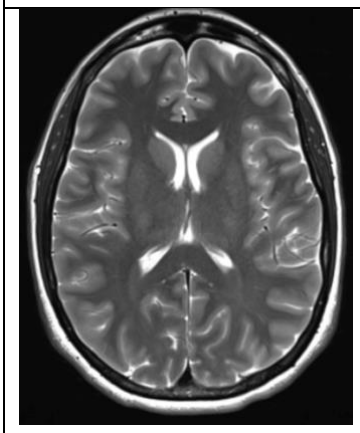
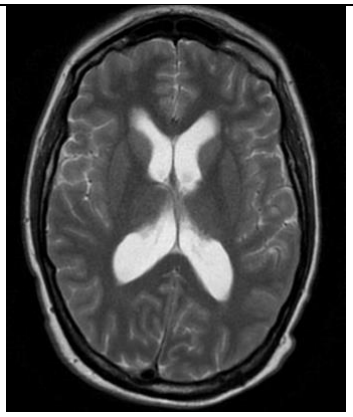
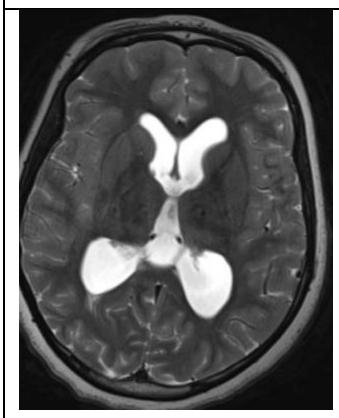
 Case courtesy of Dr Bahman Rasuli, Radiopaedia.org, rID: 72618	 Case courtesy of Dr Hani Salam, Radiopaedia.org, rID: 8581
 Case courtesy of Dr Rodrigo Dias Duarte, Radiopaedia.org, rID: 60130	 Case courtesy of Dr Andrew Lawson, Radiopaedia.org, rID: 26982

8.4 Vedlegg: Test-materiale

Test-materiale, tumor

 Case courtesy of Dr Ahmed Abdrabou, Radiopaedia.org, rID: 22898	 Case courtesy of Dr Bahman Rasuli, Radiopaedia.org, rID: 71869
 Case courtesy of Dr Heather Pascoe, Radiopaedia.org, rID: 58293	 Case courtesy of Dr Morlie L Wang, Radiopaedia.org, rID: 13281
 Case courtesy of Dr Ahmed Abdrabou, Radiopaedia.org, rID: 22898	

Test-materiale, uten tumor

 Case courtesy of Dr Bruno Di Muzio, Radiopaedia.org, rID: 41113	 Case courtesy of Dr Heba Abdelmonem, Radiopaedia.org, rID: 64968
 Case courtesy of Assoc Prof Frank Gaillard, Radiopaedia.org, rID: 42777	 Case courtesy of Dr Alexandra Stanislavsky, Radiopaedia.org, rID: 10755
 Case courtesy of Dr Jeremy Jones, Radiopaedia.org, rID: 21819	