Norges miljø- og
biovitenskapelige
universitet

Masteroppgave 2016 30 stp
Institutt for matematiske realfag og teknologi

Videreutvikling av et diagnostisk verktøy for automatisk svulstinntegning av livmorhalskreft i MR-bilder

Further Development of a Diagnostic Tool for Autodelineation of Cervical Cancers in MR Images

Elise Mühlbradt
Miljøfysikk og fornybar energi

Forord

Denne oppgaven er skrevet ved Institutt for matematiske realfag og teknologi ved Norges miljø- og biovitenskapelige universitet (NMBU) våren 2016. Oppgaven utgjør 30 studiepoeng og markerer avslutningen på en femårig mastergrad i Miljøfysikk og fornybar energi.

Aller først vil jeg rette en stor takk til min hovedveileder Cecilia Marie Futsæther for enestående hjelp, støtte og oppfølging, i tillegg til spennende diskusjoner og hyggelige samtaler. Takk til Turid Torheim som alltid har vært behjelpelig med programmering, og tilgjengelig for spørsmål og diskusjoner. Takk til Knut Kvaal for gode innspill og tilbakemeldinger, og hjelp med programvare og datamaskiner. Dere er en super gjeng, som jeg er veldig glad for å ha jobbet sammen med.

Videre vil jeg takke Heidi Lyng og Eirik Malinen for å ha skaffet datasettet, og radiologene Knut H. Holde og Kjersti Lund for å ha tegnet inn svulstmaskene. Takk til Erlend Kristoffer Frivold Andersen som har bearbeidet datasettet, og igjen takk til Turid Torheim for å utvikle svulstinntegningsprogrammet.

Til slutt vil jeg takke familie og venner som har oppmuntret og støttet meg under arbeidet med oppgaven. En spesiell takk til medstudentene på lesesalen for hjelp og støtte underveis. Takk til alle medstudenter for fem fine år på Ås, og for minner som aldri vil glemmes.

Ås, 18.mai 2016

_____

Elise Mühlbradt

Sammendrag

Svulstavgrensning og svulstinntegning innenfor medisinsk avbildning er en utfordrende, tidkrevende og stadig mer kompleks del av strålebehandling. I denne oppgaven videreutvikles et program som automatiserer svulstinntegningen, basert på MR-bilder av pasienter med livmorhalskreft. Programmet klassifiserer vokslene i MR-bildene som svulst eller ikke-svulst, med metoden lineær diskriminant analyse (LDA). I denne oppgaven testes nye metoder for (1) preprosessering av MR-bildene, (2) klassifisering av vokslene, og (3) postprosessering basert på klassifiseringsresultatene, for å øke nøyaktigheten til svulstinntegningene.

Oppgaven er basert på en studie gjennomført ved Radiumhospitalet i perioden 2001-2004. MR-avbildning av 78 kvinner med lokalavansert livmorhalskreft ble utført i forkant av behandling. Bildene består av T_1-vektede, T_2-vektede og dynamisk kontrastforsterkede (DCE) bildesekvenser. Samtlige pasienter mottok strålebehandling og kjemoterapi i etterkant. Livmorhalssvulstene ble tegnet inn av to radiologer. Radiologinntegningene ble brukt i modelltreningen og som fasit for å vurdere den automatiske inntegning.

For å forbedre analysegrunnlaget og kompensere for variasjoner i MR-bildenes intensitet mellom hver pasient, ble Median filter, Savitzky-Golay filter og Contrast Limited Adaptive Histogram Equalization (CLAHE) testet som preprosesseringsalgoritmer. En postprosesseringsmetode ble implementert, hvor en ROI (*Region of Interest*) maske tegnes rundt området hvor LDA klassifiseringen predikerte høy sannsynlighet for svulst. Dette fjerner irrelevante områder og reduserer vokselantallet, og det undersøkes om vokselklassifiseringen forbedres. Til slutt ble det undersøkt om ikke-lineære klassifiseringsmetoder skiller svulst og frisk vev mer nøyaktig enn den lineære klassifiseringsmetoden LDA. Klassifiseringsmetodene *Random Forest*, *AdaBoost*, k nærmeste nabo (kNN) og støtte vektor maskiner (SVM) ble testet.

Resultatene viser at postprosesseringen ved å velge en ROI maske ga en signifikant forbedring av vokselklassifiseringen. Hverken *Random Forest*, *AdaBoost* eller kNN klassifiseringsmetodene ga en signifikant forbedring av klassifiseringen. I tillegg brukte disse metodene vesentlig lenger tid på modelltreningen enn LDA. Uttesting av SVM klassifiseringsmetoden ble ikke fullført på grunn av for tidkrevende modelltrening. Ingen av preprosesseringsmetodene ga en forbedring av vokselklassifiseringen.

Den beste LDA klassifiseringen med ROI masken ga en gjennomsnittlig DSC og Kappa verdi på henholdsvis 0,53 og 0,51, og sensitivitet og spesifisitet på henholdsvis 0,91 og 0,81. Kappa verdien for LDA modellen var høyere enn forventet overenstemmelsen mellom radiologer, med en multirater Kappa på 0,32. Således viser det diagnostiske verktøyet for automatisk svulstinntegning av livmorhalskreft potensiale til å blir et nyttig verktøy for radiologer.

Abstract

Tumour delineation in medical imaging is a challenging, time-consuming and increasingly complex part of radiotherapy planning. The aim of this thesis is to further develop and improve an automatic tumour delineation program, based on MR-images of patients with cervical cancer. The program classifies voxels in MR-images into two classes, tumour and non-tumour, using linear discriminant analysis (LDA). In this thesis new methods were tested for (1) pre-processing of the MRI images, (2) classification of the voxels as tumour/non-tumour and (3) post-processing of the MRI images based on the classification results.

The analysis is based on a study done by the Norwegian Radium Hospital in the period 2001-2004. MRI imaging of 78 women with locally advanced cervical cancer was performed prior to treatment. The MRI images consist of T_1-weighted, T_2-weighted and dynamic contrast-enhanced (DCE) sequences. All patients received curative radiotherapy with adjuvant chemotherapy afterwards. The tumours were outlined by two radiologists. These contours were used in the modeltraining and as the ground truth to evaluate the autodelineation.

To compensate for intensity variations between the patients, the Median Filter, the Savitzky-Golay Filter and Contrast Limited Adaptive Histogram Equalization (CLAHE) were tested as pre-processing algorithms. To improve voxel classification, a post-processing method was implemented, where the results from the initial linear classification were used to draw a mask called Region of Interest (ROI) around the image region predicted to contain the tumour. This mask removed irrelevant areas and reduced the number of voxels. Finally, different classification methods were tested to investigate if non-linear classification models performed better than the linear classification model implemented in the program. The classifiers Random Forest, AdaBoost, k nearest neighbour (kNN) and support vector machines (SVM) were tested.

The analysis showed that post-processing by selecting a ROI mask significantly improved voxel classification. Neither Random Forest, AdaBoost nor the kNN classification models gave significantly better classification. In addition, these methods used significantly longer time to train the model, than LDA. The SVM classification method was not sufficiently tested because of time-consuming training. None of the pre-processing methods gave a better voxel classification.

The best LDA classification using the ROI post-processed images gave mean DSC and Kappa values of 0.53 and 0.51, respectively, and sensitivity and specificity values of 0.91 and 0.81, respectively. The Kappa values for the LDA model were higher than the expected agreement between radiologists, with a multirater Kappa of 0.32. Subsequently, the autodelineation program for cervical cancers has the potential to become a useful tool for radiologists.

Innholdsfortegnelse

1 Innledning

Livmorhalskreft er en av de vanligste kreftformene blant kvinner [1, 2], og årlig behandles rundt 3000 kvinner i Norge for celleforandringer eller forstadier til kreft i livmorhalsen [3]. I 2012 ble det på verdensbasis registrert 527 600 nye tilfeller av livmorhalskreft og 265 700 dødsfall knyttet til denne krefttypen [2]. Forskningsstudier av pasienter med livmorhalskreft viser at de fleste tilfeller av livmorhalskreft skyldes HPV viruset (*humant papillomavirus)* [1, 4].

Ved mistanke om livmorhalskreft undersøkes pasienten med CT (*Computed Tomography)*, MRI *(Magnetic Resonance Imaging)* og i noen tilfeller PET *(Positron Emission Tomography*). Strålebehandlingsplaner og doseberegninger utformes fra CT bildene [5]. Tradisjonelle CT bilder har ofte en lav bløtvev kontrast. MRI kan derimot vise deformerte strukturer, og differensiere normalt mykt vev så vel som tumor-infiltrert mykt vev [6]. I tillegg avgir ikke MRI ioniserende stråling. Denne avbildningsteknikken kan også avdekke fysiologiske radioresistente områder som deretter kan behandles med en økt strålingsdose til høyrisiko områder [7, 8]. Variasjoner i signalintensiteten grunnet bruk av ulike skannere, sekvenser og sekvens parametere er en av hovedutfordringene tilknyttet kvantitativ analyse av MRI bilder [9]. PET er en tredje avbildningsmetode som benyttes i kreftdiagnostikk. Denne teknikken brukes blant annet for å kunne skille mellom ondartede og godartede kreftsvulster [10], og unytter desintegrasjonsprosessen til kortlivede, ustabile, positronemitterende radionuklider [11].

Alder, stadium av kreft og krefttype er faktorer som påvirker valg av behandling, hvor stadium av kreft er den avgjørende faktoren [5]. Strålebehandling anbefales for en stor andel av pasientene som diagnostiseres med livmorhalskreft. I størst grad til de med et avansert stadium (*FIGO stage II-IV)* og til pasienter med tidligere stadier som ikke kan opereres [5, 12]. Utfordringen knyttet til strålingsterapi er å finne en nøyaktig avgrensing av kreftsvulsten slik at man kan maksimere strålingsdosen til kreftceller og minimalisere strålingsdosen til friskt vev [5, 7].

Det er nødvendig med manuell inntegning av livmorhalskreftsvulsten i de medisinske bildene for å kunne utføre strålingsterapi. Manuell avgrensing av svulsten er ofte tidkrevende og kan påvirkes av usikkerheter knyttet til inntegnerens spesialitet, trening og personlig variasjon [5, 7]. I tillegg finnes det usikkerheter tilknyttet bevegelse av kreftsvulsten, konfigureringsfeil på utstyr og pasient, behandlings-og doselevering, biologisk respons på strålingsdosen og pasientbevegelse [5, 7]. En nøyaktig avgrensing av GTV (*Gross Tumor Volume)* og CTV (*Clinically Treated Volume)* hentet fra MRI bilder kan redusere usikkerhetene knyttet til organgrenser/svulstgrenser og kan føre til bruk av en mindre grense, som igjen kan redusere

strålingen mottatt av frisk vev [5]. Stråling mottatt av frisk vev kan lede til strålingsforgifting og føre til skade på DNA og friske celler, eventuelt sekundær kreft [13].

Flere dataprogrammer for segmentering av organer og kreftsvulster blitt rapportert, og er nylig oppsummert av Ghose *et al*. [5]. Mange av metodene går ut på å dele opp digitale bilder inn i flere områder som er mer meningsfulle og lettere å analysere [5]. Non-rigid registering er ofte brukt i segmentering av organer i medisinsk avbildning [5, 14-16]. Dette er en registreringsmetode hvor målet er å maksimere likheten mellom to bilder ved å lage et uniformt koordinatsystem [5]. Chan *et al.* [17] og Litjens *et al.* [9] har studert segmenteringsmetoder for diagnostisering av prostatakreft. Litjens *et al.* [9] beskriver en metode som først klassifiserer pasienten i forhold til stadium av kreft, og deretter klassifiserer vokslene. I tillegg segmenteres prostata før klassifisering av kreftsvulsten.

Ujevnheter i magnetfeltet, anisotropiske MRI sekvenser, ulike blære og rektal fyllinger, pasient og organ bevegelse og svulstforandringer under behandlingen kan redusere nøyaktigheten til den automatiske inntegning og registeringen av livmorhalsen og kreftsvulstene [7]. Adaptiv re-planlegging kan brukes for å kompensere for bevegelse av organene mellom behandlingene, og bestrålingen av frisk vev kan reduseres ved å innføre automatisk segmentering gjennom behandlingen [7]. Lu *et al*. [6] beskriver en registreringsmetode for svulstdeteksjon som skanner pasienten flere ganger i løpet av behandlingsforløpet [6], og gir et sannsynlighetskart av sannsynligheten for at hver voksel tilhører svulsten [6]. Utfordringen ligger i selve avgrensingen av svulsten, og dette er det svakeste leddet i søket etter nøyaktighet innen radioterapi [7].

Turid Torheim [18] har utviklet et diagnostisk verktøy for automatisk inntegning av livmorhalskreftsvulster. Dette programmet omformer MRI bildene til en datamatrise, hvor vokselintensiteter samt romlig informasjon om vokslenes naboskap tas med. Egenskapene intensitet og struktur blir benyttet for å klassifisere vokslene som svulst eller ikke-svulst. To klassifiseringsmetoder LDA (*Lineær Diskriminant Analyse*) og QDA (*Kvadratisk diskriminant analyse*) er hittil implementert i programmet. Det diagnostiske verktøyet gir et sannsynlighetskart basert på klassifiseringsmetoden, der det vises hvor sannsynlig det er at en voksel er svulst. Videre lages en binær maske hvor alle voksler med 50% eller mer sannsynlighet blir kategorisert som svulst.

Hensikten med denne oppgaven er å videreutvikle dette diagnostiske verktøyet. Formålet er å forbedre klassifiseringens nøyaktighet, slik at det blir bedre samsvar mellom den automatiske inntegningen og radiologens inntegning. En ROI (*Region of Interest*) postprosesseringsmetode foreslås og sammenlignes med resultatene funnet via den lineære klassifikatoren. I tillegg testes ulike ikke-lineære klassifiseringsmetoder for vokselklassifisering og forskjellige preprosesseringsmetoder av MRI bildene, for å undersøke om dette gir en bedre klassifisering.

2 Teori

2.1 Livmorhalskreft

I 2013 ble 282 kvinner i Norge diagnostisert med livmorhalskreft, ved en gjennomsnittsalder på 52 år [19]. Nesten alle tilfeller av livmorhalskreft skyldes HPV viruset (*humant papillomavirus)* [1, 4]. Det finnes over 100 typer av HPV viruset, og bare et fåtall av disse gir økt risiko for livmorhalskreft [20]. Viruset smitter ved seksuell kontakt [21, 22], og en av ti kvinner på verdensbasis er til enhver tid smittet av HPV viruset [23]. I de fleste tilfeller er HPV-infeksjonen harmløs og temporær, og gir ingen symptomer [19]. Vedvarende infeksjoner og forstadier til kreft utvikles hos rundt 10% av tilfellene [21]. Bare noen av disse utvikler livmorhalskreft, hvor den største risikoen ligger hos kvinner i 35-55 års alderen [21]. Livmorhalskreft utvikler seg langsomt fra preinvasiv til invasiv kreft og det er derfor gode muligheter for å oppdage sykdommen tidlig og vellykket gripe inn [24]. Prognose for overlevelse er sterkt relatert til stadium av kreft på behandlingstidspunktet [19].

FIGO (*The International Federation of Gynecology and Obstetrics)* deler inn livmorhalskreft i ulike stadier basert på kliniske funn [19]. En forenklet oversikt over de overordnede FIGO-stadiene er vist i Tabell 2-1.

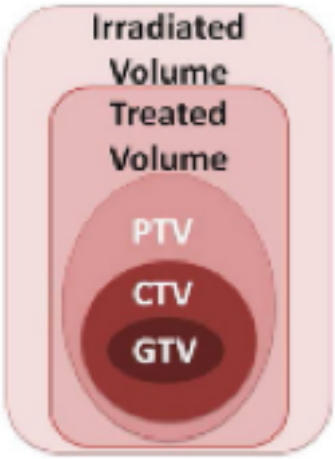
FIGO –stadium	Kliniske funn
I	Svulst begrenset til livmorhalsen
II	Svulst strekker seg til området utenfor livmorhalsen, men ikke i bekkenveggen eller ytre tredjedel av skjeden
III	Svulstutbredelse til bekkenveggen og/eller ytre tredjedel av skjeden og/eller forårsaker redusert nyrefunksjon
IV	Svulstutbredelse utenfor det lille bekkenet eller svulstinfiltrasjon av endetarm og blære eller fjernmetastaser

Tabell 2-1: Oversikt over hovedstadiene i livmorhalskreft i henhold til FIGO-inndelingen (The International Federation of Gynecology and Obstetrics) [19].

Alder, stadium av kreft og krefttype er faktorer som påvirker valg av behandling, hvor det er stadium av kreft som er den avgjørende faktoren [5]. Strålebehandling blir anbefalt for en stor andel av pasientene som utvikler livmorhalskreft, i størst grad til pasienter med et avansert stadium (*FIGO stage II-IV)* og til pasienter med tidligere stadier som ikke kan opereres [5, 12].

2.2 Avgrensing av kreftsvulster

Definisjonene fra *International Commission on Radiation Units & Measurements (ICRU)* brukes i nåværende praksis for å avgrense kreftsvulster innen strålingsterapi [25]. En svulst deles inn i tre deler; *Gross Tumor Volume* (GTV), *Clinical Target Volume* (CTV) og *Planning Target Volume* (PTV), illustrert i Figur 2-1. GTV er den delen av svulsten som bestemmes ved kliniske undersøkelser og som er synlig på CT, MR og PET bilder. Dette betyr at avgrensingen av GTV er avhengig av avbildningsmetoden og datainnsamlingsprosessen [5, 7]. CTV inkluderer GTV i tillegg til subkliniske og mikroskopiske anatomiske spredningsmønstre [7]. Disse mønstrene er vanskelig å fange opp på dagens bilder og tegnes inn ved å legge til en margin utenfor GTV [7]. Marginene til CTV baseres på antagelser og erfaringer, og har derfor stor usikkerhet [7].



Figur 2-1: Marginer brukt for å bestemme doseplanen [5]. Gross Tumor Volume (GTV), Clinically Treated Volume (CTV) og Planning Target Volume (PTV). Det behandlede volumet kan være tilnærmet lik eller større enn PTV, og er en overflate angitt av radiologien med ekvivalent dose for å oppnå formålet med behandlingen [25]. Det bestrålte volumet er vevet som mottar en signifikant strålingsdose relativt til normal strålingstoleranse [5]. Hentet fra Ghose [5], med tillatelse. Copyright © 2016 Copyright Clearence Center, Inc. All rights reserved.

Det er ofte nødvendig med manuell inntegning av livmorhalskreftsvulsten for å kunne utføre strålebehandlingen. Manuell avgrensing av svulsten er ofte tidkrevende og kan påvirkes av variasjoner og usikkerheter knyttet til inntegnerens spesialitet, trening og personlig variasjon [5, 7]. For å ta høyde for disse usikkerhetene innfører man en passende sikkerhetsmargin kalt *Planning Target Volume (PTV)* som er basert på kliniske eksperimenter [7]. Som regel inkluderer PTV en stor andel friskt vev som utsettes for en stor strålingsdose [7].

2.3 Magnetisk resonansavbildning

Magnetisk resonansavbildning også kalt MR-avbildning eller MRI (*Magnetic Resonance Imaging)*, er en avbildningsteknikk som brukes for å detektere svulster inne i kroppen. MR-avbildning anvender de magnetiske egenskapene til hydrogen for å skille sykt og friskt vev. Hydrogen forekommer hyppig i kroppen gjennom vann og lipider (fett) [26, 27]. Ved sykdom kan andelen og fordelingen av vann i vevet endre seg betydelig, og dette utnytter man under MR-avbildning [26]. På slutten av 1960 tallet ble det vist at svulstvev og friskt vev har forskjellige NMR (*Nuclear Magnetic Resonance*) spekter og at dette kunne brukes til avbildning av svulster [28]. MR-avbildning er standard prosedyre ved diagnostisering av livmorhalskreft på grunn av den gode bildekontrasten mellom normalt vev og svulst [6, 19].

2.3.1 Grunnleggende prinsipper innen MRI

En atomkjerne består av protoner og nøytroner, hvor begge disse har en individuell egenskap kalt spinn [27, 29]. For kjerner med like mange protoner og nøytroner vil summen av spinn bli lik null, fordi halvparten har spinn rettet den ene veien og den andre halvparten har spinn rettet den andre veien, ifølge Paulis eksklusjonsprinsipp [27, 30]. I noen atomer kan det være flere eller færre nøytroner enn protoner i kjernen, av et grunnstoff. Disse kalles for isotoper av grunnstoffet [27]. Atomkjerner som har en netto spinn forskjellig fra null kalles MR-aktive kjerner [27]. Prinsippet bak MR-avbildning er å bruke isotoper med netto spinn ulik null som finnes naturlig i biologisk vev [27]. I MR-avbildning bruker man som oftest hydrogenisotopen proton (^{1}H) [29]. Grunnen til dette er at omtrent 63 % av menneskekroppen består av hydrogenatomer [31], i tillegg til at hydrogen har den kraftigste responsen på eksterne magnetfelt hittil funnet i naturen [32] Hydrogenet har spinn S=½, og dermed to energinivåer (2S+1), hvor antallet energinivåer er lik [33]:

$$\#\ energinivå = 2S + 1 \tag{2.1}$$

Alle nuklider med netto spinn vil også ha et netto magnetisk dipolmoment **μ**, og det er nettopp dette som anvendes i MR-avbildningen [27]. Det magnetiske dipolmomentet for hver kjerne har vektoregenskaper, med størrelse og retning [27]. Når disse protonene ikke er utsatt for ytre magnetiske krefter vil deres magnetiske dipolmoment være tilfeldig orientert [29]. Ved påvirkning av et eksternt magnetisk felt $\mathbf{B_0}$, vil protonenes magnetiske dipolmoment rette seg etter det eksterne magnetiske feltet, enten parallelt eller antiparallelt [27, 29].

Kjerner i det lavere energinivået med spinn parallelt til det eksterne magnetfeltet kalles spinn opp kjerner, og kjerner i det høyere energinivået med magnetiske dipolmoment i antiparallell retning kalles spinn ned kjerner [27]. Energidifferansen (ΔE) mellom de to tilstandene, se Figur 2-2, er gitt ved:

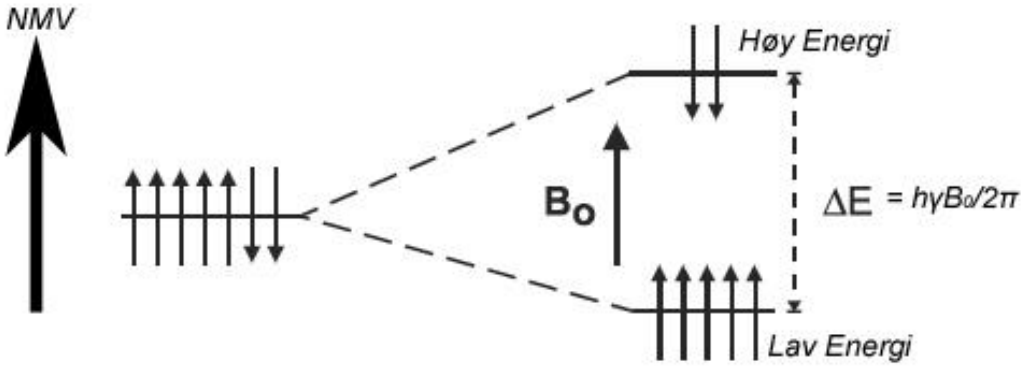
$$\Delta E = \frac{h\gamma \boldsymbol{B_0}}{2\pi} \tag{2.2}$$

hvor *h* er Plancks konstant($6.626 \times 10^{-34} Js$), γ er det gyromagnetiske forholdet til nukliden og $\mathbf{B_0}$ er det eksterne magnetfeltet [32].

Det vil alltid være flere kjerner i det lavere enn i det høyere energinivået, som tilsier at det alltid vil være flere parallelle enn antiparallelle spinnkjerner. Dette gir opphav til et netto magnetisk dipolmoment [27]. Dette magnetiske dipolmomentet skaper en magnetiseringsvektor **NMV** (*Net Magnetization Vector*) som vist i Figur 2-2 [27]. **NMV** er kilden til selve signalet som benyttes for å danne et MR-bilde [32]. Netto magnetiseringen er summen av alle dipolmomentene $\boldsymbol{\mu}$ i vevet [32],

$$\boldsymbol{NMV} = \sum \boldsymbol{\mu_i} \tag{2.3}$$

hvor $\boldsymbol{\mu_i}$ er det magnetiske dipolmomentet til proton nummer *i.*



Figur 2-2: Viser de to mulige energitilstandene til protonet under påvirkning av et eksternt magnetisk felt $\boldsymbol{B_0}$*. Spinnene til kjerner i de lavere og høyere energinivåene justerer seg henholdsvis parallelt og antiparallelt med det eksterne magnetiske feltet* $\boldsymbol{B_0}$*. ΔE er energiforskjellen mellom de to tilstandene. Flere kjerner i det lavere enn i det høyere energinivået skaper et netto magnetisk dipolmoment* $\boldsymbol{\mu}$*, som gir oppgav til en magnetiseringsvektor* ***NMV****. Laget etter figur av Brown [32]. Med tillatelse, Copyright ©2003 by John Wiley and Sons, Inc. All rights reserved.*

Fordelingen av protoner i de to energinivåene er gitt ved Boltzmann fordelingen [31]:

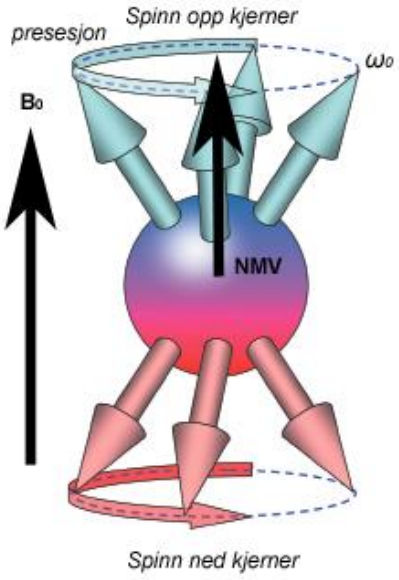
$$\frac{N^-}{N^+} = e^{\frac{-\Delta E}{kT}} \tag{2.4}$$

hvor N^- og N^+ er protoner med henholdsvis parallell (lav energi) og antiparallell (høy energi) spinn, ΔE er energidifferensen mellom tilstandene, T er vevstemperaturen i Kelvin og k er boltzmanns konstant ($1{,}3805 \times 10^{-23} J/K$).

Det eksterne magnetfeltet $\mathbf{B_0}$ gir opphav til en bevegelse kalt presesjon, som vist i Figur 2-3 [27]. Presesjon betyr at det magnetiske dipolmomentet følger en sirkulær bane rundt $\mathbf{B_0}$ med en presisjonsfrekvens ω_0 [27]. Denne frekvensen kan beregnes med Larmorlikningen [27]:

$$\omega_0 = B_0 \gamma \tag{2.5}$$

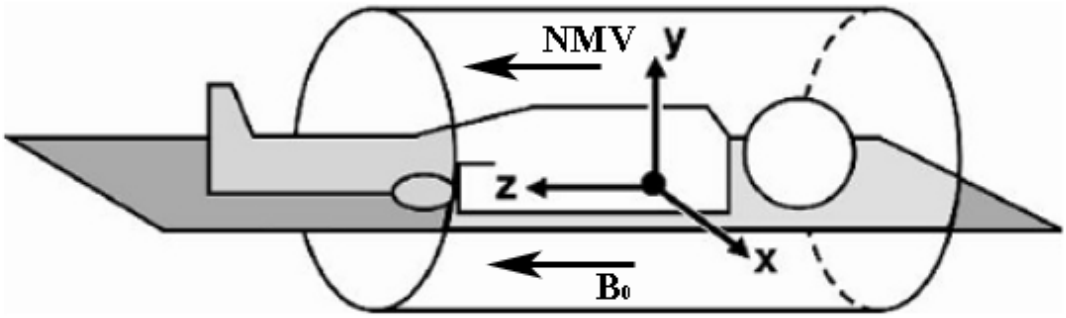
hvor B_0 er styrken på det eksterne magnetfeltet og γ er det gyromagnetiske forholdet til kjernen.



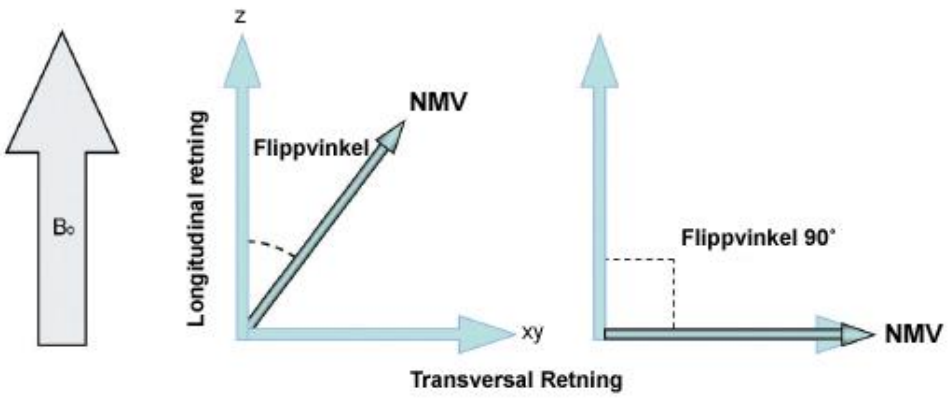
*Figur 2-3: Protonenes magnetiske dipolmoment vil rette seg etter det eksterne magnetfeltet $\mathbf{B_0}$ og presesere om $\mathbf{B_0}$ med presesjonsfrekvens ω_0 [27]. Protonenes magnetiske dipolmoment kan enten rette seg med eller imot det eksterne feltet $\mathbf{B_0}$ [27]. Det vil være et lite overskudd av protoner i det lavere energinivået rettet med det eksterne magnetfeltet $\mathbf{B_0}$, som vil skape en netto magnetiseringsvektor **NMV** [27]. Laget etter figur av Westbrook [27]. Med tillatelse, Copyright ©2011 by John Wiley and Sons, Inc. All rights reserved.*

2.3.2 Resonans

Nå som man vet hvordan protonene i kroppen reagerer på et eksternt magnetfelt kan man se nærmere på hva som skjer under selve MR undersøkelsen. Pasienten bli plassert i en spole som genererer et eksternt magnetisk felt. Denne spolen kan registrere endringer i protonenes magnetiske dipolmoment [33]. Et 3-dimensjonalt kartesisk koordinatsystem innføres, der z-aksen defineres parallelt med det eksterne magnetfeltet $\mathbf{B_0}$ og magnetiseringsvektoren **NMV**, se Figur 2-4. Fordi **NMV** signalet i z-retning er svært liten sammenlignet med det eksterne feltet $\mathbf{B_0}$, er det umulig å måle magnetiseringsvektoren når den er i likevekt og befinner seg parallelt med $\mathbf{B_0}$ [28]. For å kunne måle **NMV** er det nødvendig å flytte magnetiseringsvektoren bort fra z-aksen [28]. Dette gjøres ved hjelp av en radiofrekvent puls [28]. Når protonene i kroppen blir utsatt for en radiofrekvent puls med dens egen naturlige frekvens (Larmorfrekvensen) vil det oppstå resonans, dette kalles eksitasjon [27]. Det vil si; ved resonans vil protonene få tilført energi slik at antallet spinn ned hydrogenkjerner øker, altså det blir flere kjerner i det høye energinivået [27]. Siden **NMV** vektoren er et mål på forholdet mellom spinn opp kjerner og spinn ned kjerner, vil **NMV** under resonans bevege seg vekk fra $\mathbf{B_0}$. Magnetiseringsvektoren vil da bevege seg bort fra z-aksen og få komponenter i xy-planet (se Figur 2-5) [27]. Dette skaper en vinkel mellom **NMV** og $\mathbf{B_0}$ som kalles en flippvinkel [27, 32]. Hvor stor denne vinkelen blir er avhengig av styrken og varigheten til den radiofrekvente pulsen [27]. En flippvinkel på 90˚ er vanlig, og illustreres i Figur 2-5 [27].

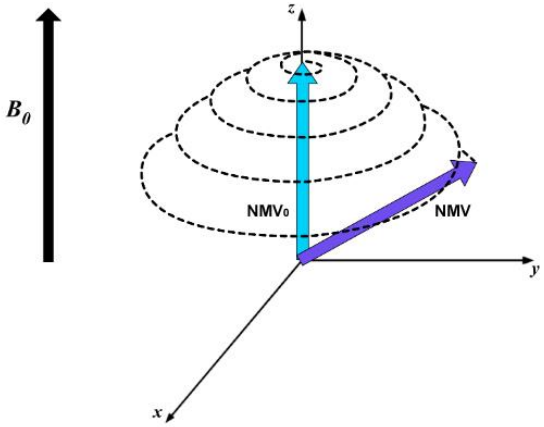


*Figur 2-4: Figuren illustrerer en person som ligger i mottakerspolen. z-aksen defineres langs kroppen, mens mottakerspolen ligger i xy-planet. Magnetiseringsvektoren **NMV** og det eksterne magnetfeltet $\boldsymbol{B_0}$ ligger langs z-aksen. Hentet fra Hashemi og Bradley [33]. Med tillatelse, Copyright © 2010, Lippincott Williams & Wilkins. All rights reserved.*



*Figur 2-5: Ved resonans vil protonene få tilført energi [27]. Dette skaper en vinkel mellom det påsatte magnetfeltet **B_0** og netto magnetisering **NMV**. Denne vinkelen kalles for flippvinkel [27]. Laget etter figur av Westbrook [27]. Med tillatelse, Copyright ©2011 by John Wiley and Sons, Inc. All rights reserved.*

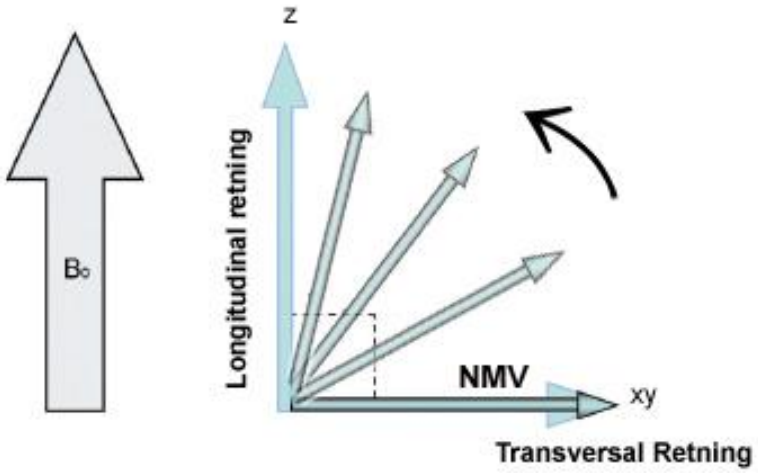
I tillegg til enkelte protoners eksitasjon vil resonans føre til at protonenes magnetiske dipolmoment legger seg i fase med hverandre og preseserer rundt **B_0** med Larmorfrekvensen ω_0, se Figur 2-6 [27]. Ved resonans vil alle magnetiske dipolmoment flytte seg til samme posisjon i presesjonveien og være i fase [27]. Presesjonen til protonene gir opphav til et tidsvarierende magnetisk felt som gjør at det induseres en strøm i mottakerspolen hvor pasienten ligger. Dette MR-signalet kan registreres og brukes for å danne MR-bildet [28].



*Figur 2-6: Figuren viser hvordan netto magnetiseringsvektoren **NMV** preseserer rundt det eksterne magnetfeltet **B_0** med Lamorfrekvensen ω_0, etter at den har blitt utsatt for en radiofrekvente puls. **NMV_0** er størrelsen til den opprinnelige likevektmagnetiseringen. Laget etter figur av de Lange og Mugler III, med tillatelse [34]*

2.3.3 MR-signalet

MR-signalet produseres når elektromagnetiske bølger i fase, beveger seg på tvers av mottakerspolen [27]. Når den radiofrekvente pulsen blir slått av vil **NMV** igjen forsøke å rette seg etter det eksterne magnetfeltet $\mathbf{B_0}$, se Figur 2-7 [27]. Denne prosessen kalles relaksasjon. Noen av kjernene i det høyere energinivået returnerer til det lavere energinivået, og hydrogenkjernene gir fra seg energi [27]. Dette gjør at magnetiseringen vil øke i longitudinal retning og avta i transversal retning [27], som vist i Figur 2-7. Den transversale magnetiseringen induserer en strøm i spolen. Denne strømmen danner MR-signalet og kalles *free induction decay* (FID) [29, 33].

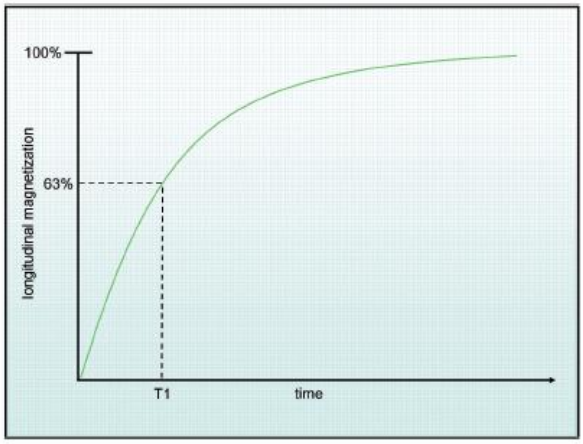


*Figur 2-7: Magnetiseringsvektoren **NMV** vil rette seg etter det eksterne feltet $\mathbf{B_0}$ etter at den radiofrekvente pulsen er slått av. Dette kalles relaksasjon. Magnetiseringsvektoren vil da øke i longitudinal retning og avta i transversal retning [27].*

Longitudinal relaksasjon (T_1 relaksasjon) oppstår ved at hydrogenkjernene avgir energi til omgivelsene og er også kalt spinn-gitter-relaksasjon [27]. Anta ar $\mathbf{B_0}$ ligger i z-retning og den radiofrekvente pulsen presser **NMV** ut i xy-planet [33], se Figur 2-7. Når den radiofrekvente pulsen opphører vil **NMV** gjenvinne magnetiseringen i z-retning eksponentielt med en tidskonstant T_1, se Figur 2-7 [27, 33]:

$$NMV_z(t) = NMV_0\left(1 - e^{-\frac{t}{T1}}\right) \tag{2.6}$$

hvor $NMV_z(t)$ er gjenvinningen av den longitudinale magnetiseringsvektoren som funksjon av tid t, NMV_0 er størrelsen til den opprinnelige likevektmagnetiseringen, t er tiden fra den radiofrekvente pulsen opphører og T_1 er tidskonstanten for den longitudinale relaksasjonskurven, og defineres som den tiden det tar før den longitudinale magnetiseringen har gjenvunnet 63 % av sin opprinnelige likevektsmagnetisering $\mathbf{NMV_0}$ i vevet, som illustrert i Figur 2-8 [27, 33].

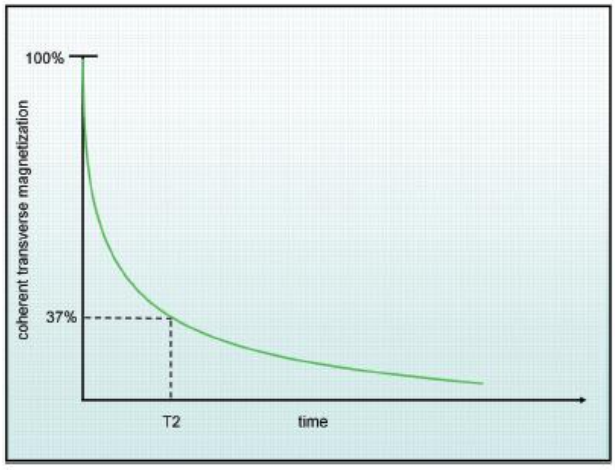


Figur 2-8: T_1 longitudinale relaksasjonskurven. Hentet fra Westbrook [27]. Med tillatelse, Copyright ©2011 by John Wiley and Sons, Inc. All rights reserved.

T_2 relaksasjonen i transversal retning oppstår ved at de lokale magnetiske feltene til naboprotonene forstyrrer hverandre og resulterer i tap av koherent transversal magnetisering [27]. Dette kalles spinn – spinn relaksasjon [27]. Tapet av den transversale magnetiseringen NMV_{xy} er gitt ved [33]:

$$NMV_{xy}(t) = NMV_0 \, e^{-\frac{t}{T2}} \tag{2.7}$$

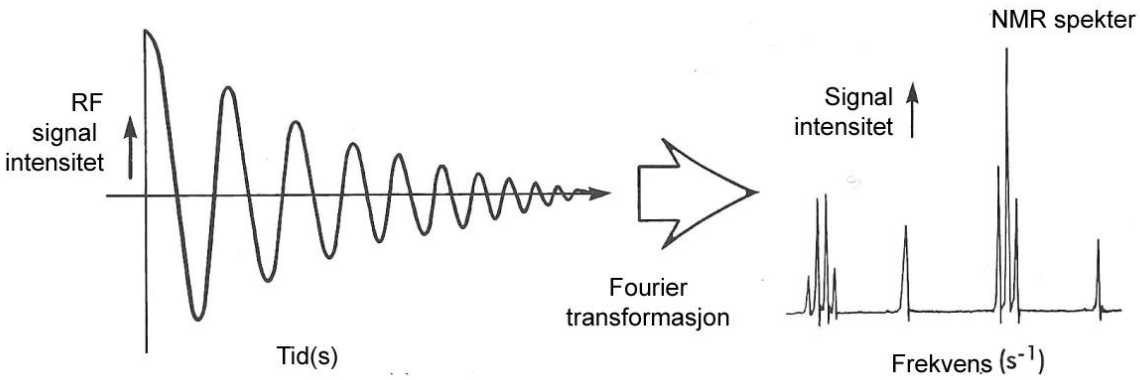
hvor NMV_0 er størrelsen til den opprinnelige likevektmagnetiseringen, t er tiden fra den radiofrekvente pulsen opphører og T_2 er tidskonstanten for den transversale relaksasjonskurven, og defineres som den tiden det tar før den transversale magnetiseringen har tapt 63 % av sin opprinnelige likevektsmagnetisering **NMV_0** i vevet, det vil si når 37% av magnetisering er igjen, som illustrert i Figur 2-9 [27, 33]



Figur 2-9: T_2 transversale relaksasjonskurven. Hentet fra Westbrook [27]. Med tillatelse, Copyright ©2011 by John Wiley and Sons, Inc. All rights reserved.

Det kan også forekomme inhomogeniteter i det ytre statiske magnetfeltet $\mathbf{B_0}$ som gjør at den transversale magnetiseringen forsvinner enda raskere [33]. Da får man en ny tidskonstant for den transversale relaksasjonskurven som kalles T_2^* [33]. Denne inkluderer både det ytre statiske feltet og spinn-spinn relaksasjon, og defineres som i ligning (2.7) hvor man benytter T_2^* istedenfor T_2 [33].

For å transformere MR-signalet til et komplett bilde bruker man FID signalet som genereres i mottakerspolen [28]. FID signalet plottes som en funksjon av tid, og gir opphav til et NMR spektrum ved hjelp av en Fourier transformasjon, se Figur 2-10 [28]. Hver av toppene i et NMR spektrum representerer en karakteristikk for de ulike vevstypene i kroppen. Dette gjør at man kan skille ulike typer vev fra hverandre [28]. Som nevnt tidligere kan dette brukes til å skille mellom frisk og sykt vev. Et bilde blir konstruert ved at hver voksel med samme intensitet representerer samme vev [28]. Utfordringen ligger i å bestemme den romlige plasseringen til vokslene i bildet [28]. Dette løses ved å benytte et magnetisk gradientfelt, som er et magnetfelt som varierer i x-,y- og z-retning [33]. Da skapes det en systematisk romlig variasjon i resonansfrekvensen, slik at man kan finne hvor signalet kommer fra og plassere bildepikselen på rett sted [28]. Ved en kombinasjon av flere radiofrekvente pulser kan man måle styrken til NMR spektrumet og i tillegg plassere vokslene slik at det blir et visuelt bilde av de forskjellige vevstypene. Denne kombinasjonen av radiofrekvente pulser kalles for pulssekvenser [28]. Det finnes flere forskjellige typer pulssekvenser, men disse beskrives ikke i denne oppgaven, se Bushong S.C [28] for detaljert beskrivelse.

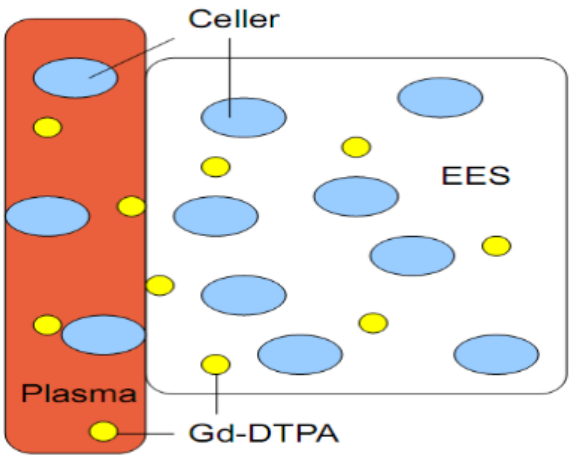


Figur 2-10: Det radiofrekvente signalets intensitet plottes som funksjon av tiden (s) og gir FID (free induction decay) signalet. Ved en Fourier transformasjon kan man få frem NMR spekteret for bildesnittet som fremstår ved å plotte signalets intensitet mot frekvensen (s^{-1}) [28]. Hentet Fra Bushong [28], med tillatelse. Copyright © Elsevier 1996, by Mosby-Year Book, Inc.

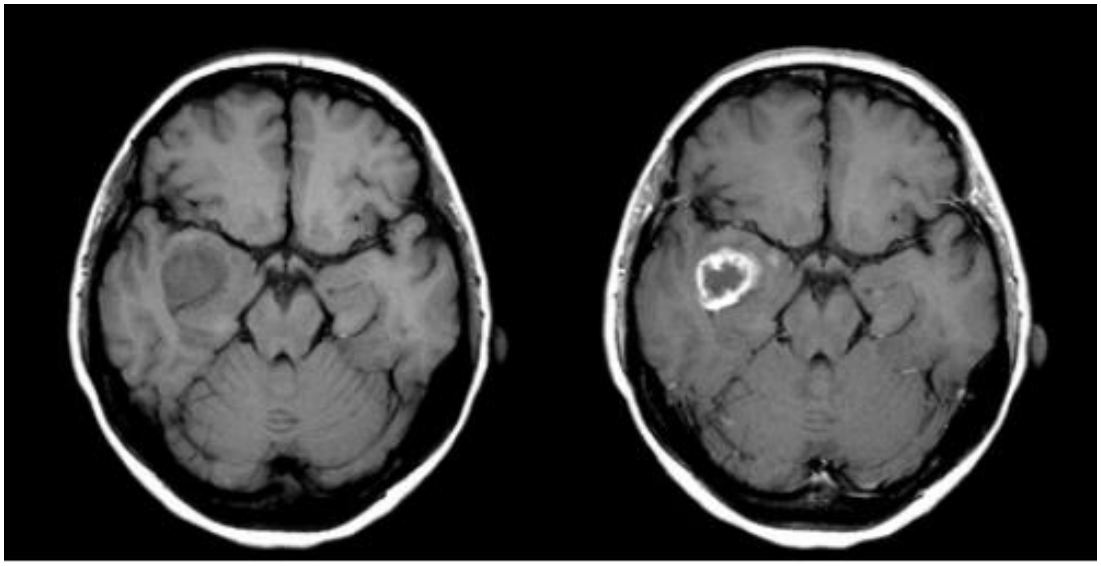
2.3.4 Dynamisk kontrastforsterket MRI

Dynamisk kontrastforsterket MRI (DCE, *Dynamic Contrast Enhanced)* utføres ved injeksjon av et kontrastmiddel i vevet, for deretter å måle de påfølgende intensitetsendringene i de T_1-vektede eller T_2-vektede bildene [35-37]. Kontrastmidlene som brukes inneholder paramagnetiske stoffer som har uparede elektroner i det ytterste skallet, og skaper dermed et kraftig lokalt magnetisk felt [36]. Det mest brukte kontrastmidlene er basert på gadolinium (Gd^{3+}), som har så mange som 7 uparede elektroner i det ytterste skallet og er derfor sterkt paramagnetisk [36]. Funksjonen til kontrastmiddelet er at det senker T_1 relaksasjonskonstanten og øker derfor signalintensiteten til det T_1 vektede bildet [36, 38]. Kontrasten mellom de forskjellige vevstypene vil da øke fordi forskjellen mellom T_1 relaksasjonstidene øker [36].

Ved DCE-MRI avbildes vevet i flere tidsintervall for å finne ut hvordan kontrastmiddelet beveger seg i det område man undersøker [38]. Kontrastmiddelet føres inn i kroppen intravenøst og fraktes med blodet til det organet eller vevet som skal avbildes [36, 39]. Det mest vanlige kontrastmiddelet er Gd-DTPA (*Gadolinium diethylene-triamine pentaacetic acid)*, og er et paramagnetisk kontrastmiddel med lav molekylær vekt [38, 39]. Gjennom mikroskopiske porer i kapillærveggene flyter kontrastmiddelet fra blodet og til et område kalt EES (*Extracellular-Extravascular Space*). EES er den delen av vevet som ikke er celler, se Figur 2-11 [38, 39]. Etter en viss tid skilles kontrastmiddelet ut gjennom nyrene [26]. Kontrastøkningen er blant annet avhengig av kontrastmiddelet, tettheten av kapillærblodårer, blodgjennomstrømning, permeabiliteten til åreveggene, og sammensetningen av EES i det spesifikke vevet [37, 38]. Fordi sykt vev har et unormal karnettverk og en høy grad av permeabilitet så tenderer kreftsvulster til å ha en høyere signalintensitet enn normalt vev. Dette fører til et kontrastforsterket bilde, se Figur 2-12 [38].

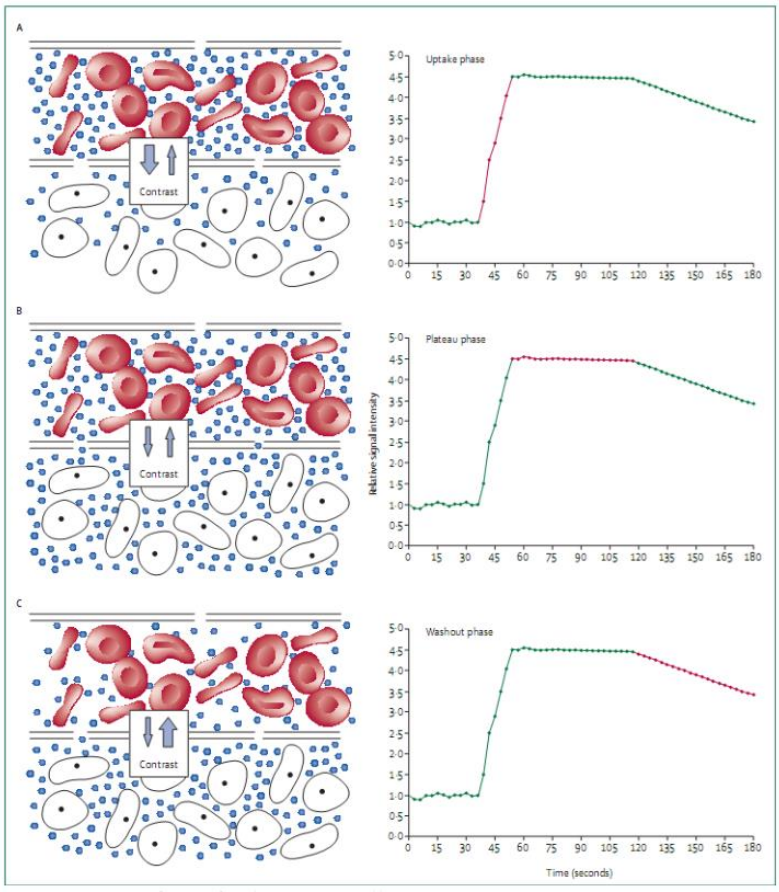


Figur 2-11: Kontrastmiddelet Gd-DTPA (gult) befinner seg i blodplasma (rødt) og i det ekstracellulære ekstravaskulære rommet EES (hvitt), men beveger seg ikke inn i cellene (blå). Blodplasma er den delen av blodet som ikke er celler. EES er den delen av vevet som ikke er celler. Bildet hentet fra Torheim[40], med tillatelse.



Figur 2-12: To bilder fra en DCE-MRI bildeundersøkelse av en hjernetumor. Bilde til venstre viser et T_1-vektet bilde før injeksjon av kontrastmiddelet, mens figuren til høyre viser effekten av et gadoliniumbasert kontrastmiddel. Kontrastmiddelet samles i tumoren, som gir en økt signalintensitet i det T_1-vektede DCE-MRI bildet [36]. Bildet er hentet fra Bjørnerud [36], med tillatelse.

Før injeksjonen av kontrastvæsken tas et før-kontrastbilde, slik at den opprinnelige T_1-verdien/T_2-verdien kan måles [38]. Etter injeksjonen av kontrastvæsken skjer en dynamisk avbildning, med typisk noen sekunders intervaller [38]. DCE-MRI avbildningen kan dermed deles inn i tre faser, se Figur 2-13 [38]. Den første fasen er opptaksfasen hvor det vil være en økt signalintensitet i område som avbildes, grunnet en netto transport av kontrastmiddel fra blodårene til EES (A Figur 2-13) [38]. Den andre fasen kalles platåfasen (B Figur 2-13) [38]. I denne fasen er det likevekt i bevegelsen av kontrastmiddel mellom blodplasma og EES, og kontrastforsterkningen er maksimal [38]. I den avsluttende utvaskingsfasen vil signalintensiteten avta fordi det er netto transport av kontrastvæske fra EES til blodet (C Figur 2-13) [38].

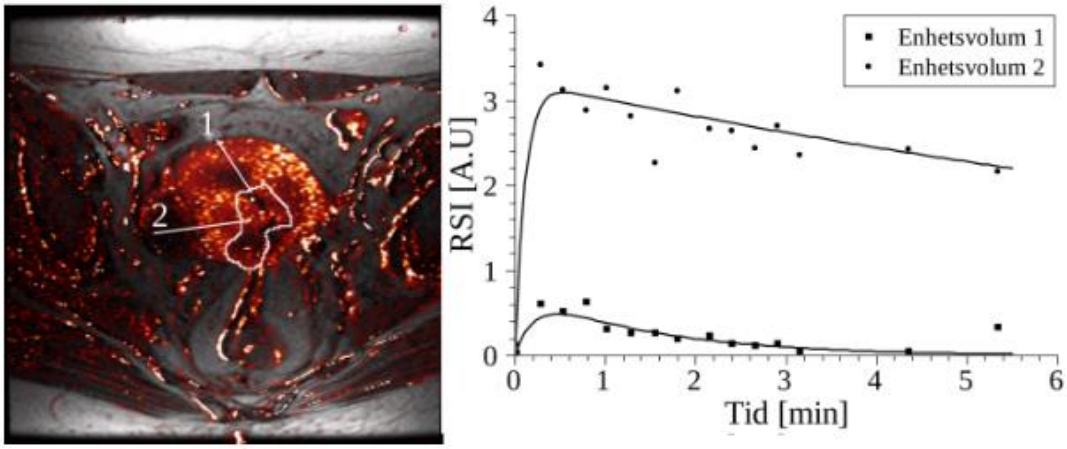


Figur 2-13: Kontrastmiddelets tre faser i vevet. A) Opptaksfasen hvor det er netto transport av kontrastmiddel fra blodårene til EES. B) Platåfasen hvor det er likevekt i transport av kontrastmiddelet. C) Utvaskingsfasen hvor det er netto transport av kontrastmiddel fra EES til blodårene. Hentet fra Zahra et al. [38], med tillatelse. Copyright © 2016 Copyright Clearence Center, Inc. All rights reserved.

Ved å sammenligne kontrastbildene med før-kontrastbildet kan man se hvordan signalintensiteten endrer seg med tid [38]. Ut ifra tid-signal intensitets plottene (Figur 2-14) kan man beregne signaløkningen i etter-kontrastbildene [38]. Hver voksel i etter-kontrastbildene tatt ved tiden *t* har en relativ signaløkning *RSI (Relative Signal Increase)* definert som [41]:

$$RSI(t) = \frac{S(t) - S(0)}{S(0)} \tag{2.8}$$

hvor *S(t)* er signalintensiteten i etter-kontrastbildet ved tiden *t* og *S(0)* er signalintensiteten i før-kontrastbildet tatt ved tiden $t = 0$.



Figur 2-14: Grafen til høyre viser datapunkter (RSI) for to enhentsvolumer i svulsten (venstre), hvor økningen i opptakskurvene indikerer opptak av kontrastvæsken, mens der kurvene avtar indikerer utvaskingen av kontrastmiddelet. Hentet fra Andersen, E. [42], med tillatelse .

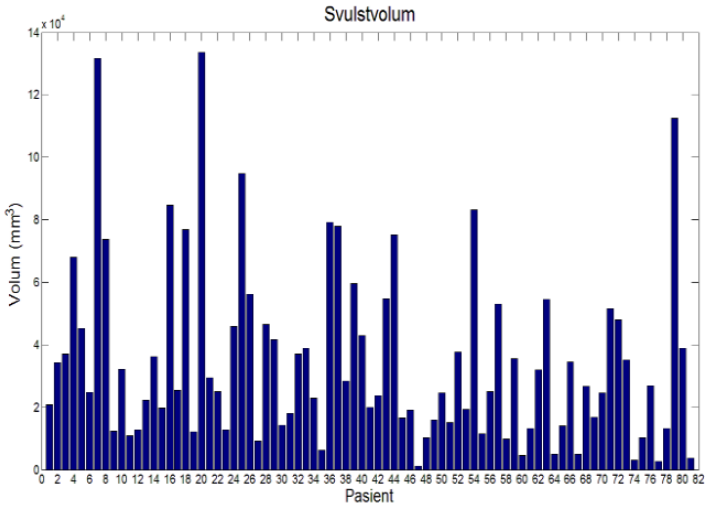
3 Materialer

3.1 Livmorhalskreft Datasettet

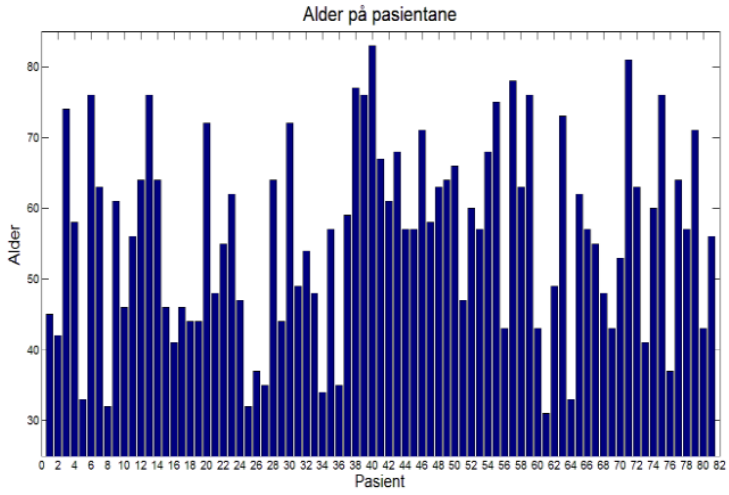
Oppgaven er basert på et datasett bestående av MR-bilder (T_1-vektede, T_2-vektede og DCE-bildeserie) fra 81 pasienter med livmorhalskreft. Avbildningen ble utført på en 1,5 Tesla MR-maskin fra GE Medical Systems [42]. Pasientene ble undersøkt i perioden 2001 til 2004 ved det Norske Radiumhospitalet (nå en del av Oslo Universitetssykehus) [43]. Tre av disse 81 pasientene ble ikke inkludert i datasettet fordi alle bildetypene (T_1-vektede, T_2-vektede og DCE-MRI) ikke ble samlet inn for disse pasientene. Pasientene mottok behandling i form av strålingsterapi og kjemoterapi etter at bildene ble tatt. Gjennomsnittsalderen for pasientgruppen er 56 år (32-85 år), hvor forskjellige stadier av kreft er presentert i Tabell 3-1. En oversikt over svulstvolum og alder for hver pasient er vist i henholdsvis Figur 3-1 og Figur 3-2.

Stadium	Antall pasienter	Andel pasienter
I	2	2,5%
II	44	54,3%
III	29	35,8%
IV	5	7,4%

Tabell 3-1: Fordeling av stadier (I- IV) til de 81 pasientene før behandling .



Figur 3-1: Svulstvolumet (mm^3) for de 81 pasientene. Gjennomsnittsvolumet er 34 913 mm^3. Hentet fra Torheim [40], med tillatelse.



Figur 3-2: Alderen til hver pasient ved undersøkelsestidspunktet. Det er totalt 81 pasienter, hvor gjennomsnittsalderen er 56 år (32-85 år). Hentet fra Torheim [40], med tillatelse.

T_1-serien og T_2-serien ble utført med en Fast-Spin-Echo sekvens, mens den dynamiske serien ble utført med en Fast-Spoiled-Gradient-Recalled sekvens med en 90° flippvinkel [42]. De T_2-vektede og T_1-vektede bildene hadde 20 eller 32 snitt per pasient, med 512x512 voksler for hvert bildesnitt.

DCE-MRI bildene hadde ni eller ti snitt per pasient, med 256x256 voksler per snitt. Kontrastmiddelet Gd-DTPA ble benyttet [42], og DCE-bildeseriene besto av et før-kontrastbilde og mellom 12 til 15 etter-kontrastbilder per pasient. De første elleve etter-kontrastbildene ble tatt med 15 sekunders intervaller, mens resten av serien ble tatt med ett minutts intervaller. De første 13 tidstrinnene (før-kontrast og de tolv første etter-kontrastbildene) ble brukt i analysene i denne oppgaven, slik at det ble samme antall tidstrinn for alle pasientene.

Den manuelle inntegningen av svulstene ble gjort av en erfaren radiolog på de T_2-vektede bildene, og godkjent av en annen radiolog. Dette ble antatt å være den grunnleggende sannheten, og brukt til å trene modellene og teste nøyaktigheten til den automatiske inntegningen.

Bildene ble avskåret slik at irrelevante og overflødige bildevoksler av for eksempel luft rundt pasienten ble fjernet, se Figur 4-2 a. T_2-vektede og T_1-vektede bilder ble dermed redusert til 343x343 voksler per bildesnitt, mens DCE bildesnittene ble redusert til 171x171 voksler per bildesnitt. Dette førte til at bildene fikk en høyere andel svulst, som igjen førte til en bedre balanse mellom de to vokselklassene (svulst og ikke-svulst). Kun bilder hvor radiologen identifiserte svulst ble brukt i analysene, igjen for å bedre vokselforholdet mellom pasientene. Ved analyser av de tre bildetypene sammen, ble T_2-vektede og T_1- vektede bilder nedskalert for å passe oppløsningen til DCE bildene.

3.2 Programvare og datamaskin

Datamaskinen DELL Precision 3620 MT Workstation med sjettegenerasjon Intel® Core™ i7-6700 prosessor (fire kjerner, 3,40 GHz, 4,0 GHz Turbo, 8 MB, med HD –grafikk 530) og operativsystem Windows 10 Pro, 64-biters, 32 GB (2 x 16 GB), 2133 MHz DDR4, ikke-ECC minne, med en 512 GB, M.2 PCIe NVme SSD-stasjon harddisk i tillegg til en ekstra 1 TB, 2,5" SATA-harddisk (7200 o/min) ble brukt. MATLAB® versjon R2015b med tilleggspakken *Statistics and Machine Learning Toolbox* versjon 10.0 ble benyttet til klassifiseringene. MATLAB® versjon R2015b med tilleggspakken *Signal Processing Toolbox* versjon 7.0 ble benyttet ved preprosessering av MR-bildene, sammen med tilleggspakken *Image Processing Toolbox* versjon 9.2.

4 Oppbygging av svulstinntegningsprogrammet

Svulstinntegningsprogrammet er utviklet av Turid Torheim [44] og bygd opp som vist i Figur 4-1. Formålet var å lage et program som på mest nøyaktig vis klarte å skille to klasser av voksler inn i svulst og ikke-svulst. Svulstinntegninger gjort av erfarne radiologer på T_2-vektede bilder ble brukt som fasit for å kunne vurdere hvor godt programmet klarer å klassifisere vokslene.

Input til programmet var MR-bildene til hver av pasientene med livmorhalskreft, hvor prosessen vist i boksene i Figur 4-1 anvendes på alle bildesnitt og tidstrinn for alle pasientene. Først preprosesseseres bildene for å kompensere for intensitetsforskjeller mellom pasientene. Deretter brettes bildene ut til en datamatrise i en utfoldingsprosess, hvor hver rad inneholder intensitetsverdier til en bestemt voksel for hvert bildet. Vokslene klassifiseres deretter som svulst/ikke-svulst ved å bruke datamatrisen og en klassifiseringsalgoritme. Til slutt postprosesseres bildene for å fjerne ubetydelig støy. Et sannsynlighetskart ble generert på bakgrunn av klassifiseringen. Output bildet er et binært bilde med en sannsynlighetsgrense på 50% mellom klassene svulst og ikke-svulst. Dette binære bildet sammenlignes med radiologens maske for å undersøke hvor god klassifiseringen er. Hvert ledd i svulstinntegningsprogrammet beskrives nøye i kapitlene under.

Figur 4-1: Oppbyggingen av svulstinntegningsprogrammet. Input til programmet er MR-bilder av livmorhalskreft. Bildene preprosesseres for å kompensere for intensitetsforskjeller mellom pasientene. Under utfolding brettes bildene ut i en datamatrise som inneholder intensitetsverdier til en bestemt voksel i bildet. Klassifiseringen skjer ved at hver voksel klassifiseres som svulst/ikke-svulst ved bruk av datamatrisen og en klassifiseringsalgoritme. Bildene postprosesseres for å fjerne ubetydelig støy. Resultater og validering blir brukt til å vurdere programmets ytelse ved å sammenligne dataprogrammets bilder (output bildet) med radiologens inntegninger.

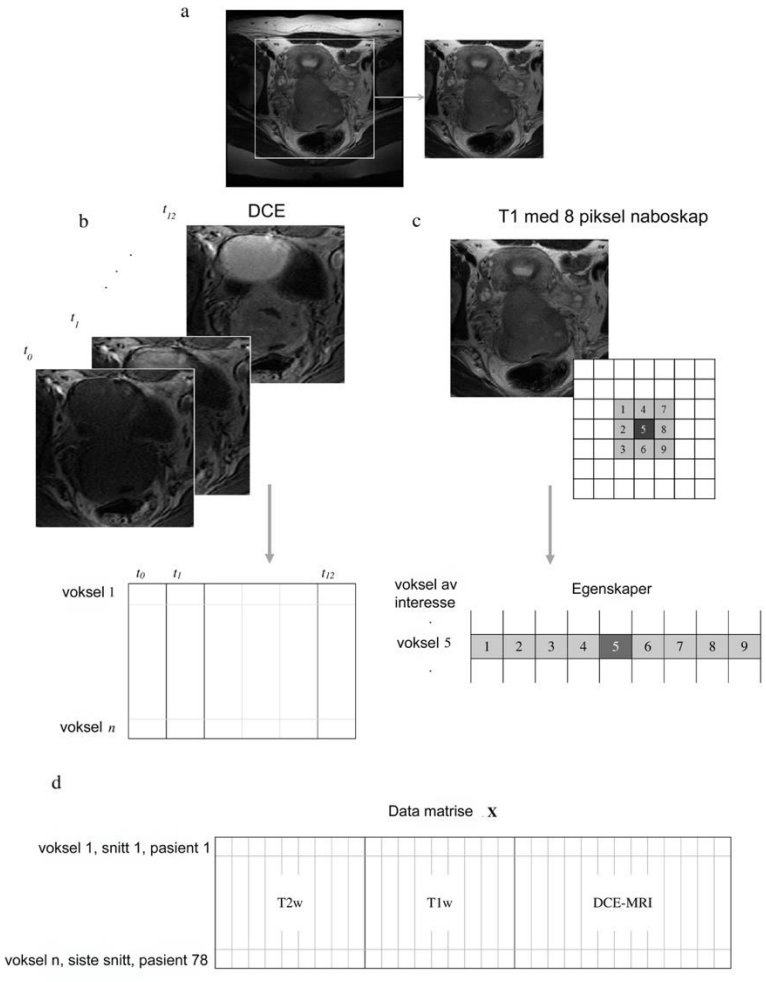
4.1 Preprosessering av MR-bildene

For å kompensere for intensitetsvariasjoner i MR-bildene mellom hver pasient, ble det gjort autoskalering av de T_1-vektede og T_2-vektede bildene til gjennomsnitt null og standardavvik en [18]. For DCE-bildene ble hele tidsserien autoskalert som én, for å beholde intensitetsøkningen mellom tidstrinnene [18].

4.2 Utfolding av MR-bildene

Hvert av snittene til de T_1-vektede og T_2-vektede bildene ble brettet ut kolonnevis [45, 46]. Dette betyr at hvert bilde ble transformert til en egenskapsvektor med lengde *n*, hvor *n* er antall voksler i hvert bildesnitt [18]. Egenskapsvektoren til alle bildesnittene ble satt sammen vertikalt i to lange egenskapsvektor av henholdsvis T_2-vektede og T_1-vektede intensitetsverdier for alle voksler fra alle pasienter [18]. DCE-bildene ble brettet ut til en egenskapsvektor for hvert tidstrinn, som ble satt sammen i en matrise med antall rader lik totalt antall voksler, og antall kolonner lik totalt antall tidstrinn, se Figur 4-2 b [18].

Egenskapsvektorene til de T_2-vektede og T_1-vektede bildene, og DCE-egenskapsmatrisen ble satt sammen til en stor datamatrise, hvor datamatrisen hadde en rad for hver voksel og en kolonne for hver egenskap til vokslene, se Figur 4-2 d [18]. I denne datamatrisen kunne man velge å benytte et åtte-voksel naboskap for et eller flere av bildetypene. Dette betyr at de åtte nærmeste vokslene rundt vokselen av interesse inkluderes i datamatrisen $\boldsymbol{X}$ for å opprettholde informasjon om det romlige forholdet mellom vokslene, se Figur 4-2 c [45, 46]. T_2-vektede og T_1-vektede bilder med åtte nabovoksler, og de tretten første tidstrinnene i DCE-serien (ingen naboskap) danner datamatrisen $\boldsymbol{X}$, som for alle de 78 pasientene besto av 31 kolonner og 486 snitt. Datamatrisen $\boldsymbol{X}$ består til slutt av 41 181 885 voksler x 31 egenskaper. 27% av vokslene i datasettet var svulst voksler, mens 73% var ikke-svulst voksler.

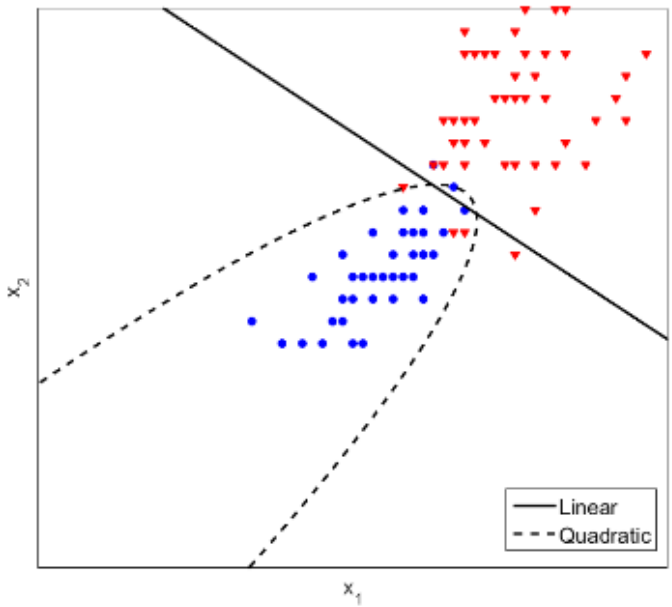


*Figur 4-2: Utfoldingsmetoden, hvor MR-bildene ble konvertert til en datamatrise **X** som ble brukt til vokselklassifisering. a) Bildet ble avskåret slik at irrelevant data som luft rundt pasienten og overflødige voksler ble fjernet. b) DCE-MRI serien for hver pasient ble utfoldet til en kolonnevektor per tidssteg. De 13 første tidsstegene (før-kontrastbildet og de 12 neste tidstrinnene) ble brukt, som resulterer i 13 kolonner i egenskapsmatrisen. c) De T_1-vektede og T_2-vektede bildene ble utfoldet enten til en kolonnevektor hver som inneholdt intensiteten til hver voksel, eller til en egenskapsmatrise hver som inkluderte nabovokslers intensitet for å ta vare på noe av informasjonen om det romlige forholdet mellom vokslene. De åtte nærmeste nabovokslene (lys grå) for hver voksel av interesse (mørk grå) ble utfoldet kolonnevis, som da gir ni kolonnevektorer. d) Datamatrisen **X** ble basert på T_2-vektede bilder med åtte nabovoksler, T_1-vektede bilder med åtte nabovoksler og de første 13 tidsstegene i DCE-MRI serien. Datamatrisen **X** besto dermed av 31 kolonner og en rad for hver voksel i de 486 snittene, som til sammen gir 14 181 885 voksler x 31 egenskaper. Hentet fra Torheim et al. [18], med tillatelse.*

4.3 Klassifisering av voksler: Lineær og kvadratisk diskriminant analyse

To klassifiseringsmetoder var foreløpig implementert for å klassifisere vokslene i MR-bildene. Den første var Fisher´s Lineær Diskriminant Analyse (LDA) [47], som forsøker å identifisere en linje som skiller to (eller flere) grupper i datasettet ved å minimalisere spredningen innad i gruppene, og maksimere spredningen mellom gruppene [47, 48], se den heltrukne linjen i Figur 4-3. LDA modellen ble trent ved å bruke radiologens inntegninger og avgrensinger [18]. Resultatet av klassifiseringen gir et sannsynlighetskart som angir sannsynligheten for at en voksel er en svulst [49]. Sannsynlighetskartene ble konvertert til binære masker ved å sette en terskel på 50% [18], som vil si at hver voksel med 50% eller høyere sannsynlighet for at den er svulst ble klassifisert som dette.

Den andre klassifiseringsmetoden som var implementert var QDA (Kvadratisk Diskriminant Analyse), som er en ikke-lineær klassifiseringsmetode som ligner på LDA [48]. Forskjellen er en kvadratisk grense mellom klassene, istedenfor en lineær, se stiplet linje i Figur 4-3.



Figur 4-3: To grupper representert med blå prikker og røde trekanter er klassifisert med en lineær modell (heltrukken linje), og en kvadratisk modell (stiplet linje) basert på de to variablene X_1 og X_2. Hentet fra Torheim et al. [18], med tillatelse.

4.4 Postprosessering av klassifiseringsresultatene

For å fjerne ubetydelig støy i klassifiseringsresultatene ble det til brukt en morfologisk operasjon på de binære bildene for å fjerne alle små elementer mindre enn 10 voksler [50].

4.5 Resultater og validering

Leave-one-patient-out kryss-validering ble brukt ved trening av klassifiseringsmodellen. For å estimere modellens ytelse ble alle bildene til en pasient tatt ut fra datamatrisen, og LDA (QDA) modellen ble trent på de resterende bilde [45, 48]. Klassifiseringsmodellen ble brukt til å tegne inn svulsten i bildene til pasienten som ble utelatt i modelltreningen. Dette ble gjentatt helt til alle pasientene hadde blitt utelatt en gang og dermed kunne ytelsen vurderes. Dette vil være lik praksis i klinikker hvor en ny pasient kommer inn og skal diagnostiseres basert på tidligere funn av svulster.

Ytelsen til klassifiseringsmodellene vurderes ved hjelp av ulike prestasjonsmålinger [51]. En forvirringsmatrise sammenligner de predikerte og faktiske klassene for klassifiseringsmodellen, og gir basisen for prestasjonsmålingene, se Tabell 4-1.

		Predikert utfall		
		Positiv	*Negativ*	*Totalt*
Målt Utfall	*Positiv*	SP	FN	P_1
	Negativ	FP	SN	Q_1
	Totalt	P_2	Q_2	100%

Tabell 4-1: Forvirringsmatrisen til en klassifikator med to klasser. SP er de sanne positive (riktig predikert som svulst), SN er de sanne negative (riktig predikert som ikke-svulst), FP er de falske positive (ikke-svulst voksler, klassifisert som svulst) og FN er de falske negative (svulst voksler, klassifisert som ikke-svulst). P_1 og Q_1 er henholdsvis den sanne andelen voksler i klassene svulst (positiv) og ikke-svulst (negativ), mens P_2 og Q_2 er henholdsvis den predikerte andelen voksler i klassene svulst (positiv) og ikke-svulst (negativ).

Dice similarity coefficient (DSC) [52, 53] ble brukt til å vurdere hvor godt masken laget av radiologen og den binære masken gitt av klassifiseringsmodellen overlapper. Den er definert som

$$DSC = \frac{2SP}{FP + FN + 2SP} \tag{4.1}$$

hvor *SP* er sanne positive (riktig predikerte svulst voksler), *SN* er sanne negative (riktig predikerte ikke-svulst voksler), *FP* er falske positive (ikke-svulst voksler, klassifisert som svulst) og *FN* er falske negative (svulst voksler, klassifisert som ikke-svulst) [52, 53]. DSC ligger mellom 0 til 1, hvor 1 er perfekt overlapp [52].

Kappa statistikk *K* ble også brukt for å sammenligne enigheten mellom den binære segmenteringen gitt av klassifiseringsmodellen og radiologens maske [18, 54]. Kappa statistikken er gitt som

$$K = \frac{2(SP \cdot SN - FP \cdot FN)}{(SP + FN)(FP + SN) + (SP + FP)(FN + SN)} \tag{4.2}$$

hvor $K = 1$ er fullstendig enighet og $K>0$ viser at enigheten er bedre enn om det skulle vært helt tilfeldig [18, 54]. $K>75$ anses som utmerket overlapp [55].

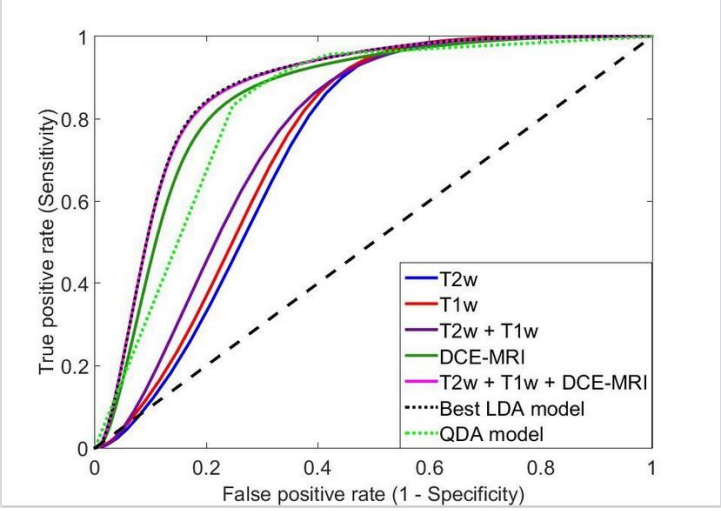
Sensitivitet og spesifisitet ble også beregnet for å vurdere hvor godt modellen klassifiserte de to vokselgruppene (svulst og ikke-svulst) [51]. Sensitiviteten er klassifikasjonsnøyaktigheten for den positive klassen (klasse svulst), og er gitt som

$$Sens = \frac{SP}{SP + FN} \tag{4.3}$$

Spesifisiteten er i motsetning til sensitiviteten, klassifikasjonsnøyaktigheten for den negative klassen (klasse ikke-svulst), og er gitt som

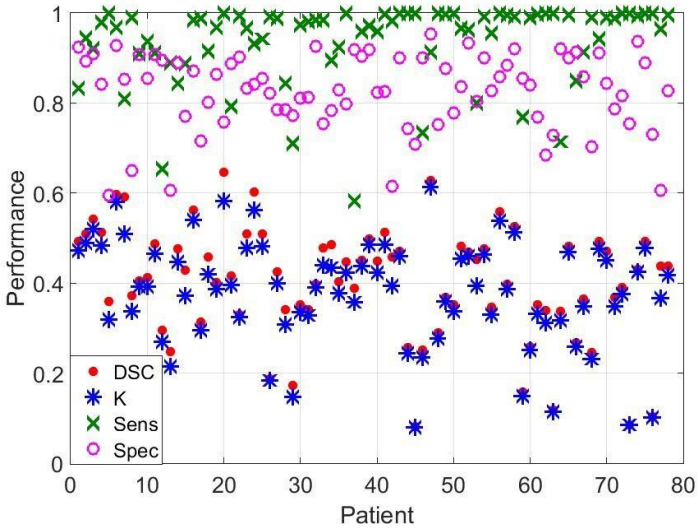
$$Spes = \frac{SN}{SN + FP} \tag{4.4}$$

En annen nyttig metode for å evaluere ytelsen til klassifiseringsmodellene er ROC-kurven (*Receiver Operating Characteristic*), vist i Figur 4-4 [56]. Denne kurven benyttes for å illustrere hvordan spesifisiteten og sensitiviteten varierer når terskelen som brukes for å skille de to klassene varierer. ROC kurven brukes til å estimere modellens ytelse ved å beregne AUC (*Areal Under Curve*). Den svarte, stiplende linjen i Figur 4-4 viser en modell som er like god som om man skulle gjettet tilfeldig hvilken gruppe vokslene tilhører, og har en AUC verdi lik 0,5. En fullstendig korrekt klassifisering vil gi en ROC kurve som har (1 – spesifisitet) lik null og sensitivitet lik en, og dermed en AUC verdi lik en.



Figur 4-4: Eksempler på ROC kurver for seks LDA modeller basert på enten T_2-vektede bilder, T_1-vektede bilder, T_2-vektede og T_1-vektede bilder, DCE-MRI bilder eller alle tre bildetypene kombinert. ROC kurvene representerer LDA modellen med den høyeste AUC verdien for hver kategori. Kurven "Best LDA model", representerer LDA modellen basert på T_1-vektede bilder med åtte naboer inkludert i tillegg til DCE-MRI serien, og autoskalering som preprosessering av bildene. "QDA model" representerer QDA modellen basert på det samme bilde datasettet som den beste LDA modellen. Hentet fra Torheim et al. [18], med tillatelse.

DSC, *K*, sensitivitet og spesifisitet ble beregnet for hvert bildesnitt og tidstrinn, og et gjennomsnitt for hver pasient ble beregnet slik at man kunne sammenligne resultater mellom pasientene. Alle disse parameterne for hver pasient er vist i Figur 4-5. Denne figuren viser hvor godt klassifiseringsmodellen fungerer for hver enkelt pasient.



Figur 4-5: Spredningsplot for den gjennomsnittlige DSC, K, sensitivitet og spesifisitet for hver pasient for den beste LDA modellen, se Vedlegg A: Vokselklassifisering med LDA, Tabell A-1 i uthevet skrift. Figuren viser at modellytelsen varierer fra pasient til pasient. Hentet fra Torheim et al. [18], med tillatelse.

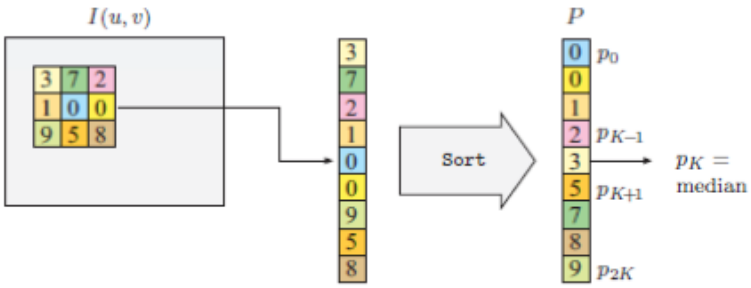
5 Metoder og videreutvikling av svulstinntegningsprogrammet

Formålet med videreutviklingen av svulstinntegningsprogrammet var å øke nøyaktigheten til det diagnostiske verktøyet. Programmet beholdt samme struktur som vist i Figur 4-1, hvor nye metoder for preprosessering, vokselklassifisering og postprosessering testes.

5.1 Preprosesseringsmetoder

5.1.1 Median filter

Median filteret er en metode for å fjerne støy i bilder [57]. Median filteret erstatter hver bildepiksel (voksel) med medianen av alle pikslene i den gitte filterregionen [57]. Medianen til sekvensen $2K+1$ med p_i verdier, er definert som den midterste verdien p_k etter at den gitte sekvensen $P = (p_o,....,p_{2k})$ er sortert i stigende rekkefølge, som vist i Figur 5-1 [57].



Figur 5-1: Beregning av et 3 x 3 piksel median filter. Ni nærliggende pikselverdier tas fra originalbildet I(u,v) og sorteres i stigenes rekkefølge inn i en sekvens P, og den midterste verdien p_k er den resulterende medianverdien. Hentet fra Burger er al. [57], med tillatelse.© Springer-Verlag London Limited 2009.

Så lenge sekvensen *P* består av et oddetall antall voksler vil det ikke bli laget noen nye pikselverdier i bildet [57]. Om *P* sekvensen derimot inneholder et partall antall voksler så defineres medianen som gjennomsnittet mellom de to midterste verdiene, og nye pikselverdier kan skapes [57].

I denne oppgaven ble MATLAB®´s median filter anvendt på alle bildesnitt for alle bildetypene og alle tidstrinn for hver pasient. Hver output voksel inneholder medianverdien til 3x3 ($2K + 1 = 9$ hvor $K = 4$) nabovoksler rundt den gjeldene vokselen i input bildet. Median filteret ble kun testet med LDA klassifiseringsmetoden. Median filteret erstatter vokslene på kantene av bildet med verdien null slik at alle filterregioner er like store. Dette fører til at medianverdiene for filterregioner som inneholder verdien null på kantene kan forvrenges.

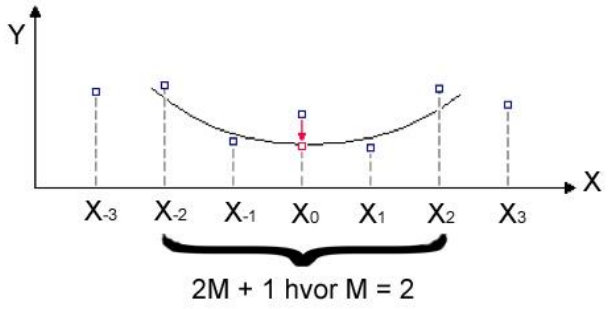
5.1.2 Savitzky-Golay filter

Savitzky-Golay filteret er en metode for datautjevning som er basert på en lokal tilnærming med minste kvadraters polynomer [58]. Dette er et lavpassfilter hvor hensikten er å fjerne så mye støy som mulig, uten at det går ut over underliggende informasjon i dataene [58].

Ideen bak Savitzky-Golay filteret er å tilpasse forskjellige polynomer til dataene som ligger rundt de ulike datapunktene [59]. La oss for eksempel si at man har et intervall av 2*M* + 1 datapunkter, hvor *M* er bredden til intervallet, se Figur 5-2. Datapunktene innenfor intervallet tilpasses en valgt polynom ved bruk av minste kvadraters metode [58], hvor det kun er punktet i midten av intervallet som glattes. Hvert punkt i datasettet glattes så ved å flytte intervallet et punkt til siden [60]. Denne prosessen kalles konvolering, og kan beskrives matematisk ved [58]:

$$X_j = \frac{\sum_{i=-M}^{i=M} C_i X_{j+i}}{N} \tag{5.1}$$

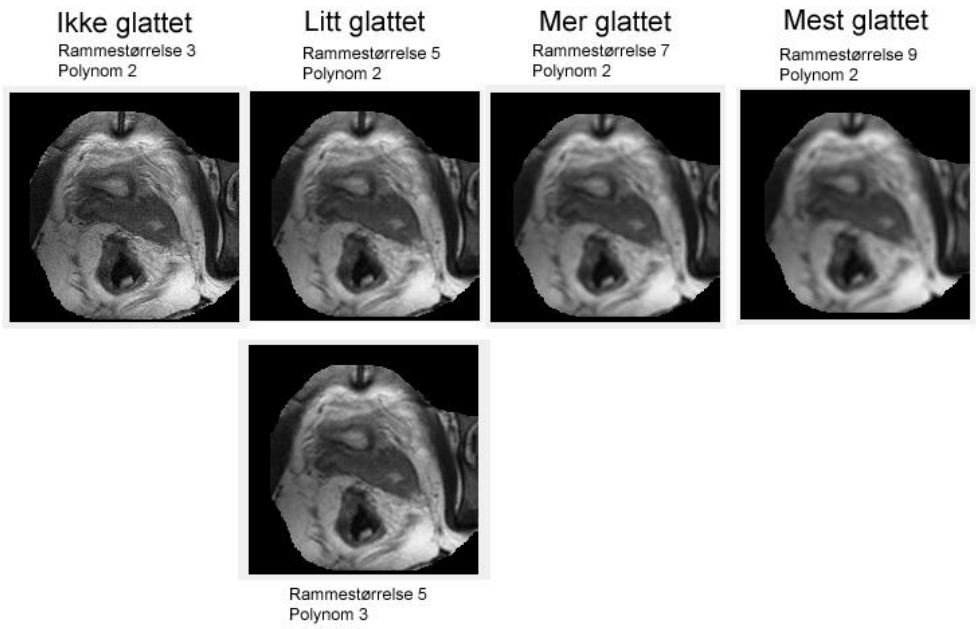
Hvor X_j er det glattede datapunktet, M er bredden til intervallet, C_i er konvoleringskoeffisientene, X_{j+i} er datapunktene i intervallet (hvor j representerer den løpende indeksen til de originale datapunktene) og N er antall datapunkt.



Figur 5-2: Savitzky Golay filteret glatter det midterste punktet i intervallet 2M+1 ved å tilpasse et valgt polynom ved bruk av minste kvadraters metode innenfor dette intervallet. Her er bredden av intervallet M=2 og intervallet har fem datapunkter. Hvert datapunkt glattes ved å flytte intervallet et punkt til siden. Hentet fra H. Lohninger http://www.statistics4u.com/fundstat_eng/cc_filter_savgolay.html, med tillatelse.

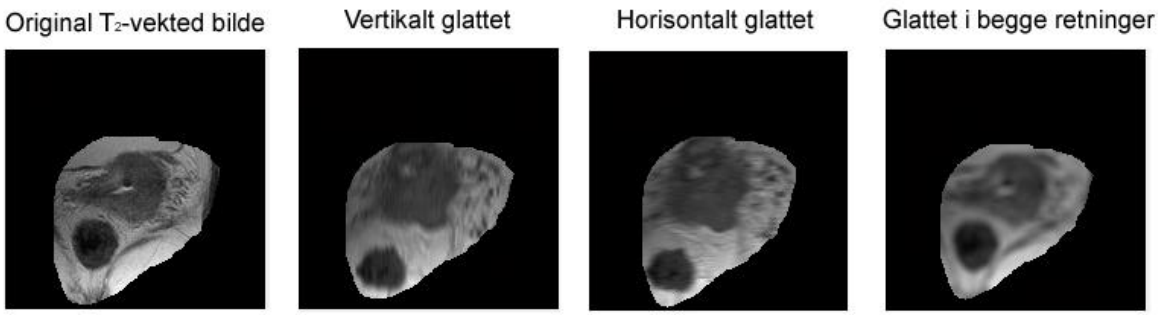
Hvis man antar at datapunktene har samme avstand mellom hverandre i *X*-retning vil ikke konvoleringskoeffisientene være avhengig av X_j eller avstanden i mellom dem, som vil føre til at konvoleringskoeffisientene C_i kun må beregnes en gang [61]. Det vil si at input verdiene i et intervall 2*M* + 1 kan effektivt kombineres med et fast sett av vektede koeffisienter som kan beregnes en gang for et gitt polynom *P* og et antatt intervall av lengden 2*M*+1 [62].

Figur 5-3 viser glatting med Savitzky-Golay filteret med forskjellige rammestørrelser og polynomer på MR-bildene. Med rammestørrelse menes et 2*M*+1 intervall som beskrevet over, og dette kan ikke være et partall. Hvis rammestørrelsen 2*M*+1 er, for eksempel, syv og polynomgraden er to, slik som vist i tredje bilde fra venstre i Figur 5-3, vil de syv nærmeste vokslene tilpasses en andregradspolynom ved en lineær minste kvadraters metode for vokselen i midten.



Figur 5-3: Figuren viser hvordan forskjellige innstillingene på rammestørrelse og polynom endrer graden av glatting med Savitzky-Golay filteret i det T_2-vektede bildet. Med rammestørrelse menes hele intervallet, slik at for rammestørrelse 3 er bredden M = 1.

Siden dette filteret i utgangspunktet er tilpasset et endimensjonalt datasett, må bildene glattes i henholdsvis horisontal og vertikal retning. Det enkelte bilde blir først glattet i vertikal retning, altså i radretning, og deretter blir det i tillegg glattet i horisontal retning (kolonneretning) ved å benytte de glattede punktene fra radene, se Figur 5-4.

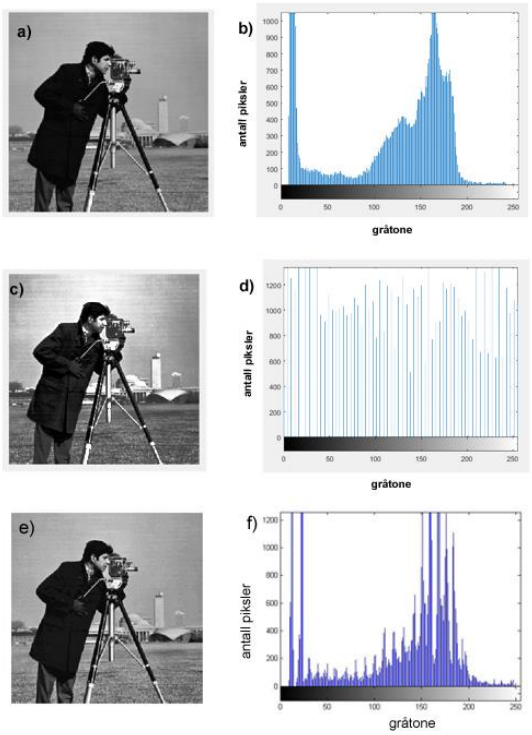


Figur 5-4: Savitzky-Golay filter brukt på et T_2-vekted bilde i vertikal, horisontal, og i tillegg i begge retninger. Rammestørrelsen er 11 (M = 5) og polynomsgraden er 2.

I denne oppgaven ble Savitzky-Golay filteret anvendt i begge retninger på alle bildesnitt og alle tidstrinn for hver pasient. Rammestørrelsene fem, syv og ni og polynomene to og tre, ble testet for DCE-bildeserien. For T_1-vektede og T_2-vektede bilder ble rammestørrelsene fem og ni, og polynom to testet. Rammestørrelsen er den parameteren som har størst virkningen på graden av glatting, som vist i Figur 5-3. Dette filteret ble kun testet med LDA klassifiseringen.

5.1.3 Contrast Limited Adaptive Histogram Equalization (CLAHE)

Histogrammet til et gråtonebilde representerer hvor mange piksler (voksler) hver gråtone har i et enkelt bilde [63], se Figur 5-5 a og b. Histogramutjevning generelt, er en prosess hvor man ønsker å finne en transformasjonsfunksjon $y = f(x)$ som mapper mellom input og output gråskalaverdier, slik at man får en uniform fordeling av histogrammet [63], se figur Figur 5-5 c og d.

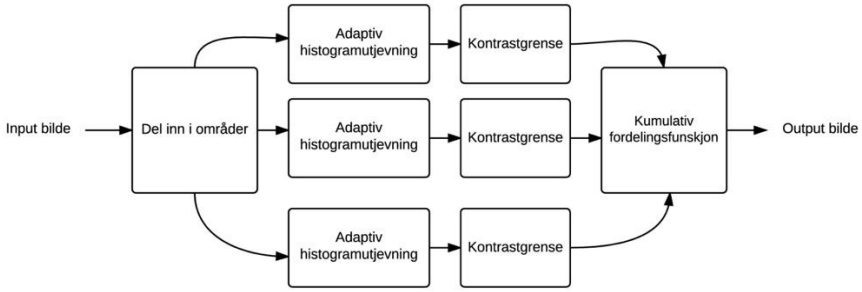


Figur 5-5: a) Viser originalversjonen av et gråtonebilde med det tilhørende histogrammet gitt i b). Dette histogrammet har en topp til venstre, som representerer de mørkeste pikslene i bildet. Det har også en topp midt i histogrammet, som representerer de lyse gråtonene i bakgrunnen. Bildet inneholder få helt lyse piksler, som vises helt til høyre i histogrammet. c) Viser det histogramutjevnede bildet med det tilhørende histogrammet gitt i d). Dette er et tilnærmet uniformt histogram, hvor det er omtrentlig like mange mørke som lyse piksler i bildet. e) Viser bildet hvor det er utført CLAHE med 8x8 piksler, 0,01 kontrastgrense, 32 bins og uniform distribusjon, og det tilhørende histogrammet gitt i f). På grunn av kontrastgrensen vil ikke det lyse området over hodet til fotografen være like tydelig som i det histogramutjevnede bildet i c).

Histogramutjevning er basert på antagelsen om at bildekvaliteten er ensartet over alle områder og en unik gråtone kartlegging gir tilsvarende forbedringer for alle områder i bildet [64]. Når fordelingen av gråtoner endrer seg fra et område til et annet, er ikke denne antagelsen gyldig [64]. *Adaptive histogram equalization* løser dette problemet, ved å finne transformasjonen til hver piksel basert på den lokale gråskalafordelingen (nabopiksler) [65-67]. Det vil si at hver piksel kontrastforsterkes med en transformasjonsfunksjon basert på intensitetsverdiene til de nærliggende pikslene.

I noen tilfeller kan også gråtonefordelingen i bildet være konsentrert i en liten del av histogrammet. I disse tilfellene kan kartleggingskurven bli bratt i noen områder, som betyr at to svært nærliggende gråtoner kan få veldig forskjellige gråtoner og bildet vil få en høy kontrast. Dette problemet løses ved å begrense kontrasten som er tillatt gjennom histogramutjevningen. Kombinasjonen av denne kontrastbegrensede tilnærmingen og den nevnte adaptive histogramutjevningen gir det som omtales som CLAHE (*Contrast Limited Adaptive Histogram Equalization*) [68, 69] og vises i Figur 5-5 e og f. Figuren viser tydelig hvordan kontrastgrensen til CLAHE begrenser det lyse område øverst i bildet, sammenlignet med det histogramutjevnede bildet vist i Figur 5-5 c.

CLAHE filteret er bygd opp som vist i Figur 5-6, hvor filteret deler bildet inn i flere ikke-overlappende områder med omtrent lik størrelse. Først blir histogrammet for hvert område beregnet, ved bruk av adaptiv histogramutjevning. Deretter, med en ønsket grense for kontrastutvidelse, blir histogrammene omfordelt slik at histogrammenes høyde ikke overstiger kontrastgrensen. Til slutt blir den kumulative fordelingsfunskjonen for det resulterende kontrastbegrensede histogrammet bestemt for gråskalafordelingen [50]. CLAHE fordeler pikslene ved en lineær kombinasjon av de fire nærmeste områdene.



Figur 5-6: Oppbyggingen av Contrast Limited Adaptive Histogram Equalization (CLAHE). Input bildet fordeles i områder med omtrent lik størrelse, hvor områdene ikke overlapper hverandre. Først beregnes histogrammet for hvert område, ved bruk av adaptiv histogramutjevning. Deretter, med en ønsket grense for kontrastutvidelse, omfordeles histogrammene slik at histogrammenes høyde ikke overstiger kontrastgrensen. Til slutt bestemmes den kumulative fordelingsfunksjonen for det resulterende kontrastbegrensede histogrammet for gråskalafordelingen, som benyttes på output bildet [50]. CLAHE fordeler pikslene ved en lineær kombinasjon av de fire nærmeste områdene.

Contrast Limited Adaptive Histogram Equalization (CLAHE) ble testet som en preprosesseringsalgoritme for MR-bildene hvor de ulike parameterne som ble testet er vist i Tabell 5-1. Alle parameterne ble testet i kombinasjon, i tillegg til å bli testet på både DCE, T_1-vektede og T_2-vektede bilder. Dette filteret ble kun testet med LDA klassifiseringen.

Parameter	**Input verdier**
Antall piksler [M,N]	[8,8], [32,32] og [64,64]
Kontrastgrense	0,005, 0,01 og 0,02
Antall bins i histogrammet	256, 64 og 32
Distribusjon	Uniform og rayleigh

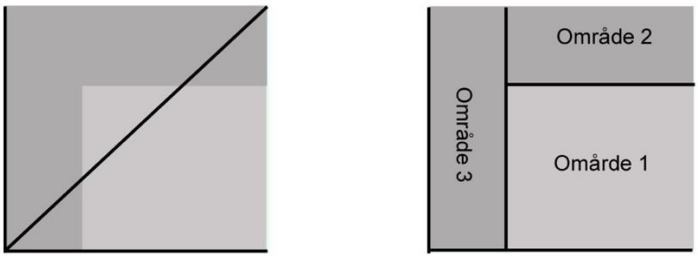
Tabell 5-1: De ulike parameterne for CLAHE (Contrast Limited Adaptive Histogram Equalization) filteret som ble testet. Alle parameterne ble testet i kombinasjon, i tillegg til å bli testet på både DCE, T_1-vektede og T_2-vektede bilder.

5.2 Klassifiseringsmetoder

I denne oppgaven var det to klasser av voksler, svulst og ikke-svulst. Hensikten med å teste datasettet for andre klassifiseringsmetoder, var å undersøke om ikke-lineære klassifiseringsalgoritmer identifiserte svulstvokslene mer nøyaktig enn LDA modellen som allerede var implementert. De ikke-lineære metodene som ble testet er overvåkede metoder, som betyr at de hadde en responsvektor **Y**. Denne responsvektoren var den grunnleggende sannheten, inntegnet av radiologene, og ble brukt i modelltreningen, samt til å vurdere nøyaktigheten til programmets svulstinntegninger. Klassifiseringsmetodene som ble implementert og testet beskrives nøyaktig i kapitlene under.

5.2.1 Random Forest

Random Forest er en form for beslutningstre, som er en metode for både regresjon og klassifikasjon [70]. Metoden er spesielt god i tilfeller hvor man ikke har en lineær struktur og hvor man har en høy kompleksitet mellom variablene og responsene, som illustrert i Figur 5-7 [70]. Datasettet deles inn i bestemte områder, og for et klassifikasjonstre, predikeres det at hver observasjon hører til den klassen hvor treningsobservasjonene forekommer hyppigst i det tilhørende området [70]. Ved tolking av resultatene er man ikke bare interessert i den predikerte klassen som tilhører et bestemt område, men også antallet treningsobservasjoner fra hver klasse som faller inn under det området [70].



Figur 5-7: Figuren til venstre viser hvordan en lineær analyse av dataene ikke fanger opp skillet mellom de to gruppene (mørkegrått og lysegrått) like godt som beslutningstreet til høyre. Innenfor de forskjellige områdene vil det utnevnes en klasse, som blir bestemt ut i fra de treningsobservasjonene som forekommer hyppigst.

Hovedproblemet med vanlige beslutningstrær er høy varians [70]. En *bootstrap* analyse blir gjort for å redusere variansen, og gjøres ved å ta et gjennomsnitt av observasjonene [70]. La oss si at vi har sett av *n* uavhengige variabler $Z_1,...,Z_n$, hvor hver av disse variablene har varians σ^2. Variansen til gjennomsnittet av variablene $\bar{Z}$ er gitt ved $\frac{\sigma^2}{n}$, som betyr at variansen reduseres ved å ta gjennomsnittet av variablene. For at dette skal ha en effekt må man ha tilgang til flere treningssett, noe man som regel ikke har [70]. Da kan man isteden lage *N* forskjellige *bootstrappede* treningssett fra det tilgengelige treningssettet, og trene modellen på disse [70]. *N* forskjellige beslutningstrær konstrueres ved å bruke *N* *bootstrappede* treningssett [70]. Hver gang en deling blir vurdert under konstruksjonen av trærne, blir en tilfeldig prøve av *m* variabler valgt som delingskandidater fra det fulle settet av *p* variabler [70]. Når en gren splittes kan algoritmen kun vurdere disse *m* variablene [70]. En ny prøve med *m* variabler blir valgt ut for hver splitt, og typisk velger man

$$m = \sqrt{p} \tag{5.2}$$

hvor *m* er antallet tilfeldige variabler som vurderes ved hver splitt av totalt *p* variabler.

Med andre ord har ikke algoritmen mulighet til å vurdere alle variablene i hver splitt [70]. Dette gjør at dominerende variabler ikke vil påvirke splittene i like stor grad, og man unngår mange like trær [70]. Bruk av en liten verdi for *m* når man bygger en *Random Forest* vil typisk være gunstig når man har et stor antall korrelerende variabler [70].

I denne oppgaven ble Radom Forest klassifiseringsmetoden testet for datamatrisen **X** med livmorhalskreft pasientene. Ved bruk av denne algoritmen må *N* antall beslutningstrær velges for modellen og det er valgt $\sqrt{p}$ delingskandidater. Modellen ble testet for 10, 50, 100 og 200 trær for å vurdere klassifiseringen i henhold til antall trær.

5.2.2 AdaBoost

Boosting er en metode som fokuserer på de områdene som er vanskelige å klassifisere [71]. I *boosting* gror trærne fortløpende, i motsetning til *Random Forest* hvor trærne gror uavhengig av hverandre og deretter settes sammen [70]. Hvert tre skapes ved å bruke informasjon fra det forrige [70]. I *boosting* lager man ikke flere forskjellige *bootstrappede* treningssett, men i stedet er hvert tre tilpasset en modifisert versjon av det originale datasettet [70].

AdaBoost genererer en sekvens av svake klassifiseringer, altså der hvor klassifiseringsmodellen kun klarer å predikere marginalt bedre enn en tilfeldig tildeling av klasse [71]. For hver iterasjon finner algoritmen den beste klassifiseringen basert på de gjeldene vektede datapunktene [71]. Dette vil si at når datapunkter er feilaktig klassifisert i den *k´te* iterasjonen, vil de bli høyere vektet i den *(k+1)te* iterasjonen, mens datapunkter som er korrekt klassifisert vil få mindre vekting i den påfølgende iterasjonen [71]. Dette betyr at datapunkter som er vanskelige å klassifisere vil bli vektet høyere helt til algoritmen identifiserer en modell som klarer å klassifisere disse datapunktene korrekt [71]. *AdaBoost* algoritmen har egenskapen at den lærer mens den analyserer datapunktene, og vil fokusere på de områdene i datasettet som er vanskelige å klassifisere [71].

En trinnvis vekting beregnes basert på feilraten til hver iterasjon [71]. La oss si at vi har et sett datapunkter, som skal klassifiseres i klassene +1 og -1. Hvert datapunkt har samme vekt ved startpunktet [71]. For hver iterasjon fra k=1 til K vil man tilpasse en svak klassifikator ved å bruke de vektede datapunktene og beregne feilklassifiseringen err_k for hver k [71]. Deretter beregner man den trinnvise vektingen for den *k´te* iterasjonen [71], gitt som

$$\ln(\frac{1 - err_k}{err_k}) \tag{5.3}$$

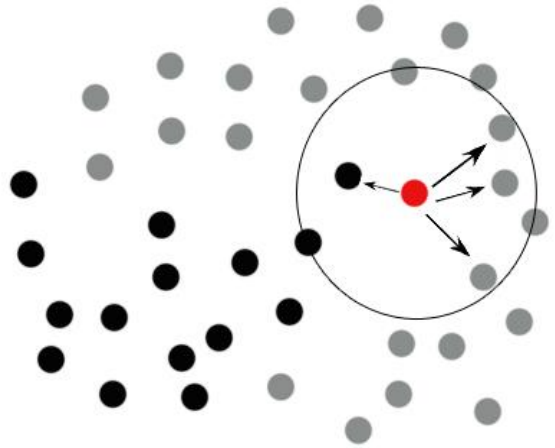
hvor err_k er feilklassifiseringen til den *k´te* iterasjonen.

Ved å bruke denne feilklassifiseringen kan man vekte de datapunktene som er feilaktig klassifiser mer enn de datapunktene som har blitt klassifisert korrekt [71]. Datapunktene klassifiseres så ved å multiplisere den *k´te* trinnvise vektingen med den *k´te* modell prediksjonen og deretter legge sammen disse over k. Hvis summen er positiv, vil datapunktet få klasse +1, og er den negativ vil den få klasse -1 [71].

I denne oppgaven ble *AdaBoost* testet for et maksimum antall splitt på 100. I tillegg ble det testet ulike minimum bladstørrelser, som er antallet voksler i hver ende av treet, disse ble satt til 100, 100, 100 000, 500 000 og 1 000 000. Antall trær ble satt til 10, 50, 100 og 200.

5.2.3 K nærmeste nabo (kNN)

K-nærmeste nabo (kNN) er en klassifiseringsalgoritme som beregner avstanden mellom en ukjent prøve og hver av prøvene i treningssettet [72]. Ved klassifiseringen av en ny prøve finner man de *K* nærmeste naboene i treningssettet og klassifiserer den ukjente prøven til gruppen som har flest medlemmer blant disse naboene, se Figur 5-8 [72].



Figur 5-8: De svarte og grå punktene viser prøver fra to forskjellige grupper. Klassifiseringsmetoden K nærmeste nabo bruker de K nærmeste naboene (altså avstanden mellom prøvene) fra treningssettet og klassifiserer den ukjente prøven (rødt punkt) til den gruppen som har flest medlemmer blant disse naboene. Ved én nærmeste nabo vil det røde punktet klassifiseres til den svarte gruppen, mens ved tre nærmeste nabo vil den klassifiseres til den grå gruppen. Dette er en typisk situasjon som kan være vanskelig å klassifisere ved bruk av LDA/QDA [72], og hvor kNN metoden vil fungere bra [72].

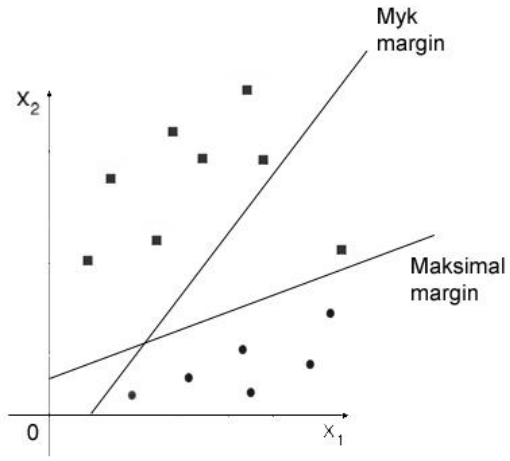
Fordelen med denne type klassifisering er at man ikke må gjøre noen antagelse om formen til de forskjellige klassene [72]. Sjansen for at et nytt datapunkt er nærmere en klasse enn en annen klasse er blant annet avhengig av antallet datapunkt som finnes av de forskjellige klassene i datasettet [72]. Ideelt burde fordelingen av klassene i treningssettet representere den forutbestemte sannsynligheten for at det ukjente datapunktet hører til en av klassene [72]. Hvis for eksempel en sjelden klasse er overrepresentert i treningssettet vil denne klassen plukke opp for mange av de ukjente datapunktene [72]. Derfor burde fordelingen av klasser i treningssettet gjenspeile den antatte fordelingen av klasser i de ukjente datapunktene.

En annen ulempe er at beregningene må gjøres for alle ukjente datapunkt [72]. Hvert ukjent datapunkt må sammenlignes med alle datapunktene i treningssettet, det vil si at alle disse sammenhengene må huskes og prosesseres når klassifikasjonen skjer [72]. Om et stort datasett blir brukt kan denne jobben bli meget omfattende [72].

I denne oppgaven testes kNN algoritmen for en, to og tre naboer, ved bruk av en euklidsk avstand mellom datapunktene.

5.2.4 Støtte Vektor Maskiner (SVM)

Støtte vektor maskiner (SVM) brukes til binære klassifiseringer, som vil si at det kun finnes to klasser [70], og kan bruke både lineære og ikke-lineære grenser mellom klassene. Den lineære grensen kan være en maksimal margin eller en myk margin [70]. En maksimal margin setter en lineær grense mellom klassene, slik at alle observasjonene for en klasse faller på hver sin side av grensen, se Figur 5-9 [70]. Dette gjør at enkelte observasjoner som kan karakteriseres som uteliggere innad i gruppene i treningssettet, påvirker den lineære grensen betraktelig [70]. Dette kan løses ved å benytte en myk margin, som tillater at noen treningsobservasjoner faller på feil side av den lineære grensen, se Figur 5-9 [70].



Figur 5-9: To forskjellige klasser med to variabler X_1 og X_2, er vist med runde og firkantede punkter. Ved den maksimale marginen faller de runde og firkantede punktene på hver sin side av den lineære grensen. Ved en myk margin faller den ene av de firkantede punktene på feil side av grensen, som gjør at grensen endrer seg betraktelig [70].

Ved en maksimal margin blir grensen mellom de to gruppene valgt der avstanden mellom de to gruppene er størst. For et todimensjonalt system er denne grensen en linje, for et tredimensjonalt system er denne grensen et plan og for et flerdimensjonalt system kalles dette et hyperplan [70, 73]. Dette hyperplanet *f(x)* har en beslutningsregel som deler planet inn i to regioner [73]:

$$f(x) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0 \tag{5.4}$$

hvor $\boldsymbol{x}$ er en treningsobservasjon i planet, $\boldsymbol{w}$ er en vektet vektor som er vinkelrett på hyperplanet og b er en konstant som legger hyperplanet vekk fra origo [73].

Fortegnet til *f(x)* angir hvilket område datapunktet tilhører [73]. For å bestemme de såkalte støtte vektorene velges en margin slik at hyperplanet som skiller de to regionene er

maksimert [73]. Marginen er lik d_- og d_+ (se Figur 5-10) hvor fortegnene + og – skiller de to områdene. Den totale marginen mellom klassene er d_- pluss d_+ [73]. Støttevektorene er datapunktene i treningssetet som vil påvirke posisjonen til hyperplanet om de flyttes på, altså punktene på marginlinjene som oppfyller et av kravene [70, 73]:

$$f(x_+) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = +1 \quad (5.5)$$
$$f(x_-) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = -1 \quad (5.6)$$

Det vil si at alle punktene som har verdien

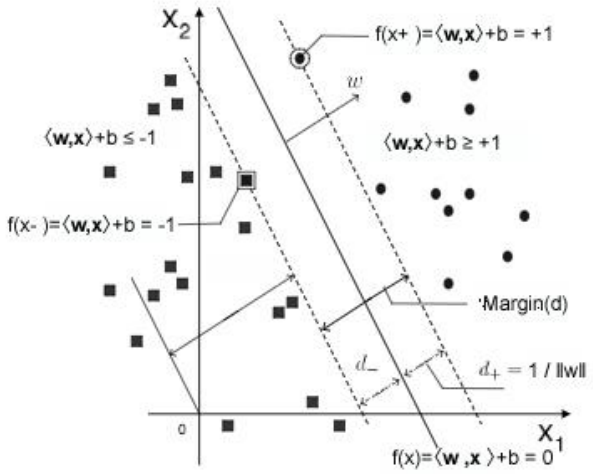
$$\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b \geq +1 \quad (5.7)$$

vil havne på den positive siden av hyperplanet og klassifiseres med verdien +1 [73], mens alle punktene som har verdien

$$\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b \leq -1 \quad (5.8)$$

vil havne på den negative siden av hyperplanet og klassifiseres med verdien -1 [73]. For to klasser +1 og -1, som er illustrert i Figur 5-10, vil avstanden mellom klassene være

$$Margin(d) = \frac{2}{\|\boldsymbol{w}\|} \quad (5.9)$$



*Figur 5-10: Hyperplanet f(x) = <**w,x**> + b = 0 deler rommet inn i to klasser, med en margin $d_{\pm}$. Alle punktene som har verdier f(x) = ± 1 er støttevektorer (og) . Alle punktene som har verdier f(x) ≥ +1 hører til klassen med runde punkter, mens punktene merket med firkanter har verdier f(x) ≤ -1 og hører til den andre klassen. Her er det brukt en lineær funksjonsgrense og en maksimal margin. Hentet fra Härdle et al. [73], med tillatelse fra forfatteren.*

Ikke alle klasser er lineært separable og kan dermed ikke skilles ved en rett linje. Ved ikke-lineær klassifisering kan man benytte annen, tredje og til og med høyere ordens ligninger for å finne en grense som best mulig skiller gruppene [70]. Dette gjøres ved å transformere de opprinnelige treningsobservasjonene til et høyeredimensjonalt rom kalt *feature*-rommet *H*, hvor de kan skilles ved en rett linje. Dette gjør at man kan bruke lineære klassifiseringsmodeller på ikke-lineære sammenhenger, ved hjelp av en kernel [70]. Hovedpoenget ligger i det indre produktet mellom observasjonene, og en lineær støtte vektor klassifikator kan bli representert som [70, 73] :

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \tag{5.10}$$

hvor det er *n* parametere α_i, i = 1, ..., *n, x* er en ny prøve, x_i er en treningsobservasjon og β_0 er en parameter.

For å estimere parameterne α_1,...., α_n og β_0 trenger man det indre produktet mellom alle parene av treningsobservasjoner i treningssettet $\langle x_i, x_{i'} \rangle$, hvor x_i er en treningsobservasjon og $x_{i'}$ er en annen treningsobservasjon [70]. Det indre produktet mellom den nye prøven x og hver av treningsobservasjonene x_i må også beregnes [70]. Fordelen er at α_i er forskjellig fra null kun for støtte vektorene [70]. Det vil si at hvis en treningsobservasjon ikke er en støtte vektor, så er α_i null [70]. Det betyr at alt man trenger for å representere en lineær klassifikator *f(x)*, og beregne dens parametere, er det indre produktet [70], slik at hver gang det indre produktet dukker opp i beregningen av støtte vektor klassifiseringer så bruker vi en generalisering av det indre produktet på formen:

$$K(x_i, x_{i'}) \tag{5.11}$$

Hvor *K* er en funksjon som vi refererer til som kernelen. En kernel er den funksjon som kvantifiserer likheten mellom to observasjoner [70], for eksempel den lineære kernelen:

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{5.12}$$

hvor $x_i$og $x_{i'}$ er to forskjellige treningsobservasjoner og p er dimensjonen til vektorrommet.

Man kan også benytte en polynom kernelen [70]:

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d \tag{5.13}$$

hvor *d* er antall frihetsgrader.

Om *d* er større enn 1 blir kernelen ikke-lineær og vil føre til en mer fleksibel bestemmelsesgrense mellom klassene [70]. Et annet alternativ er en radial kernel på formen

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \tag{5.14}$$

hvor γ er en positiv konstant.

Dermed blir den ikke-lineære støtte vektor klassifikatoren på formen:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) \tag{5.15}$$

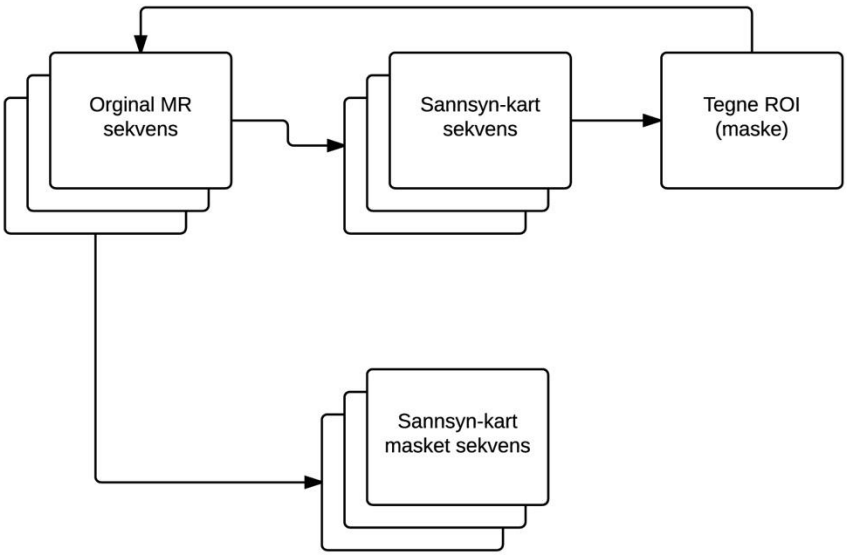
hvor *S* er antall støtte punkt som har α_i forskjellig fra null.

I denne oppgaven ble en SVM klassifiseringsalgoritme med en lineær kernel testet.

5.3 Postprosesseringsmetoder

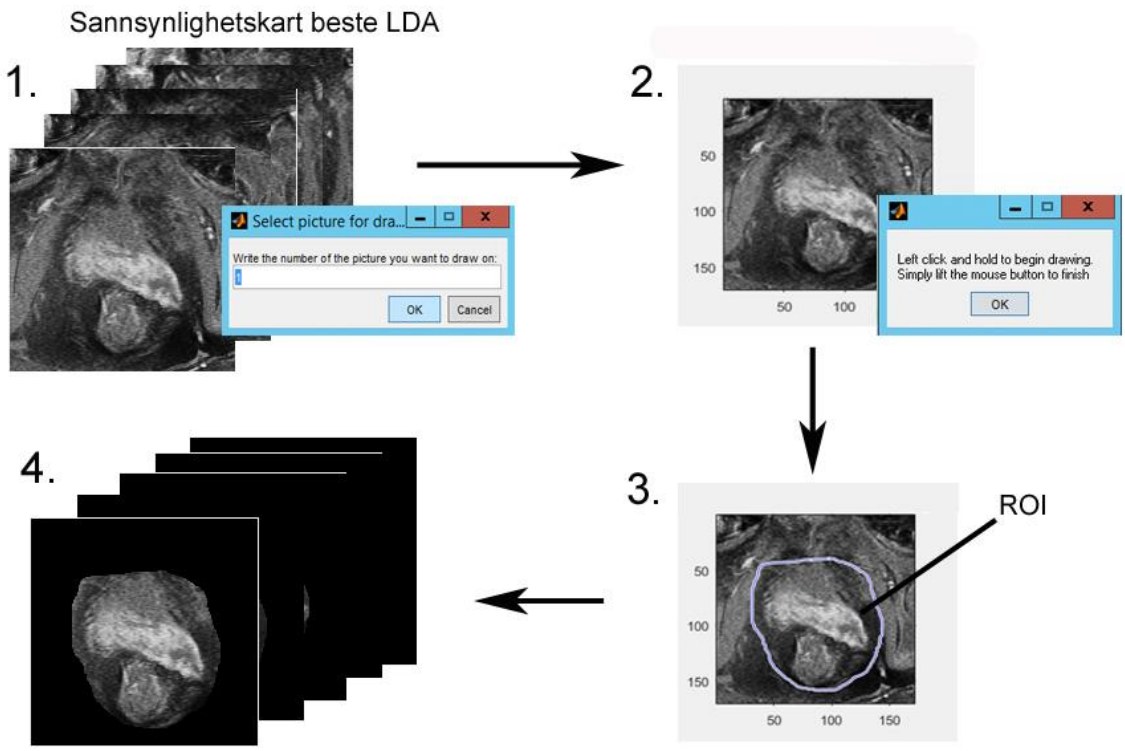
Postprosesseringen er en behandling av bildene etter klassifiseringen er utført. Her ønsker man å finne en metode som kan forbedre inntegningen av kreftsvulsten og fjerne feilklassifiserte voksler. I denne oppgaven er det testet en ROI (*Region of Interest*) metode, hvor en maske blir tegnet rundt vokslene som mest sannsynlig er svulst, ved å analysere sannsynlighetskartene generert av den lineære klassifiseringen.

5.3.1 ROI (*Region Of Interest*)



Figur 5-11: Den originale MR sekvensen går igjennom en lineær eller en ikke-lineær klassifisering og resulterer i et sannsynlighetskart for hvor svulsten befinner seg. Radiologen kan gå inn å velge ut et bilde hvor det kan tegnes inn en ROI (Region of Interest). Denne regionen brukes videre som en maske som legges opp på de originale bildene. Dette resulterer i bilder med færre voksler som igjen kan kjøres igjennom en klassifisering for å få ut et nytt sannsynlighetskart.

Ved bruk av ROI (*Region of Interest*) kan radiologene i postprosesseringen velge ut et ønskelig område av sannsynlighetskartet som er generert av den beste LDA modellen. En forenklet modell av denne prosessen er vist i Figur 5-11. Hensikten er å kombinere kunnskapen til radiologene og automatiseringen til programmet. Programmet gir et sannsynlighetskart ved hjelp av enten en lineær eller ikke-lineær vokselklassifikasjon, se Figur 5-12 1.) Sannsynlighetskartene for hver pasient samles i en sekvens slik at man enkelt får en oversikt over bildene. På ett av bildene i denne sekvensen kan det tegnes inn et egendefinert område (ROI) som inneholder områder med høy sannsynlighet for svulst, se Figur 5-12 2, 3.) Dette område lagres som en maske, som legges på alle orignalbildene i pasientens bildesekvens, se Figur 5-12 4.) Dette gjør at bildet får en høyere prosentandel svulst, som igjen fører til en endring i balansen mellom de to klassene (svulst og ikke-svulst) i datasettet. Disse maskede bildene har færre voksler og kan enkelt klassifiseres på nytt, enten ved en lineær klassifiseringsmetode eller ved ikke-lineære klassifiseringsmetoder. Vokslene utenfor det valgte område klassifiseres automatisk som friskt vev.



Figur 5-12: Figuren viser hvordan programmet samarbeider med radiologen for å lage en ROI maske. 1) Viser sannsynlighetskartene for den beste LDA modellen, hvor radiologene kan bla seg igjennom bildesekvensen og velge det bildet de ønsker å tegne masken på. 2) Viser det valgte bildet og instruks på hvordan man skal tegne ROI masken. 3) Viser hvor masken er tegnet, alt utenfor denne grensen vil automatisk klassifiseres som ikke-svulst. 4) Den samme masken legges på hele bildesekvensen, og radiologen kan bla igjennom for å sjekke at masken stemmer for alle bildesnittene.

I denne oppgaven ble det tegnet en ROI maske på alle bildesnitt for alle pasienter ved å se på sannsynlighetskartet gitt av den beste LDA modellen. Dette reduserte vokselantallet som ble inkludert i analyser med 65 %, hvor 81% av vokslene var svulst og 19% av vokslene var friskt vev.

5.4 Validering

For å redusere tidsbruken knyttet til modelltreningen ble det brukt en *leave-ten-patients-out* kryss-validering, istedenfor en *leave-one-patient-out* kryss-validering som er brukt i det opprinnelige programmet. Dette førte til at treningstiden ble redusert betraktelig spesielt for ikke-lineære klassifiseringsmetoder, fordi treningen kun måtte gjøres åtte ganger istedenfor 78 ganger. Ved å sammenligne ytelsesparameterne fra de to ulike valideringene ved den beste LDA modellen fant man at forskjellen var liten, som vist i Tabell 5-2.

Ytelsesparameter	*Leave-one-patient-out* kryss-validering	*Leave-ten-patients-out* kryss-validering	Differanse
DSC	0,5301	0,5292	0,0009
Kappa	0,5095	0,5085	0,0010
Sens	0,9259	0,9255	0,0004
Spes	0,9090	0,9088	0,0002
AUC	0,8134	0,8134	0,0000

Tabell 5-2: Forskjellen (Differansen) på ytelsesparameterne for leave-one-patient-out kryss-validering og leave-ten-patients-out kryss-validering for den beste LDA modellen, med ROI masken. Denne modellen inkluderte T_1-vektede og T_2-vektede bilder med 8 naboer, og DCE bilder med 0 naboer, hvor alle bildene er autoskalert.

Leave-ten-patients-out kryss-valideringen ble gjort ved å ta ut de ti første pasienter i datasettet og trene klassifiseringsmodellen på de resterende bildene. Klassifiseringsmodellen ble brukt til å tegne inn svulsten til de ti pasientene som ble utelatt i modelltreningen. Deretter gjentas dette for de ti neste pasientene i datasettet, helt til alle pasientene i grupper av ti hadde blitt utelatt en gang. Ytelsen av modellen ble deretter vurdert på samme måte som i det opprinnelige programmet, hvor DSC, Kappa, Sens, Spes og AUC ble beregnet for modellen.

5.5 Statistikk

I denne oppgaven ble det brukt en analyse av variansen for å teste om gjennomsnittet til ulike ytelsesparameter for klassifiseringsmodeller er signifikant forskjellige eller ikke. Om nullhypotesen (H_0) ble avvist, vil det føre til konklusjonen om at gjennomsnittene ikke er signifikant like [74]. Allikevel vil ikke dette nødvendigvis bety at gjennomsnittene var signifikant forskjellige [74].

Når man utfører flere sett av *t*-tester for å undersøke forskjellen mellom to gruppers gjennomsnitt, kan sannsynligheten for at de er signifikant forskjellige skyldes et stort antall tester [74]. Det vil si at for noen av testene, vil denne metoden feilaktig påstå at noen av gruppenes gjennomsnitt er signifikant forskjellige. Dette kommer av at *t*-testene bruker data fra de samme prøvene, og er dermed ikke uavhengige av hverandre [74]. Dette gjør at det er vanskeligere å kvantifisere nivået av signifikans for flere tester.

Anta en enkel *t*-test, sannsynligheten for at nullhypotesen (H_0) blir avvist når den faktisk er sann er liten, si 5%. Anta videre at det utføres seks uavhengige *t*-tester. Hvis signifikansnivået for hver test er 5%, vil sannsynligheten for at testene riktig unngår å avvise nullhypotesen, når den er sann for hvert tilfelle være $(0.95)^6 = 0,735$ [74]. Og sannsynligheten for at en av testene feilaktig avviste nullhypotesen er 1-0,735 = 0,265, noe som er mye høyere enn 0,05 [74].

For å kompensere for flere tester brukes MATLAB® *multiple comparison* prosedyre [74]. Denne prosedyren utfører flere parvise sammenligninger av gjennomsnittet til hver gruppe, og anvender Tukey´s *honestly significant difference criterion* som sier avvis H_0: $a_i=a_j$ hvis

$$|t| = \frac{|\bar{y}_i - \bar{y}_j|}{\sqrt{MSE(\frac{1}{n_i}+\frac{1}{n_j})}} > \frac{1}{\sqrt{2}} q_{\alpha,k,N-k} \tag{5.16}$$

hvor $q_{\alpha,k,N-k}$ er den øvre 100*(1- α) prosenten av den studentiserte fordelingen med signifikansnivå α med parameter *k* og *N – k* frihetsgrader. *K* er antall grupper og *N* er det totale antallet observasjoner. $\bar{y}_i$ og $\bar{y}_j$ er gjennomsnittet til to forskjellige grupper og n_i og n_j er antallet observasjoner innenfor de respektive gruppene. *MSE* er *Mean Square Error.*

Når to eller flere gjennomsnitt sammenlignes i denne oppgaven vil det alltid være på et 0,05 (5%) signifikansnivå og korrigert for multiple tester.

6 Resultat

6.1 Test av ROI postprosesseringsmetoden

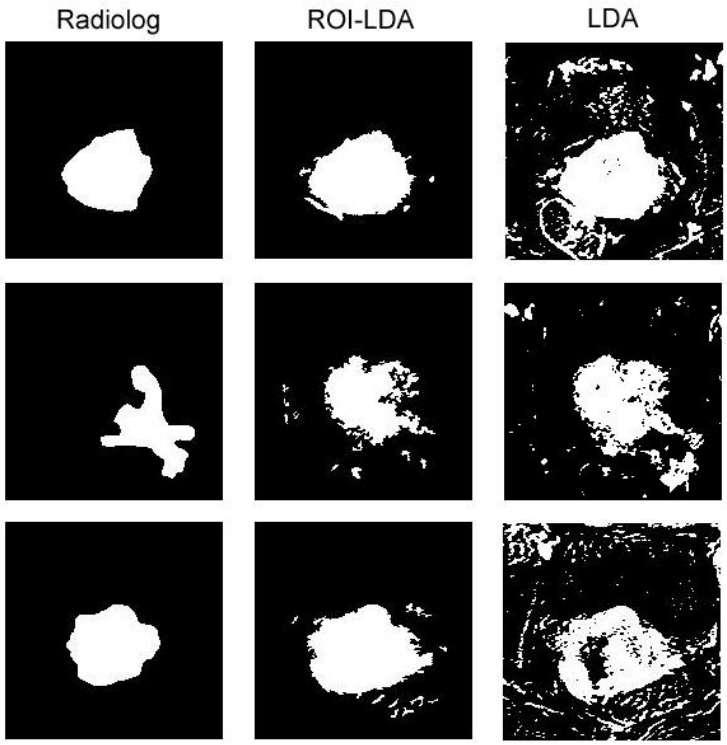
6.1.1 Klassifisering med LDA

ROI (*Region of Interest*) postprosesseringen ble først testet for LDA klassifiseringsmetoden, for å undersøke om ROI masken førte til en bedre klassifisering. LDA klassifiseringen med ROI masken (ROI-LDA) ble sammenlignet med tidligere resultater for LDA klassifiseringen uten ROI masken (LDA). Først sammenlignes ROI-LDA modellene basert på T_1-vektede og T_2-vektede bildensnitt, med LDA modellene basert på samme bildetyper. Deretter sammenlignes ROI-LDA og LDA modellene som inkluderer DCE bildeserien. Til slutt sammenlignes den beste modellen fra LDA klassifiseringen med og uten ROI masken.

Alle resultatene fra ROI-LDA klassifiseringen presenteres i Tabell B-1, Vedlegg B: Vokselklassifisering med ROI-LDA. Tabell B-1 viser at alle ROI-LDA modellene som var basert på T_1-vektede eller T_2-vektede bilder hadde en DSC verdi mellom 0,39 – 0,44 og en *K* verdi mellom 0,36 – 0,42. Dette indikerer en overenstemmelse med radiologene som var bedre enn tilfeldig. Likeledes presenteres alle resultatene fra LDA klassifiseringen uten ROI masken i Tabell A-1, Vedlegg A: Vokselklassifisering med LDA. Denne tabellen viser at alle modellene basert på T_1-vektede eller T_2-vektede bilder hadde en DSC verdi mellom 0,19 – 0,22 og en *K* verdi mellom 0,15 – 0,21. Ved å sammenligne DSC og *K* verdiene for LDA klassifiseringen med og uten ROI masken fant man at ROI-LDA ga signifikant høyere DSC og *K* parametere. Videre viste resultatene fra modellene med T_1-vektede eller T_2-vektede bilder at ROI-LDA klassifiseringen hadde en noe lavere sensitivitet (0,85-0,94), men ikke signifikant forskjellig fra LDA klassifiseringen (0,90-0,95). Spesifisiteten var derimot signifikant høyere for de maskede bildene (0,85-0,87) enn for bildene uten masken (0,50-0,53). Dette indikerer at ROI-LDA modellen klassifiserte begge gruppene ikke-svulst og svulst mer korrekt, enn LDA modellen basert på bildene uten ROI masken.

I likhet med LDA klassifiseringen ble ytelsen til ROI-LDA klassifiseringen betraktelig forbedret når DCE bildeserien ble inkludert. Ved sammenligning av alle resultatene i Tabell A-1 og Tabell B-1 som inkluderte DCE bildeserien hadde ROI-LDA en DSC mellom 0,51 – 0,53, mens LDA modellene hadde en DSC på 0,38 – 0,42. *K* verdiene var i område 0,49 – 0,51 for ROI-LDA modellene, mens de lå mellom 0,35 – 0,40 for LDA modellene uten ROI masken. Dette betyr at DSC og *K* for ROI-LDA modellene var signifikant høyere, og ga dermed en bedre klassifisering av svulsten. Dette vises i Figur 6-1, hvor radiologens maske sammenlignes med vokselklassifisering fra ROI-LDA og LDA modellene. For den første pasienten (første rad), var den binære masken god for ROI-LDA modellen (DSC=0,74), mens den var noe lavere for LDA modellen (DSC = 0,56). Tilsvarende tilfelle vises i rad tre hvor ROI-LDA (DSC=0,71)

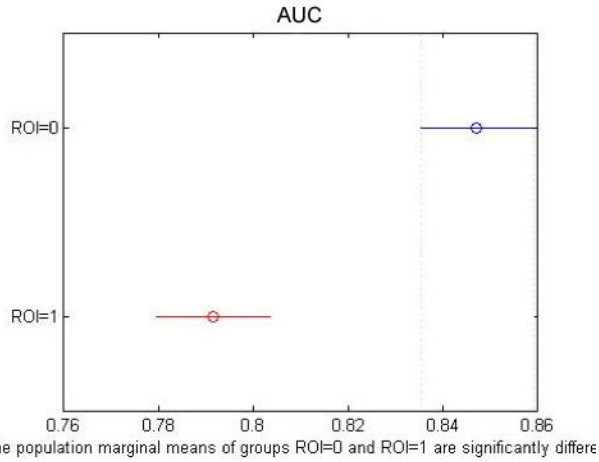
klassifiserte de fleste vokslene korrekt, i forhold til LDA (DSC = 0,51) som hadde flere feilklassifiserte voksler. Den andre raden viser et bildesnitt hvor enigheten mellom radiologen og LDA modellene var lavere, både for ROI-LDA (DSC =0,30) og LDA (DSC=0,35).



Figur 6-1: Valgte snitt fra tre forskjellige pasienter for å illustrere variasjoner i LDA klassifiseringen. Den første kolonnen viser radiologens maske, den andre kolonnen viser LDA klassifiseringen med ROI postprosessering og den tredje kolonnen viser LDA klassifiseringen uten ROI postprosessering. Tre forskjellige pasienter ble valgt, med forskjellig grad av korrekt klassifisering.

Både ROI-LDA og LDA modellene fikk en jevnere klassifisering av gruppene ikke-svulst og svulst når DCE bildeserien ble inkludert. Spesifisiteten og sensitiviteten lå mellom 0,90 – 0,91 for ROI-LDA modellene, mens for LDA modellene lå disse parameterne mellom 0,84 – 0,94, se Tabell A-1 og Tabell B-1. Spesifisiteten og sensitiviteten varierte noe mer for ROI-LDA klassifiseringen, enn de gjorde for LDA klassifiseringen. Dette betyr at det var større variasjon i hvor godt ROI-LDA klassifiserte gruppene svulst og ikke-svulst når DCE bildeserien ble inkludert, men verdiene var ikke signifikant forskjellig fra LDA klassifiseringen.

AUC verdiene for LDA modellene med og uten en ROI masken var henholdsvis 0,76 – 0,81 og 0,84 – 0,87. Det vises i Figur 6-2 at AUC verdiene for ROI-LDA klassifisering var signifikant lavere enn for LDA klassifisering.



Figur 6-2: Multiple comparison test (MATLAB®) for AUC verdiene for ROI-LDA (ROI = 1) og for LDA (ROI = 0) klassifiseringer. Figuren viser at det var signifikant forskjell på de to modellenes AUC verdi.

Som vist i Vedlegg F – H hadde ingen av preprosesseringsteknikkene Median filter, Savitzky-Golay filter og CLAHE en signifikant effekt på ytelsen til ROI-LDA modellene. For å kunne sammenligne resultatene i denne oppgaven med den beste LDA modellen presentert i Torheim et al. [18] brukes autoskalering som preprosessering av T_1-vektede, T_2-vektede og DCE-bildene videre.

For en grundigere vurdering om ROI-LDA klassifiseringen forbedret svulstinntegningen for hver enkelt pasient, ble den beste modellen med LDA klassifiseringen sammenlignet med tilsvarende ROI-LDA modell. Ved å sammenligne ytelsen for LDA og ROI-LDA (uthevet skrift) modellene vist i Tabell 6-1, vises de i Figur 6-3 at parameterne DSC, Spes og Kappa er signifikant høyere for ROI-LDA modellen. Det er ikke signifikant forskjell på sensitiviteten til klassifiseringen for bilder med og uten ROI masken. ROC kurven for den beste ROI-LDA og LDA modellen vises i Figur 6-4. ROI-LDA modellen hadde en lavere AUC verdi (0,81) sammenlignet med LDA modellen, som hadde en AUC verdi på 0,87.

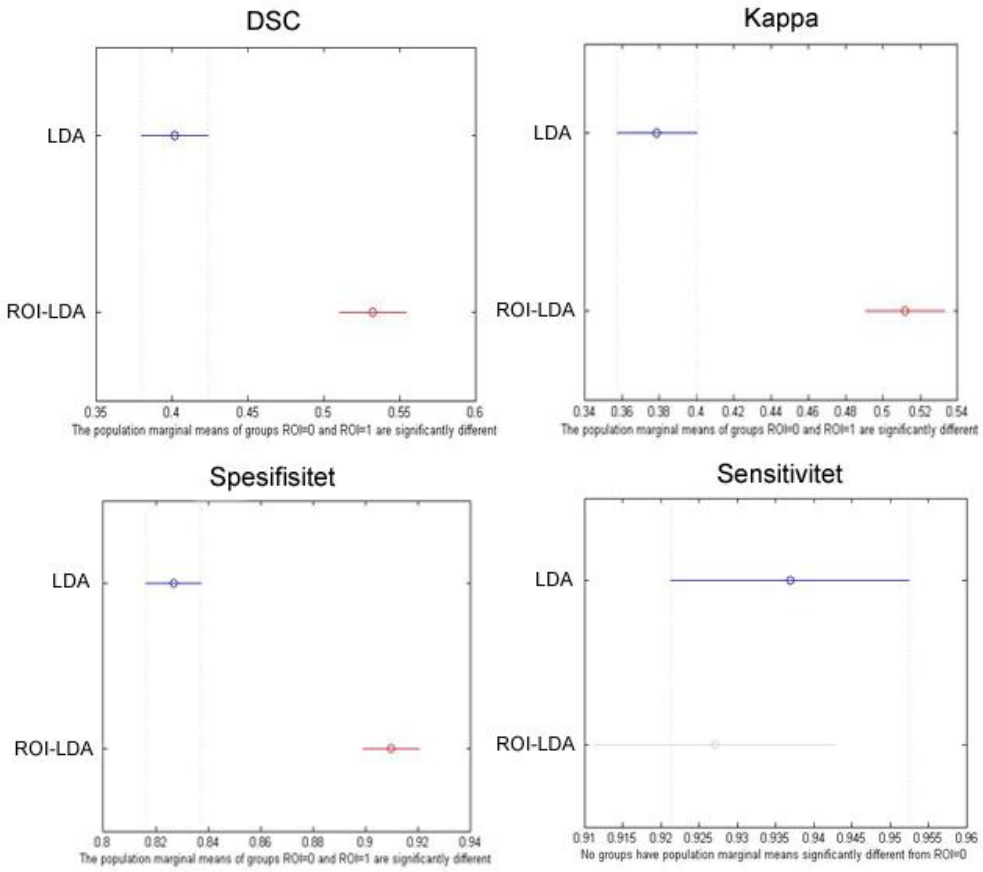
LDA:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
T1w + DCE	Autoskalert	0	0,42	0,39	0,94	0,82	0,87

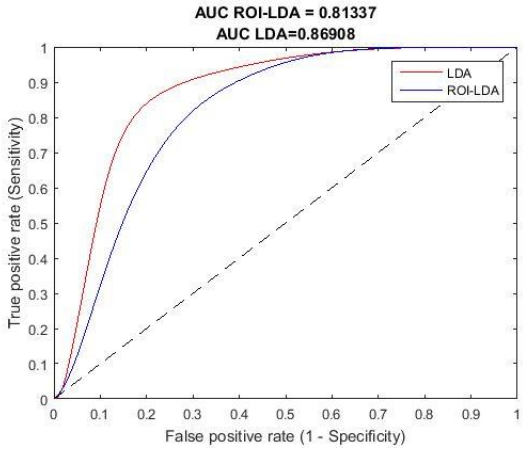
ROI-LDA:

DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,53	0,51	0,93	0,91	0,81
T1w + DCE	**Autoskalert**	**T1w: 8** **DCE: 0**	**0,53**	**0,51**	**0,93**	**0,91**	**0,81**

Tabell 6-1: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for den beste LDA modellen for bilder med og uten ROI (Region of Interest) masken. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

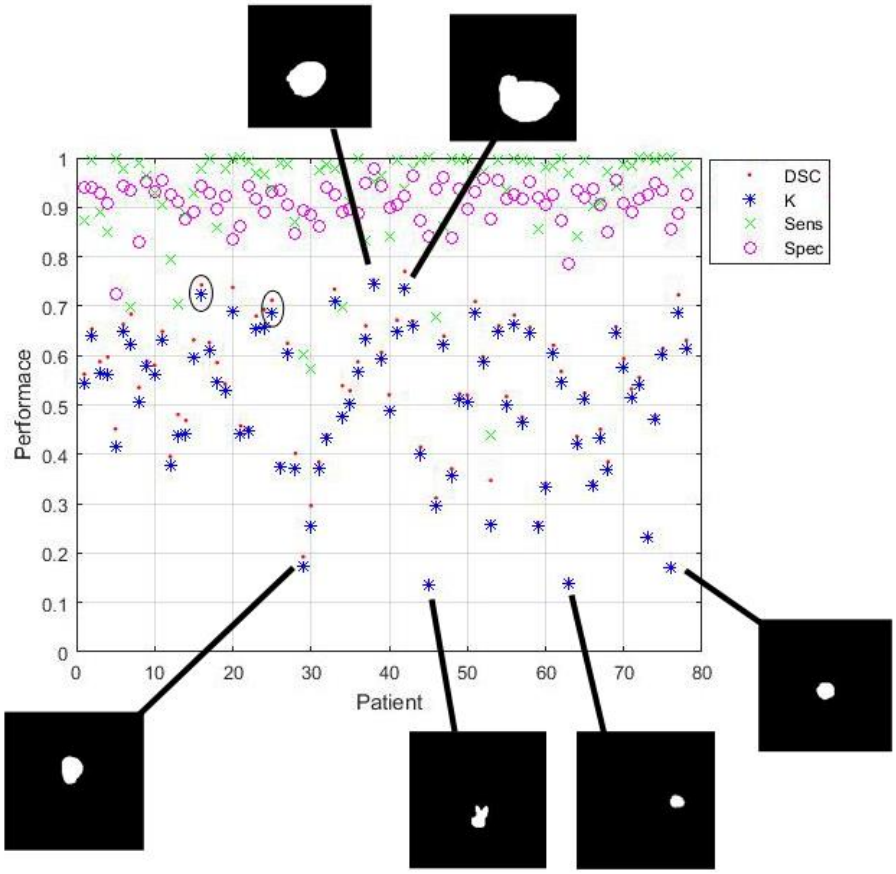


Figur 6-3: Multiple comparison test (MATLAB®) for de ulike parameterne DSC (Dice similarity coefficient), K(Kappa statistikk), Spes (spesifisitet) og Sens (sensitivitet) for den beste LDA modellen og for den beste ROI-LDA modellen (uthevet skrift), vist i Tabell 6-1. Figuren viser at parameterne DSC, Spes og Kappa er signifikant høyere for ROI-LDA modellen. Det er ikke signifikant forskjell på sensitiviteten til klassifiseringen for bilder med og uten ROI masken.



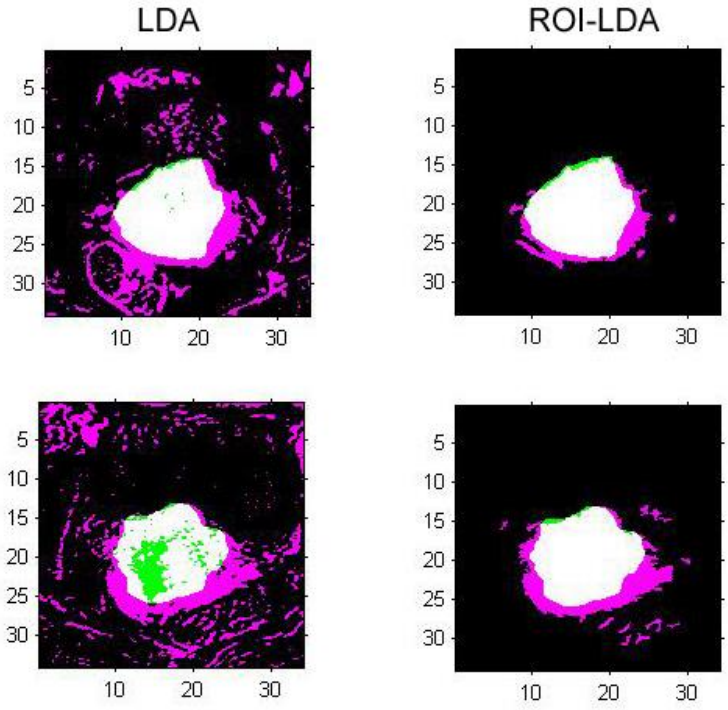
Figur 6-4: ROC kurven for den beste ROI-LDA (uthevet skrift) og LDA modellen, vist i Tabell 6-1. Modellene inkluderer T_1-vektede bilder med 8 naboer og DCE bilder med 0 naboer, hvor alle bildene er autoskalert.

Selv om ROI-LDA modellen utførte klassifiseringen bedre enn LDA modellen var det fortsatt variasjoner i klassifiseringsytelsen mellom pasientene for ROI-LDA modellen. Dette er illustrert i Figur 6-5. Den gjennomsnittlige DSC verdien, for ROI-LDA modellen, for hver pasient varierte fra 0,14 til 0,77. De fleste pasientene (97%) hadde en sensitivitet over 0,6, og 87% av pasientene hadde en DSC og Kappa verdi mellom 0,3-0,7. Figuren viser at Kappa verdiene er lave for de små svulstene.



Figur 6-5: Spredningsplot for den gjennomsnittlige DSC, Kappa, Sens og Spes for hver pasient fra den beste ROI-LDA modellen, se Tabell 6-1 i uthevet skrift og Figur 6-4 for ROC kurven. De laveste DSC og Kappa verdiene representerer alle små svulster, mens de høyeste DSC og Kappa verdiene representerer større svulster, som vist med radiologens maske. Klassifiseringen til pasient 16 og pasient 25 er merket med en sirkel og vist i Figur 6-6.

To av de best klassifiserte svulstene vises i Figur 6-6 og merket med en sirkel i Figur 6-5. Figur 6-6 viser hvilke voksler som ble klassifisert riktig og feil i forhold til radiologen, og var grunnlaget for beregningene av DSC, Kappa, Sensitiviteten og Spesifisiteten. Legg merke til at ROI-LDA klassifiseringen hadde feilklassifiserte voksler i kanten av svulsten (rosa og grønne områder), hvor det er forventet høy usikkerhet. Det vises også tydelig at ROI masken forbedrer vokselklassifiseringen.



Figur 6-6: Viser hvor godt LDA klassifiseringen uten ROI (LDA) og LDA klassifiseringen med ROI (ROI-LDA) overlapper med radiologens maske for to bildesnitt. De rosa områdene indikerer FP (Falsk positiv), det modellene feilaktig klassifiserer som svulst. De grønne områdene indikerer FN (Falsk negativ), det modellene feilaktig klassifiserer som friskt vev. De svarte områdene indikerer SN (Sann negativ), de områdene som modellene riktig klassifiserer som frisk vev, og de hvite områdene indikerer SP (Sann positiv), de områdene som modellene riktig klassifiserer som svulst. Forskjellen på DSC verdien for de to ulike klassifiseringsmetodene er 0,18 for den første raden og 0,20 for den andre raden. Se Figur 6-5 for klassifiseringsresultatene til pasient 16 (øverste rad) og pasient 25 (nederste rad) for ROI-LDA.

6.1.2 Klassifisering med QDA

ROI postprosessering ble også testet med QDA klassifiseringen (ROI-QDA) og sammenlignet med QDA modellen uten ROI masken (QDA), for å teste om inntegningen av ROI masken ga bedre ytelsesverdier for QDA klassifiseringen. Videre ble ROI-QDA modellen sammenlignet med ROI-LDA modellen for å undersøke om det var signifikant forskjell på ytelsen til den lineære og ikke-lineære klassifiseringsmodellen.

Fra tidligere undersøkelser er det kun utført én analyse for QDA klassifikatoren. Denne modellen presenteres i Tabell 6-2, sammen med tilsvarende modell for ROI-QDA og ROI-LDA. Alle disse modellene inkluderte T_1-vektede (T1w) bilder og T_2-vektede (T2w) bilder med 8 naboer, og DCE bilder med 0 naboer. Øvrige modeller for ROI-QDA vises i Vedlegg C: ROI-QDA klassifisering, Tabell C-1.

QDA:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,28	0,26	0,99	0,69	0,90

QDA-ROI:

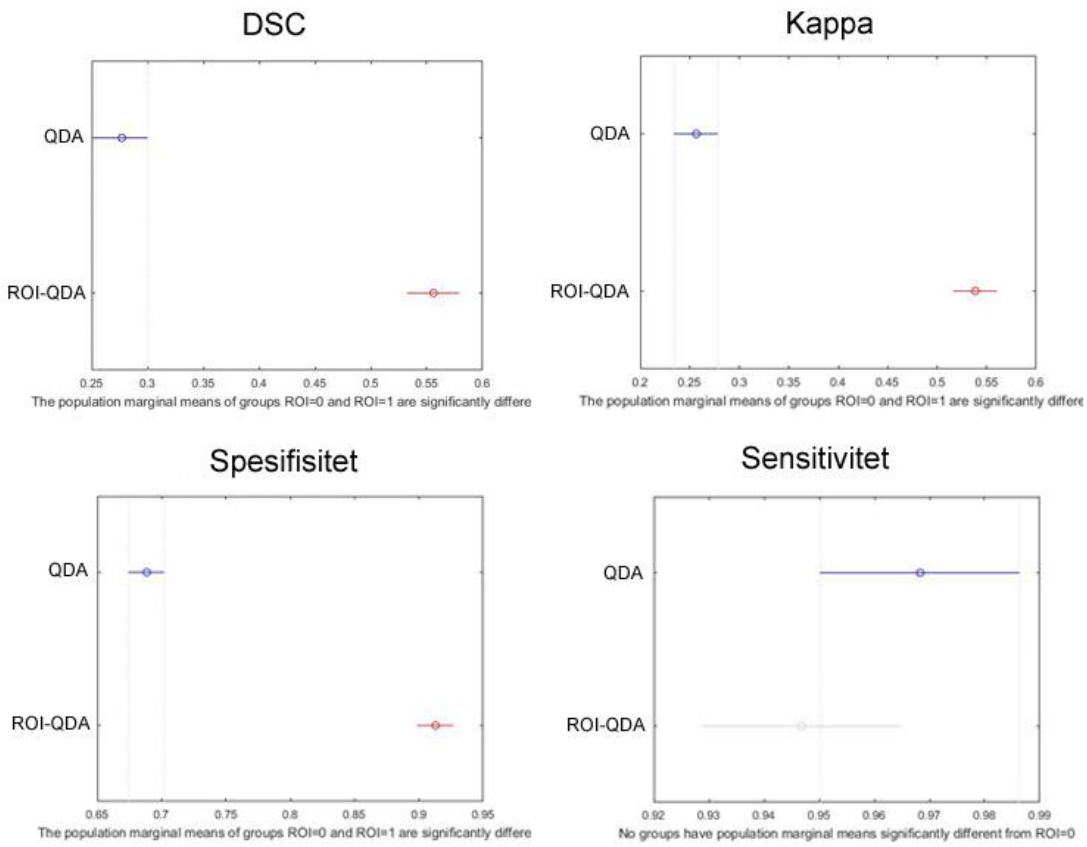
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,56	0,54	0,95	0,91	0,88

ROI-LDA:

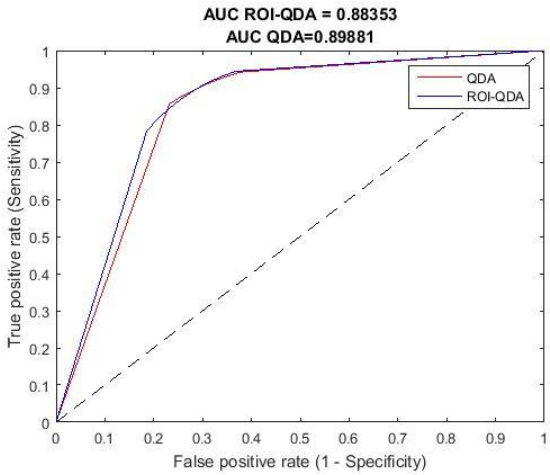
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,53	0,51	0,93	0,91	0,81

Tabell 6-2: DSC (Dice similarity coefficient), K (Kappa statistikk, Sens (sensitivitet), Spes (spesifisitet),) og AUC (arealet under kurven) for en av de beste ROI-LDA modellene og tilsvarende modeller for QDA uten ROI (Region of Interest) masken og ROI-QDA. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

Ved å sammenligne ytelsen for QDA og ROI-QDA modellene i Tabell 6-2, vises det i Figur 6-7 at DSC, *K* og Spes var signifikant høyere for ROI-QDA. Dette indikerer at bildene med ROI masken ga en bedre klassifisering av svulsten. Det er ingen signifikant forskjell mellom sensitiviteten til ROI-QDA og QDA modellen. AUC verdien til de to modellene var tilnærmet like og ROC kurven for ROI-QDA og QDA modellene vises i Figur 6-8.

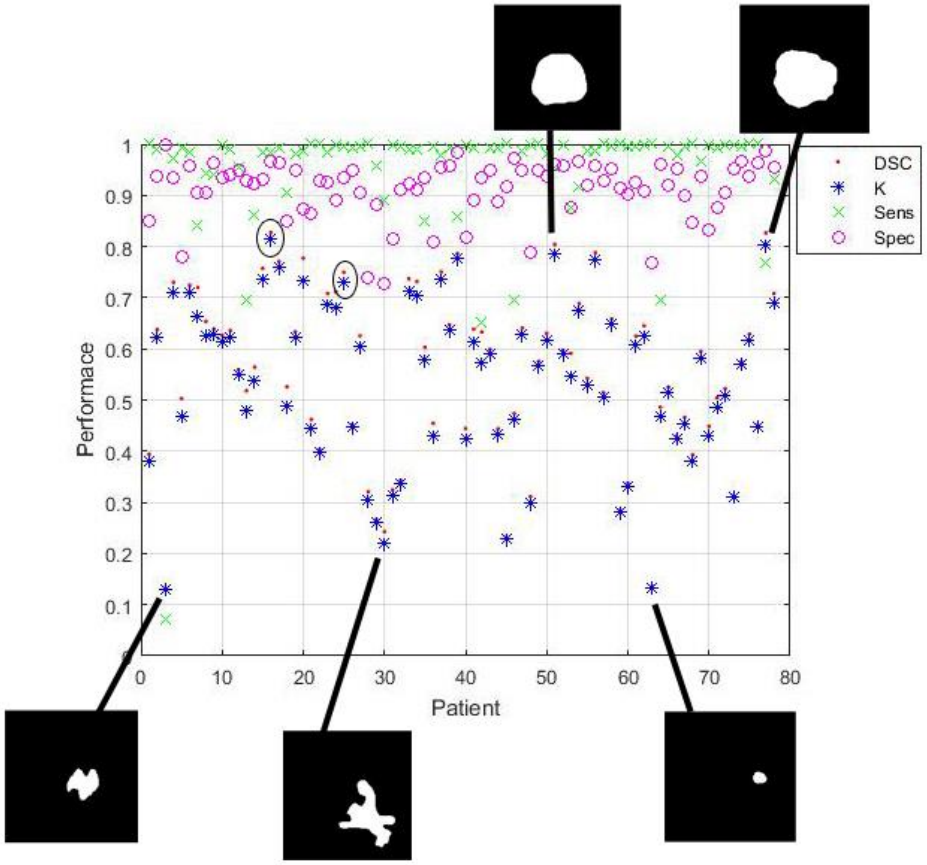


Figur 6-7: Multiple comparison test (MATLAB®) for de ulike parameterne DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet) og Spes (spesifisitet) for QDA og ROI-QDA modellene, presentert i Tabell 6-2. Figuren viser at parameterne DSC, Kappa og Spes var signifikant høyere for bildene med ROI masken. Det var ingen signifikant forskjell mellom sensitiviteten for bildene med og uten ROI masken.



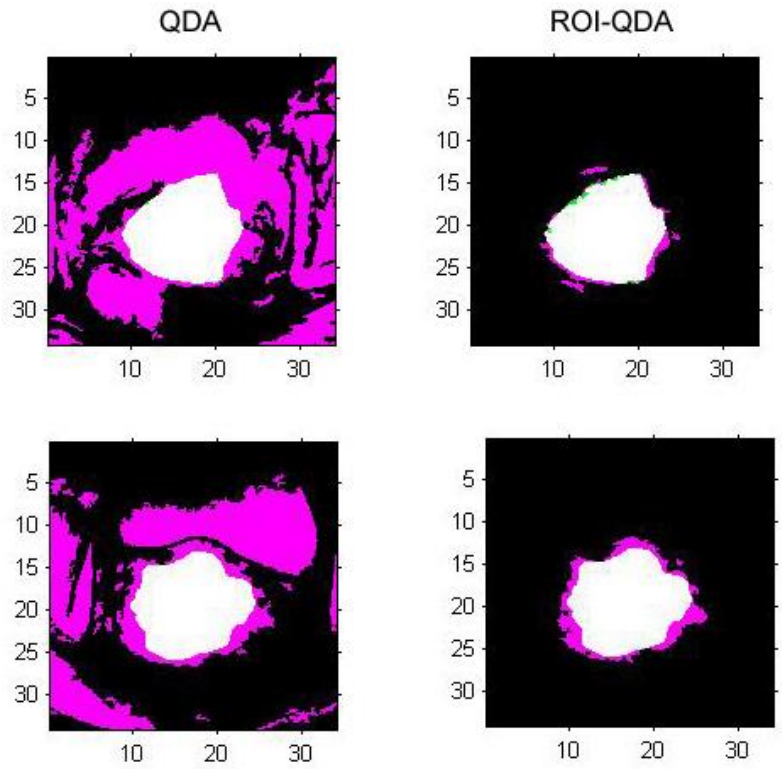
Figur 6-8: ROC kurvene for QDA modellene for bildene med eller uten ROI (Region of Interest) masken, presentert i Tabell 6-2. Modellene inkluderer T_1-vektede og T_2-vektede bilder med 8 naboer, og DCE bilder med 0 naboer.

Selv om ROI-QDA modellen utførte vokselklassifiseringen bedre enn QDA modellen var det fortsatt en del variasjoner mellom pasientene for ROI-QDA modellen. Dette illustreres i Figur 6-9. Figuren viser spredningsplottet for ROI-QDA modellen presentert i Tabell 6-2. Eksempelvis varierte den gjennomsnittlige DSC verdien for hver pasient fra 0,13 til 0,83. De fleste pasientene for ROI-QDA modellen (92%) hadde en sensitivitet over 0,8 og 79% av pasientene hadde en DSC og Kappa verdi mellom 0,4-0,8.



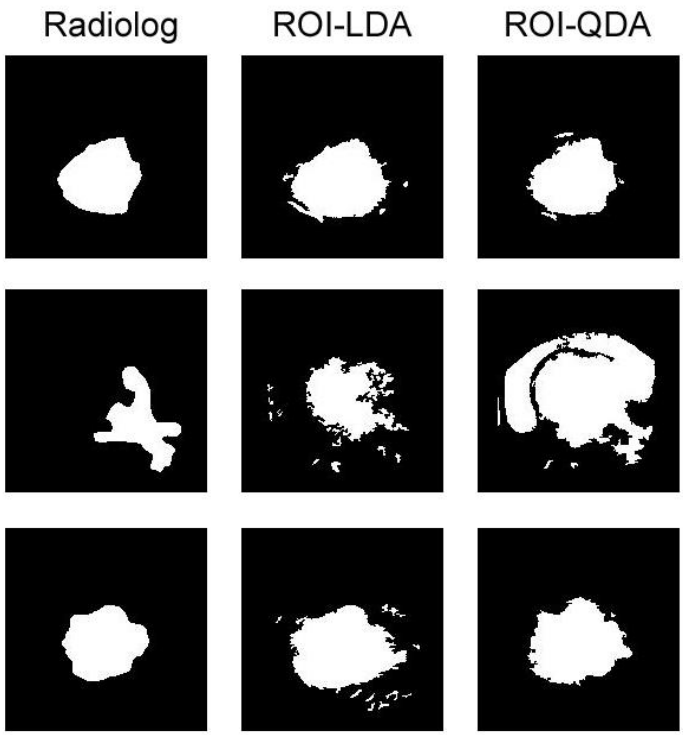
Figur 6-9: Spredningsplot for den gjennomsnittlige DSC, Kappa, Sens og Spes for hver pasient for ROI-QDA modellen, presentert i Tabell 6-2. Se Figur 6-8 for ROC kurven. Radiologens maske vises for svulstene som ROI-QDA modellen predikerte med lavest og høyest Kappa verdi. Klassifiseringen til pasient 16 og pasient 25 (merket med sirkel) er vist i Figur 6-10.

To av de best klassifiserte svulstene for ROI-QDA vises i Figur 6-10 og merket med en sirkel i Figur 6-9. Figur 6-10 viser hvilke voksler som ble klassifisert riktig og feil i forhold til radiologen, og var grunnlaget for beregningene av DSC, Kappa, Sensitiviteten og Spesifisiteten. I likhet med ROI-LDA klassifiseringen ser man at feilklassifiseringen lå i kantene på svulstene, og figuren viser tydelig at ROI masken forbedret vokselklassifiseringen.

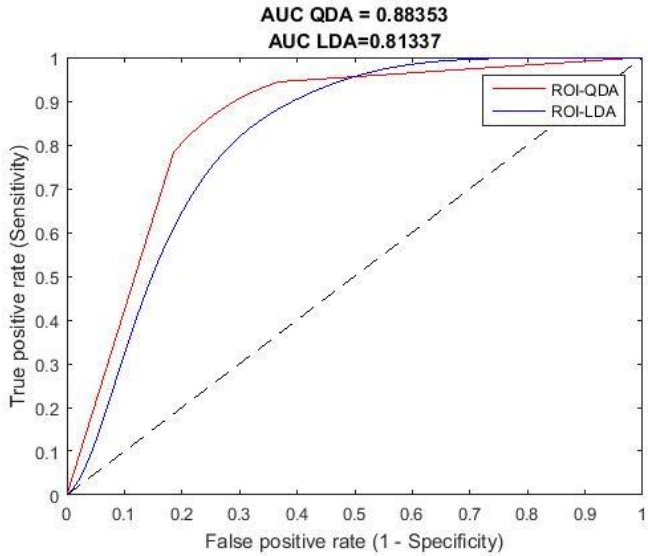


Figur 6-10: Viser hvor godt QDA klassifiseringen uten ROI masken (QDA) og QDA klassifiseringen med ROI masken (ROI-QDA) overlapper med radiologens maske for to bildesnitt. De rosa områdene indikerer FP (Falsk positiv), det QDA modellen feilaktig klassifiserer som svulst. De grønne områdene indikerer FN (Falsk negativ), det QDA modellen feilaktig klassifiserer som friskt vev. De svarte områdene indikerer SN (Sann negativ), de område som QDA modellen riktig klassifiserer som frisk vev, og de hvite områdene indikerer SP (Sann positiv), de områdene som QDA modellen riktig klassifiserer som svulst. Forskjellen på DSC verdien for de to ulike klassifiseringsmetodene er 0,55 for den første raden og 0,35 for den andre raden Se Figur 6-9 for klassifiseringsresultatene til pasient 16 (øverste rad) og pasient 25 (nederste rad) for ROI-QDA (høyre kolonne).

Videre sammenlignes den beste ROI-LDA modellen med tilsvarende ROI-QDA modell, vist i Tabell 6-2. Ingen av ytelsesparameterne var signifikant forskjellige for de to modellene. Forskjellen mellom modellklassifiseringen for noen svulster vises i Figur 6-11. For den første pasienten (første rad), var modellklassifiseringen høyere for ROI-QDA (DSC = 0,83) enn for ROI-LDA (DSC = 0,74). Den andre raden viser et bildesnitt hvor enigheten mellom radiologen og ROI-QDA modellen var lavere (DSC = 0,24) enn for ROI-LDA modellen (DSC =0,30). Den tredje raden viser et bildesnitt hvor modellklassifiseringen var omtrent like god for ROI-QDA (DSC = 0,74) som for ROI-LDA (DSC = 0,71). AUC verdien (se Figur 6-12 for ROC kurven) var noe høyere for ROI-QDA modellen enn for ROI-LDA modellen.



Figur 6-11: Valgte snitt fra tre forskjellige pasienter for å illustrere variasjonene i ROI-LDA og ROI- QDA klassifiseringene. Den første kolonnen viser radiologens maske, den andre kolonnen viser LDA klassifiseringen med ROI masken, og den tredje kolonnen viser QDA klassifiseringen med ROI masken. Tre forskjellige pasienter ble valgt, med forskjellig grad av korrekt klassifisering.



Figur 6-12: ROC kurven for ROI-QDA og ROI-LDA modellene, presentert i Tabell 6-2. Modellene inkluderer T_1-vektede og T_2-vektede bilder med 8 naboer, og DCE bilder med 0 naboer, hvor alle bildene er autoskalert.

6.2 Test av ikke-lineære klassifiseringsmetoder ved bruk av ROI bilder

På grunnlag av resultatet over ble kun bildene med ROI masken brukt i videre analyser, og testet med ikke-lineære klassifiseringsmodeller. Hensikten var å finne ut om andre klassifiseringsmetoder ga en bedre klassifisering av svulsten enn ROI-LDA/ROI-QDA metodene. Alle bildene ble i tillegg postprosessert ved bruk av en morfologisk operasjon på de binære bildene for å fjerne små elementer mindre enn 10 voksler.

6.2.1 Random Forest klassifisering

Random Forest metoden ble først undersøkt for å bestemme om det var signifikant forskjell i klassifisering ved valg av 10, 50, 100 eller 200 trær. Deretter sammenlignes den beste ROI-LDA modellen som inkluderte alle bildetypene, med samme modell for *Random Forest.* Dette var også den modellen med de høyeste ytelsesparameterne for *Random Forest* klassifiseringen.

Modellene med høyest ytelsesverdier for *Random Forest* klassifiseringsmetoden med 10, 50, 100 og 200 trær presenteres i Tabell 6-3. Disse modellene inkluderte T_1-vektede (T1w) bilder med 8 naboer, T_2-vektede (T2w) bilder med 8 naboer og DCE bilder med 0 naboer. Øvrige modeller vises i Vedlegg E: ROI-*Random Forest* klassifisering, i Tabell E-1.

10 trær:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,55	0,53	0,51	0,99	0,88

50 trær:

DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,58	0,56	0,55	0,99	0,91

100 trær:

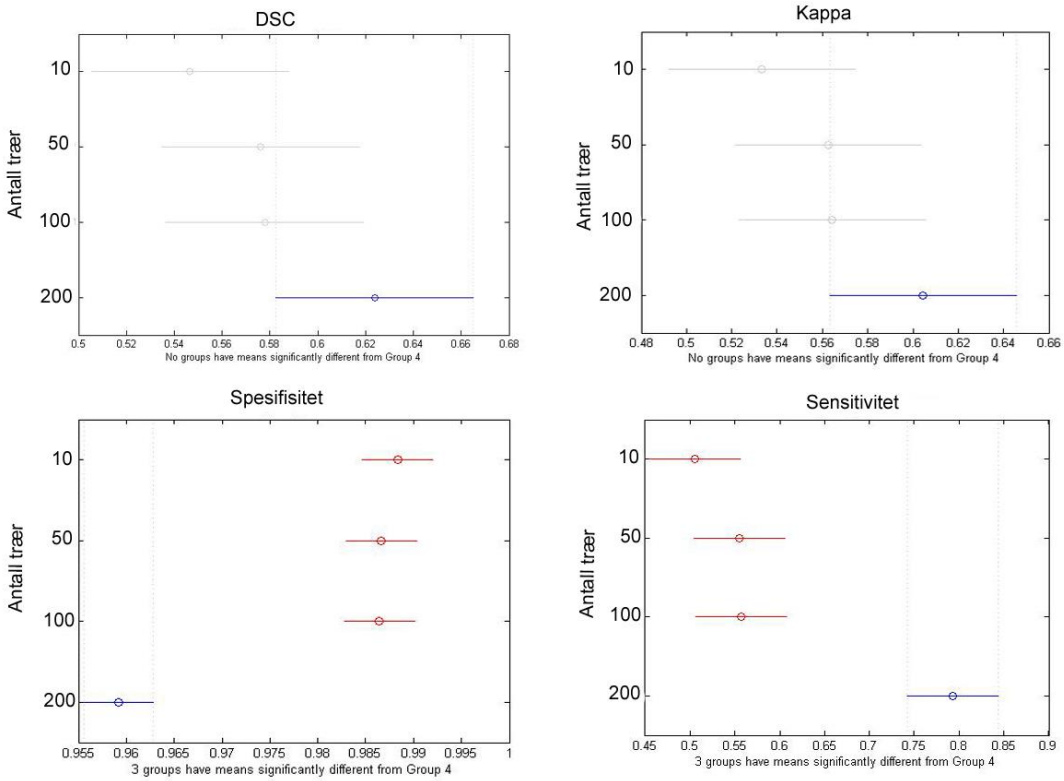
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,58	0,56	0,56	0,99	0,91

200 trær:

DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,62	0,60	0,79	0,96	0,74

Tabell 6-3: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for Random Forest modellene med høyest ytelsesparametere for ulike antall trær (10,50,100 og 200) med ROI (Region of Interest) masken. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

Ved å sammenligne DSC, Kappa, Sens og Spes for modellene presentert i Tabell 6-3 vises det i Figur 6-13, at DSC og Kappa verdiene ikke var signifikant forskjellige for noen av modellene. Spesifisiteten var derimot signifikant lavere for modellen med 200 trær i forhold til modellene med 10, 50 og 100 trær. Figur 6-13 viser også at modellene med 10, 50 og 100 trær hadde signifikant høyere sensitivitet enn modellen med 200 trær. Dette betyr at modellene med 10, 50 og 100 trær ligner mer på hverandre, enn modellen med 200 trær.



Figur 6-13: Multiple comparison test (MATLAB®) for de ulike parameterne DSC (Dice similarity coefficient), K (Kappa statistikk), Spes (spesifisitet) og Sens (sensitivitet) for Random Forest modellene med 10, 50, 100 og 200 trær, presentert i Tabell 6-3. Figuren viser at parameterne DSC og Kappa ikke var signifikant forskjellig for noen av modellene. Spesifisiteten var signifikant bedre for modellene med 10, 50 og 100 trær i forhold til modellen med 200 trær, mens sensitiviteten var signifikant bedre for modellen med 200 trær i forhold til modellen med 10, 50 og 100 trær.

For å vurdere om *Random Forest* klassifiseringen klarte å identifisere svulstene mer nøyaktig enn ROI-LDA klassifiseringen, velges én av *Random Forest* modellene. Siden ingen av *Random Forest* modellene var signifikant forskjellige fra hverandre velges den med høyest AUC verdi. Modellene med 100 og 50 trær hadde lik AUC verdi. Derfor velges modellen med 50 trær fordi den brukte 12 timer kortere tid på å trene modellen og var i tillegg en enklere modell. For å sammenligne de to klassifiseringsmetodene ble den beste modellen for ROI-LDA brukt, sammen med tilsvarende modell for *Random Forest* med 50 trær som vist i Tabell 6-4. Disse modellene inkluderte T_1-vektede (T1w) og T_2-vektede (T2w) bilder med 8 naboer, og DCE bilder med 0 naboer.

ROI-LDA:

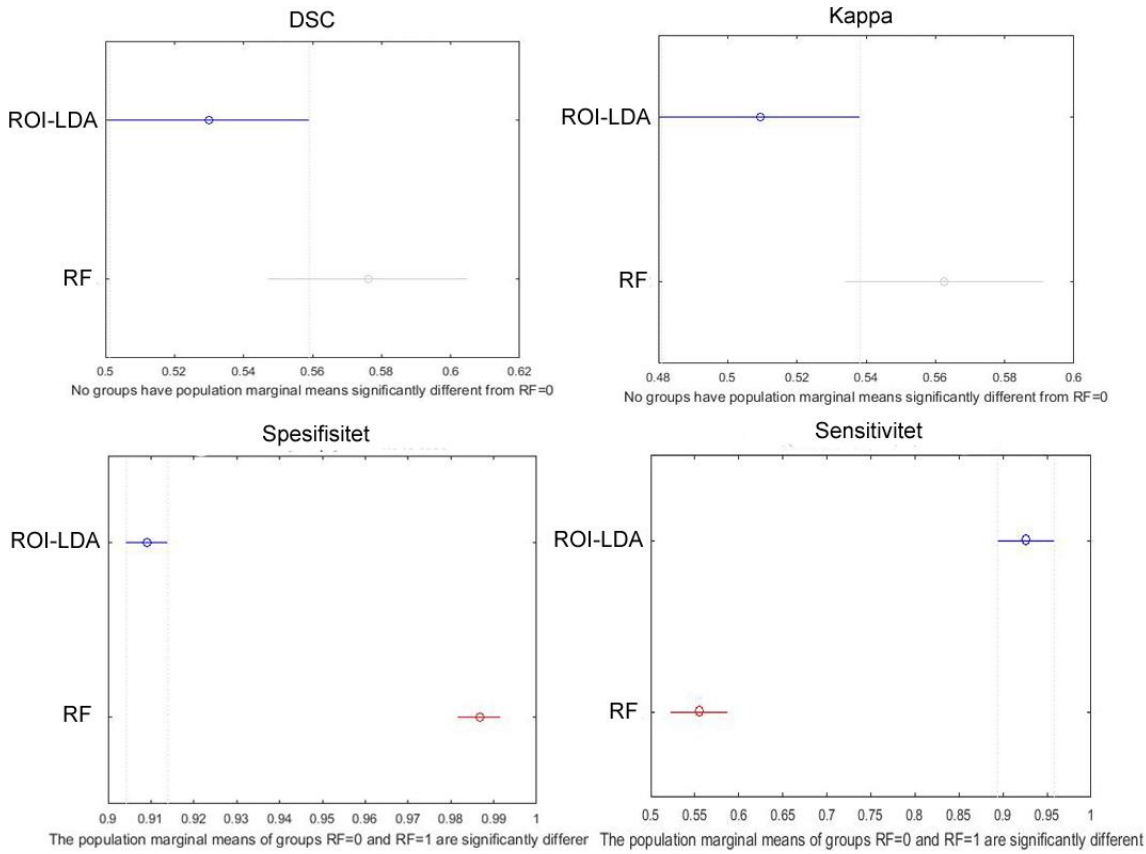
Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,53	0,51	0,93	0,91	0,81

***Random Forest* med 50 trær:**

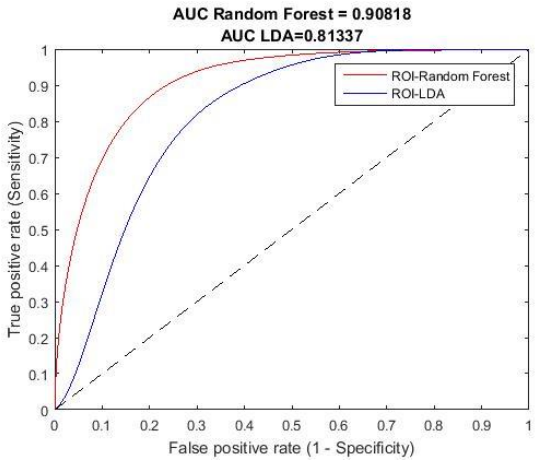
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,58	0,56	0,55	0,99	0,91

Tabell 6-4: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for den beste ROI-LDA modellen og tilsvarende modell for Random Forest med 50 trær og ROI (region of Interest) maske. Disse modellene ble basert på T_2-vektede (T1w), T_1-vektede (T2w) og DCE-MRI bilder. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

For å undersøke om *Random Forest* klassifiseringsmodellen hadde bedre ytelse enn ROI-LDA klassifiseringsmodellen sammenlignes DSC, Kappa, Sensitiviteten og Spesifisiteten til modellene i Tabell 6-4, vist i Figur 6-14. Denne figuren viser at DSC og Kappa verdiene til *Random Forest* modellen med 50 trær ikke var signifikant forskjellig fra ROI-LDA klassifiseringsmetoden. Spesifisiteten var derimot signifikant høyere for *Random Forest* enn for ROI-LDA, mens sensitiviteten var signifikant høyere for ROI-LDA klassifiseringen enn for *Random Forest* klassifiseringen. Dette betyr at *Random Forest* algoritmen klassifiserte en høyere andel voksler som svulst, enn det ROI-LDA klassifiseringsmetoden gjorde. I tillegg brukte *Random Forest* algoritmen i overkant av 10 timer lenger tid på å trene modellen enn ROI-LDA algoritmen. AUC verdien var høyere for *Random Forest* klassifiseringsmodellen med 50 trær enn for ROI-LDA klassifiseringsmodellen, og ROC kurvene vises i Figur 6-15.

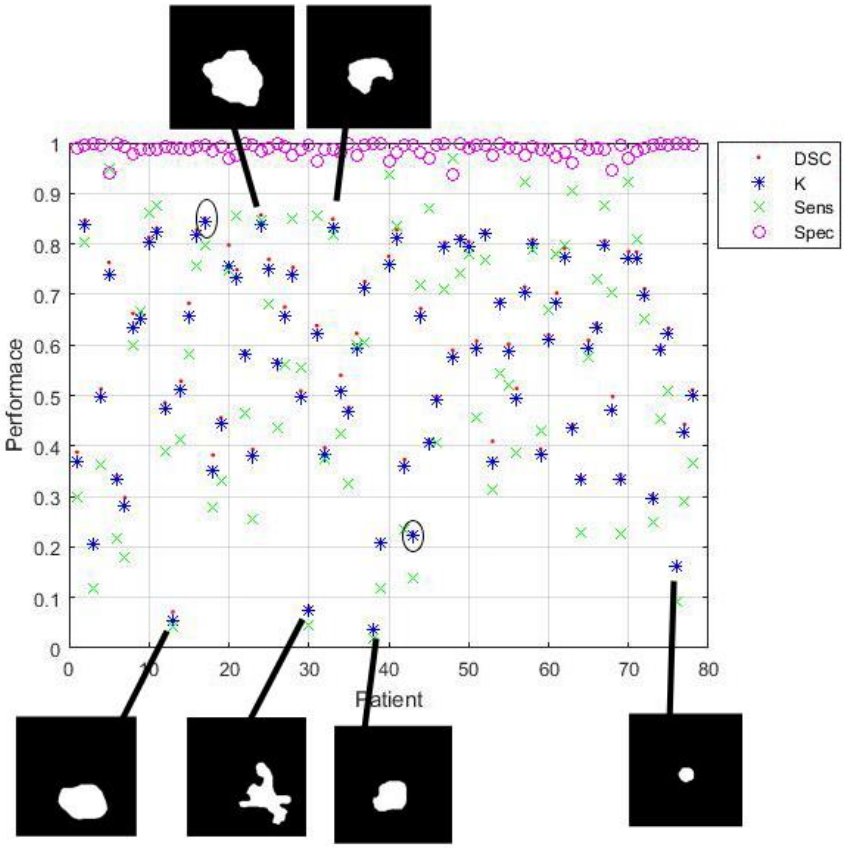


Figur 6-14: Multiple comparison test (MATLAB®) for de ulike parameterne DSC (Dice similarity coefficient), K (Kappa statistikk), Spes (spesifisitet) og Sens (sensitivitet) for den beste ROI-LDA modellen og tilsvarende modell for Random Forest (RF) med 50 trær, presentert i Tabell 6-4. Figuren viser at parameteren DSC og Kappa ikke var signifikant forskjellig fra hverandre. Spesifisiteten var signifikant høyere for Random Forest modellen med 50 trær enn for LDA modellen, mens sensitiviteten var signifikant høyere for LDA modellen enn for Random Forest modellen.



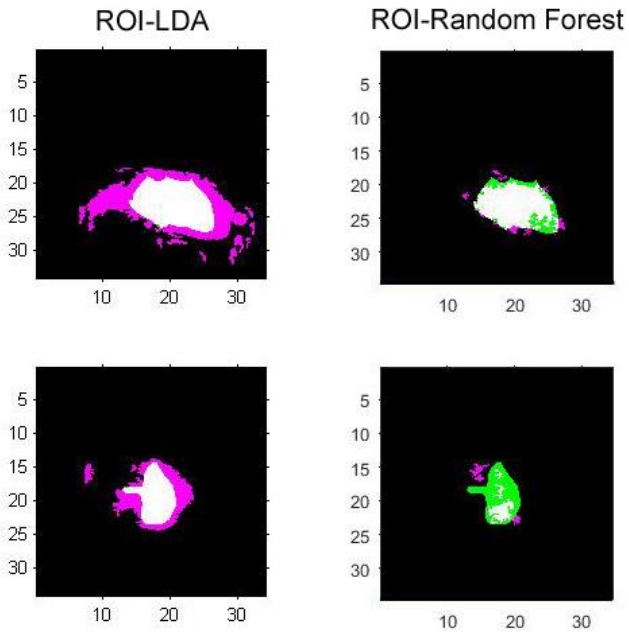
Figur 6-15: ROC kurven for den beste ROI-LDA modellen og tilsvarende modell for Random Forest med 50 trær, presentert i Tabell 6-4. Modellene inkluderte T_1-vektede og T_2-vektede bilder med 8 naboer, og DCE bilder med 0 naboer, hvor alle bildene er autoskalert.

Figur 6-16 visualiserer ytelsesvariasjonen mellom pasientene for *Random Forest* klassifiseringsmodellen med 50 trær presentert i Tabell 6-4. Den gjennomsnittlige DSC verdien for hver pasient varierte mellom 0,04 – 0,86. De fleste pasientene (72%) hadde en sensitivitet over 0,4 og 68% av pasientene hadde en DSC og Kappa verdi mellom 0,5 og 0,9.



Figur 6-16: Spredningsplot for den gjennomsnittlige DSC, Kappa, Sens og Spes for hver pasient for Random Forest modellen med 50 trær, se Tabell 6-4 og Figur 6-15 for ROC kurven. Radiologens maske vises for svulstene som Random Forest modellen predikerte med lavest og høyest Kappa verdi. Klassifiseringen til pasient 17 og pasient 43 (merket med sirkel) vises i Figur 6-10Figur 6-17.

For å illustrere variasjonene mellom de to klassifiseringsmetodene ble det valgt ut to svulster, vist i Figur 6-17, som merket med en sirkel i Figur 6-16. Figur 6-17 viser at enkelte svulster klassifiseres bedre med *Random Forest* metoden, mens andre svulster klassifiseres bedre med ROI-LDA metoden. Den viser også hvilke voksler som ble klassifisert riktig og feil i forhold til radiologen. Den øverste svulsten hadde en høyere DSC verdi (0,85) for *Random Forest* klassifiseringsmodellen, enn for ROI-LDA klassifiseringen (DSC = 0,63). For den nederste svulsten hadde ROI-LDA klassifiseringen en høyere DSC (0,70) enn *Random Forest* klassifiseringen (DSC = 0,23). Figur 6-17 viser også tydelig at sensitiviteten var høyere for ROI-LDA modellen (få FN voksler, grønne felter), mens spesifisiteten var høyere for *Random Forest* modellen (få FP voksler, rosa felter).



Figur 6-17: Viser hvor godt ROI-LDA (kolonne til venstre) og Random Forest klassifiseringen med 50 trær (kolonne til høyre) overlapper med radiologens maske for to bildesnitt. De rosa områdene indikerer FP (Falsk positiv), det modellen feilaktig klassifiserer som svulst. De grønne områdene indikerer FN (Falsk negativ), det modellen feilaktig klassifiserer som friskt vev. De svarte områdene indikerer SN (Sann negativ), de områdene som modellen riktig klassifiserer som frisk vev, og de hvite områdene indikerer SP (Sann positiv), de områdene som modellen riktig klassifiserer som svulst. Forskjellen på DSC verdien for de to ulike klassifiseringsmetodene er 0,22 for den første raden og 0,47 for den andre raden. Se Figur 6-16 for klassifiseringsresultatene til pasient 17 (øverste rad) og pasient 43 (nederste rad) for ROI-Random Forest (høyre kolonne).

6.2.2 AdaBoost klassifiseringsmetode

AdaBoost klassifiseringsmetode ble testet for ulike parametere, blant annet for maksimum antall splitt og minimum bladstørrelser. Hovedproblemet med denne klassifiseringsmodellen var at alle vokslene ble klassifisert i samme klasse. Uansett valg av parametere ble alle ytelsesparameterne for klassifiseringen tilnærmet lik null, med unntak av spesifisiteten som ble tilnærmet lik en. Dette tilsier at denne klassifiseringsmetoden klassifiserte tilnærmet alle vokslene som ikke-svulst.

6.2.3 SVM klassifiseringsmetoden

SVM klassifiseringsmetoden ble ikke tilstrekkelig testet på grunn av tidsbegrensinger, altså det tok for lang tid å trene modellen. En av ulempene med denne algoritmen var at maksimalt antall støttevektorer ikke kunne bestemmes manuelt, og derfor ble det vanskelig å få ned tiden det tok å trene modellen.

6.2.4 k Nærmeste Nabo (kNN)

kNN klassifiseringsmetoden ble først undersøkt for å sjekke om det var signifikant forskjell på valg av én, to eller tre antall nabovoksler i algoritmen. Deretter sammenlignes den beste ROI-LDA modellen, med tilsvarende modell for kNN klassifiseringen. Den beste modellen for ROI-LDA var også de modellene for kNN metoden som hadde de høyeste ytelsesparameterne.

Modellene med høyest ytelsesverdier for kNN klassifiseringsmetoden for én, to og tre nabovoksler presenteres i Tabell 6-5. Disse modellene inkluderte T_1-vektede (T1w) og T_2-vektede (T2w) bilder med 8 naboer, og DCE bilder med 0 naboer. Øvrige modeller vises i Vedlegg D: ROI-kNN klassifisering, Tabell D-1.

kNN én nabo:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,62	0,60	0,79	0,96	0,74

kNN to naboer:

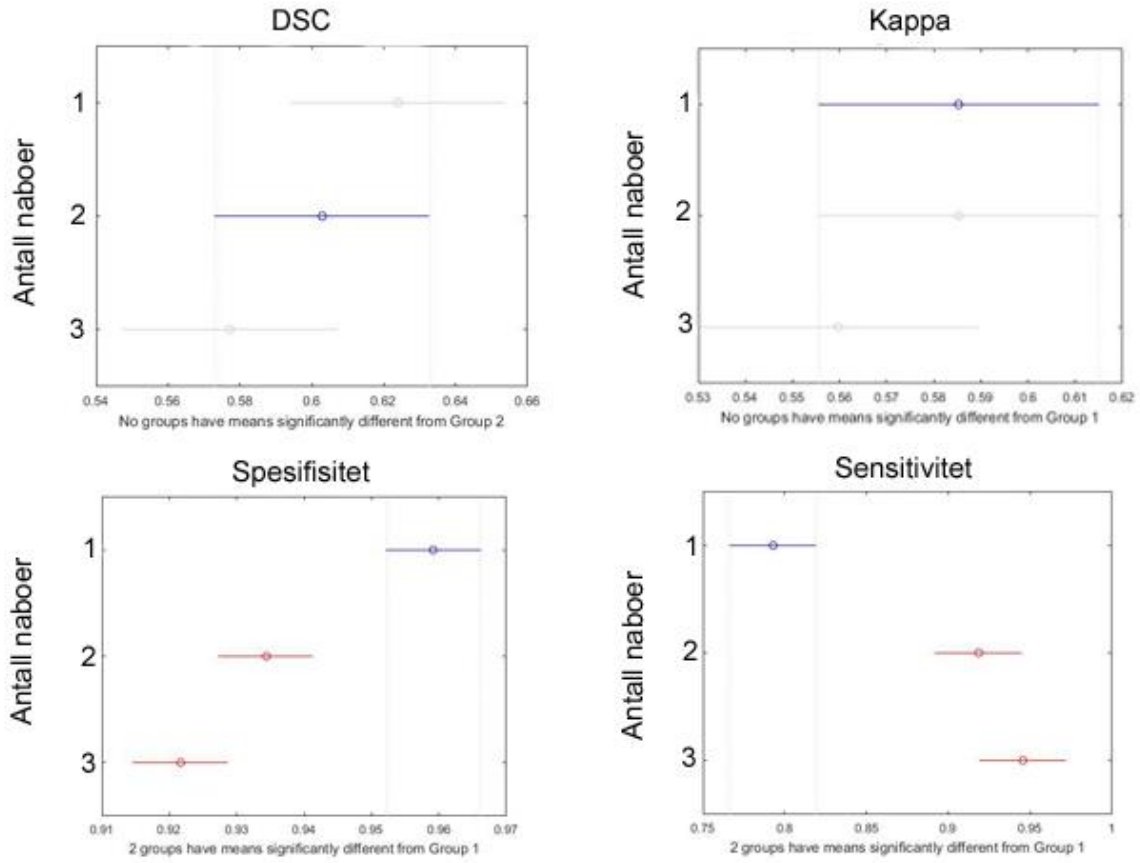
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,60	0,59	0,92	0,93	0,81

kNN tre naboer:

DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,58	0,56	0,95	0,92	0,83

Tabell 6-5: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for kNN modellene med høyest ytelsesparameter med en, to eller tre antall naboer og ROI (region of Interest) postprosessering. Disse modellene ble basert på T_2-vektede, T_1-vektede og DCE bilde. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

Ved å sammenligne DSC, Kappa, Sens og Spes for modellene presentert i Tabell 6-5 vises det i Figur 6-18, at DSC og Kappa verdiene ikke var signifikant forskjellige. Spesifisiteten var derimot signifikant høyere for modellen med én nabovoksel i forhold til modellene med to og tre nabovoksler. Figur 6-18 viser også at modellene med to og tre nabovoksler hadde signifikant høyere sensitivitet enn modellen med én nabovoksel, og klassifiserte dermed svulst vokslene bedre. Dette betyr at modellene med to og tre nabovoksler ligner mer på hverandre, enn modellen med én nabovoksel.



Figur 6-18: Multiple comparison test (MATLAB®) for de ulike parameterne DSC (Dice similarity coefficient), K (Kappa statistikk), Spes (spesifisitet) og Sens (sensitivitet) for kNN modellen med høyest ytelsesparameter med en, to og tre nabovoksler, presentert i Tabell 6-5. Figuren viser at parameterne DSC og Kappa ikke var signifikant forskjellig for noen av modellene. Spesifisiteten var signifikant bedre for modellen med én nabo, mens sensitiviteten var signifikant bedre for to og tre antall naboer.

For å vurdere om kNN klassifiseringen klarte å identifisere svulstene mer nøyaktig enn ROI-LDA klassifiseringen ble kNN modellen med høyest AUC verdi valgt, siden ingen av kNN modellene var signifikant forskjellige fra hverandre. Den beste modellen for ROI-LDA klassifiseringen og tilsvarende modell for kNN klassifiseringen med tre nabovoksler vises i Tabell 6-6. Disse modellene inkluderte T_1-vektede og T_2-vektede bilder med 8 naboer, og DCE bilder med 0 naboer.

ROI-LDA:

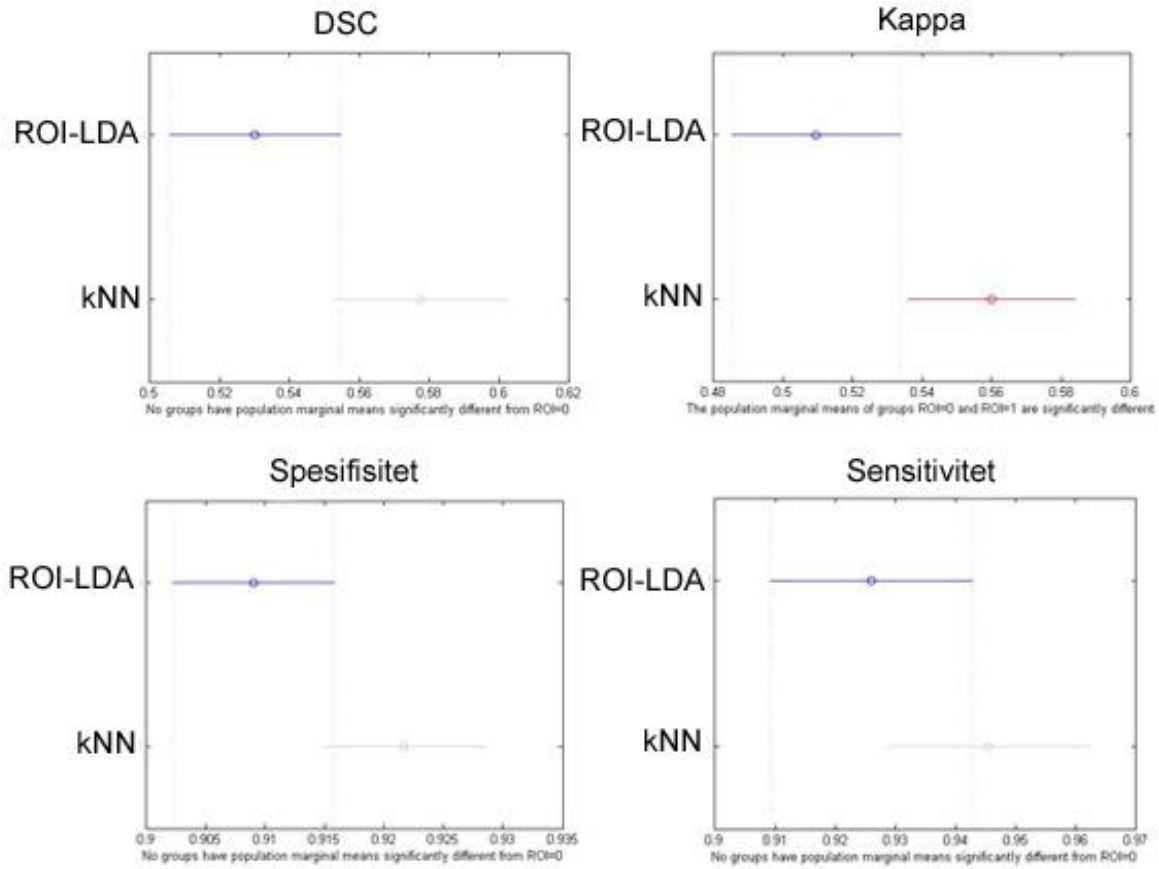
Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,53	0,51	0,93	0,91	0,81

ROI-kNN med tre nabovoksler:

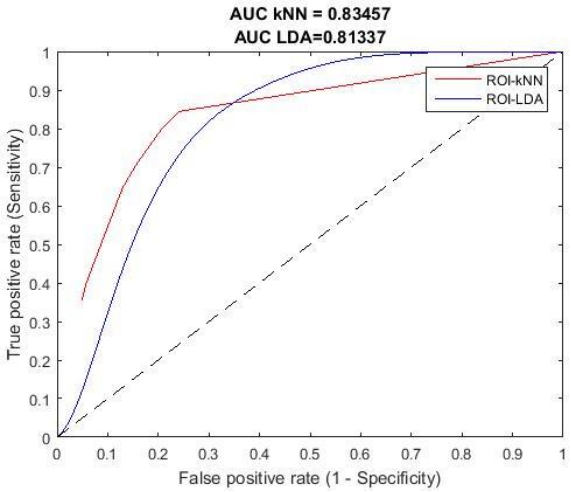
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,58	0,56	0,95	0,92	0,83

Tabell 6-6: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for den beste ROI-LDA modellen og tilsvarende modell for kNN med tre nabovoksler og ROI (region of interest) masken. Modellene ble basert på T_2-vektede, T_1-vektede og DCE-MRI bilder. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

For å undersøke om kNN klassifiseringsmodellen hadde bedre ytelse enn ROI-LDA klassifiseringsmodellen sammenlignes DSC, Kappa, Sensitiviteten og Spesifisiteten til modellene vist i Tabell 6-6, for hver pasient. Figur 6-19 viser at DSC verdiene til kNN modellen med tre nabovoksler var noe høyere, men ikke signifikant forskjellig fra ROI-LDA modellen. Kappa verdien derimot viste seg å være signifikant høyere for kNN klassifiseringsmodellen. Hverken sensitiviteten eller spesifisiteten var signifikant høyere for kNN klassifiseringsmodellen. Det som er verdt å merke seg er at alle ytelsesparameterne var noe høyere for kNN klassifiseringsmodellen enn for ROI-LDA modellen. På den andre siden brukte kNN algoritmen med tre nabovoksler i overkant av 76 timer lengre tid på å trene modellen enn ROI-LDA algoritmen. ROC kurvene for den beste ROI-LDA modellen og tilsvarende modell for kNN med tre nabovoksler vises i Figur 6-20. AUC verdien var høyere for kNN klassifiseringsmodellen med tre nabovoksler enn for ROI-LDA klassifiseringsmodellen.

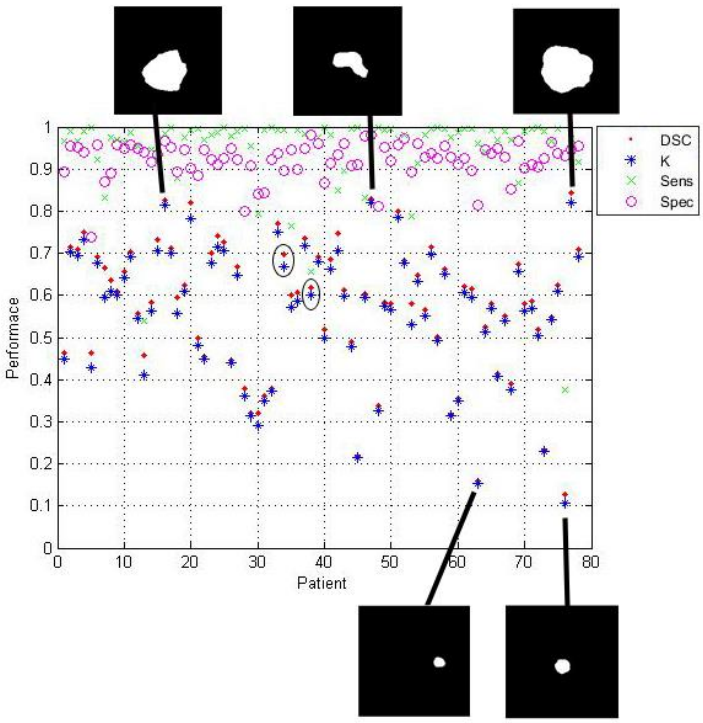


Figur 6-19: Multiple comparison test (MATLAB®) for de ulike parameterne DSC (Dice similarity coefficient), K (Kappa statistikk), Spes (spesifisitet) og Sens (sensitivitet) for den beste ROI-LDA modellen og tilsvarende modell for kNN med tre nabovoksler, vist i Tabell 6-6.. Figuren viser at Kappa verdien var signifikant høyere for kNN modellen med tre naboer. Hverken DSC, spesifisiteten eller sensitiviteten var signifikant forskjellig.



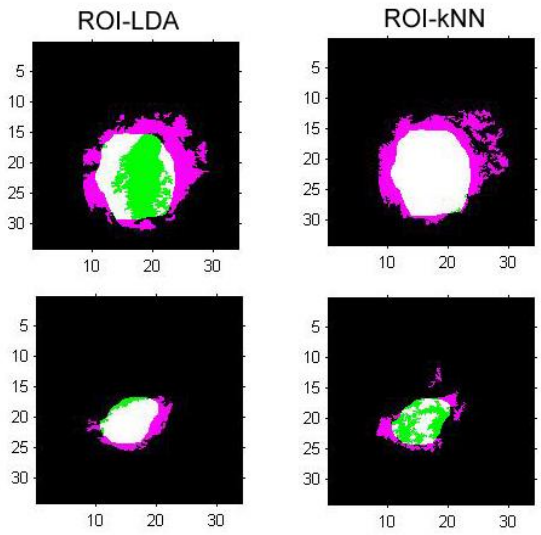
Figur 6-20: ROC kurvene for den beste ROI-LDA modellen og tilsvarende modell for ROI-kNN med tre naboer, vist i Tabell 6-6. Modellene inkluderer T_1-vektede og T_2-vektede bilder med 8 naboer, og DCE bilder med 0 naboer, hvor alle bildene er autoskalert.

For å visualisere variasjonen mellom pasientene i kNN klassifiseringsmodellen med tre nabovoksler, vist i Tabell 6-6, plottes DSC, Kappa, Sens og Spes for hver av pasientene, som vist i Figur 6-21. Den gjennomsnittlige DSC verdien for hver pasient varierte fra 0,13-0,84. De fleste pasientene (95%) hadde en sensitivitet over 0,8 og 80% av pasientene hadde en DSC og Kappa verdi mellom 0,4 og 0,8.



Figur 6-21: Spredningsplott for den gjennomsnittlige DSC, Kappa, Sens og Spes for hver av pasientene for kNN modellen med tre nabovoksler, se Tabell 6-6, og Figur 6-20 for ROC kurven. Klassifiseringen til pasient 34 og pasient 38 (merket med sirkel) vises i Figur 6-22.

For å illustrere variasjonene mellom de to klassifiseringsmetodene (ROI-LDA og kNN) ble to svulster valgt, vist i Figur 6-22, som er markert med en sirkel i Figur 6-21. Figuren viser hvordan noen svulster klassifiseres bedre med kNN metoden, mens andre svulster klassifiseres bedre med ROI-LDA metoden. Den viser også hvilke voksler som ble klassifisert riktig og feil i forhold til radiologen. Den øverste svulsten viser en svulst som hadde høyere DSC verdi (0,70) for kNN klassifiseringen, enn for ROI-LDA klassifiseringen (DSC = 0,54). Den nederste svulsten viser en svulst hvor ROI-LDA klassifiseringen hadde en høyere DSC verdi (0,75), enn kNN klassifiseringen (0,62)



Figur 6-22: Viser hvor godt LDA klassifiseringen (kolonne til venstre) og kNN klassifiseringen med tre nabovoksler (kolonne til høyre) overlapper med radiologens maske for to bildesnitt. De rosa områdene indikerer FP (Falsk positiv), altså der modellen feilaktig klassifiserer som svulst. De grønne områdene indikerer FN (Falsk negativ), der modellen feilaktig klassifiserer som friskt vev. De svarte områdene indikerer SN (Sann negativ), de områdene som modellen riktig klassifiserer som frisk vev, og de hvite områdene indikerer SP (Sann positiv), altså de områdene som modellen riktig klassifiserer som svulst. Forskjellen på DSC verdien for de to ulike klassifiseringsmetodene er 0,16 for den første raden og 0,13 for den andre raden. Se Figur 6-21 for klassifiseringsresultatene til pasient 34 (øverste rad) og pasient 38 (nederste rad) for ROI-kNN (høyre kolonne).

7 Diskusjon

Et dataprogram for automatisk inntegning av livmorhalskreftsvulster har blitt videreutviklet. Dette programmet kombinerer informasjon for flere typer MR-bilder. Ulike preprosesseringsmetoder og klassifiseringsmetoder har blitt implementert og testet i programmet. I tillegg ble en postprosesseringsmetode ved bruk av ROI (*Region of Interest*) utviklet. Ytelsen til programmet evalueres ved bruk av svulstinntegninger laget av erfarne radiologer. Formålet var å undersøke om metodene for preprosessering, ikke-lineære klassifiseringsmetoder og postprosessering av MR-bildene ga en mer nøyaktig klassifiseringen av svulstene, enn de tidligere implementerte metodene. Nedenfor følger en vurdering av de implementerte metodene.

7.1 Preprosessering

Tre preprosesseringsalgoritmer ble implementert og testet i programmet for automatisk inntegning av livmorhalskreftsvulster. Formålet med preprosesseringen var å kompensere for intensitetsvariasjoner mellom vokslene fra forskjellige bildeopptak.

Preprosessering av bildene med Median filteret ga tilnærmet lik klassifisering som ROI-LDA modellen uten preprosessering når T_1-vektede, T_2-vektede, DCE bildene, eller en kombinasjon av disse, ble benyttet. Det samme gjelder for autoskaleringen. Grunnen til at disse ga tilnærmet like resultater er at medianen til en vokselgruppe er tilnærmet lik gjennomsnittet av den [57]. Savitzky-Golay filteret ga ingen signifikant forbedring av ROI-LDA klassifiseringen, selv når en høy grad av glatting ble benyttet. Filteret ga generelt lavere sensitivitetsverdier, noe som tilsier at ved glatting ble klassifikasjonsnøyaktigheten for svulst lavere. CLAHE filteret hadde utfordringer knyttet til valg av parameterinnstillinger, som kunne kombineres på mange ulike måter. For de ulike parameterinnstillingene som ble testet ga ingen en signifikant forbedring av klassifiseringen.

Ingen av preprosesseringsalgoritmene ga signifikant forskjellige resultater fra ROI-LDA klassifiseringen uten preprosessering. Det var spesielt sensitiviteten og spesifisiteten som hadde en negativ eller positiv effekt av preprosesseringen, hvor ingen av metodene ga en positiv effekt for begge parameterne. Det betyr at når sensitiviteten øker, og spesifisiteten minker (eller motsatt), vil ikke klassifiseringen forbedres. Dette fordi feilklassifiseringen økte i en av klassene, mens feilklassifiseringen ble lavere i den andre klassen.

7.2 Postprosessering

I denne oppgaven ble det foreslått en postprosesseringsmetode hvor det tegnes inn en ROI (*Region of Interest*) maske i sannsynlighetskartet funnet fra LDA klassifiseringen. Masken tegnes rundt området med høyest sannsynlighet for å inneholde svulst. Denne masken legges over alle bildene (T_1-vektede, T_2-vektede og DCE) til hver enkelt pasient, slik at vokslene innenfor ROI masken inkluderes i en ny klassifisering. Dette endrer forholdet mellom svulst og ikke-svulst voksler i datasettet, og reduserte antall voksler i datasettet slik at modellen trenes raskere. Resultatene viser at ytelsesparameterne *DSC, K* og Spes ble signifikant høyere for LDA/QDA klassifiseringen basert på ROI maskede bilder. Det var ingen vesentlig forskjell på sensitiviteten for de to klassifiseringene, noe som tilsier at den var like god. Dette betyr at LDA/QDA klassifiseringen ga en bedre ytelse når den ble basert på bildene med ROI masken. Original bildene (uten ROI masken) ga lavere spesifisitet, som indikerer at flere voksler feilaktig ble klassifisert som svulst. Siden alle vokslene utenfor ROI masken automatisk ble klassifisert som friskt vev, bidro dette til å øke spesifisiteten for de ROI maskerte bildene.

Den eneste ytelsesparameteren som ble signifikant lavere for de ROI maskerte bildene var AUC (*Areal Under Curve*). Dette kommer av at AUC verdiene beregnes ved hjelp av ROC-analysen, som kun brukte vokslene inkludert i ROI masken. Sensitiviteten og spesifisiteten beregnes per pasient, og deretter beregnes et gjennomsnitt av disse. Siden de ROI maskerte bildene hadde ulike antall voksler per pasient, ble disse verdiene ulike. Denne forskjellen ble tydeligere med de ROI maskede bildene, fordi antall voksler per pasient varierte når man valgte en ROI maske med ulik størrelse for hver pasient. Dette betyr at det var en forskjell på analysene, og resultatene er dermed ikke direkte sammenlignbare.

Hensikten med ROI masken er å kombinere kunnskapen til radiologiene og effektiviteten til den automatiske klassifiseringen. I denne oppgaven ble ROI masken tegnet av forfatteren med bakgrunn i sannsynlighetskartet for den beste LDA modellen. Dette kan ha ført til usikkerheter knyttet til inntegningen av masken, hvor svulstvoksler kan ha havnet utenfor ROI masken.

7.3 Klassifiseringsmetoder

Resultatene fra de ulike klassifiseringsmetodene indikerte at det ikke var signifikant forskjell mellom metodene. Ingen av de ikke-lineære metodene ga signifikant bedre klassifisering av svulstene enn den lineære (ROI-LDA) klassifiseringsmetoden. Uavhengig av klassifiseringsalgoritme, bygde modellene med høyest ytelse på T_1-vektede og T_2-vektede bilder med 8 naboer og DCE bilder med 0 naboer. Selv om ingen av klassifiseringsmetodene utmerket seg i stor grad, var ytelsen høy for samtlige. Det var også variasjoner mellom pasientene i hvor nøyaktig de beste modellene predikerte svulstvokslene.

Resultatene basert på *Random Forest* metoden med 50 trær viser ingen signifikante forskjeller fra ROI-LDA klassifiseringen, med unntak av spesifisiteten og sensitiviteten. Fordi spesifisiteten til *Random Forest* modellen var vesentlig høyere enn sensitiviteten, indikerer dette at de fleste vokslene havnet i klassen ikke-svulst. Selv om ROI-LDA klassifiseringen hadde en lavere sensitivitet enn *Random Forest* klassifiseringen, hadde den også mindre forskjell mellom sensitivitet og spesifisitet, som indikerer en bedre balanse mellom klassifiseringen av svulst og ikke-svulst voksler. DSC og Kappa verdiene for *Random Forest* modellen varierte i tillegg mer for hver enkelt pasient. Dette betyr at for noen pasienter identifiserte denne algoritmen svulsten meget nøyaktig, med maksimal DSC verdi på 0,86. For andre pasienter derimot, fant algoritmen svært få voksler som kunne klassifiseres som svulst, med en minimum DSC verdi på 0,05. Sammenlignet med ROI-LDA metoden hadde denne en minimum og maksimum DSC verdi på henholdsvis 0,14 og 0,77. Grunnen til den store variasjonen i DSC verdier for *Random Forest* algoritmen kan være overtilpasning [70]. Hvert tre i *Random Forest* modellen er uavhengig av hverandre [70], noe som tilsier at hvert tre passer godt til datapunktene i treningssettet. Når man videre klassifiserer vokslene som ikke er en del av treningssettet vil svulster som ligner på svulster i treningssettet klassifiseres godt, mens svulster som ikke ligner vil ha en dårlige klassifisering. I motsetning til dette trekker ROI-LDA metoden en rett linje mellom de to vokselgruppene. Denne modellen vil være mer robust for ulike svulster, og vil dermed klassifisere svulstene for hver pasient jevnere. Overtilpasning kan også bli et problem når man senere ønsker å teste metoden på andre datasett [70], som inneholder svulster med andre typer former og strukturer. I tillegg brukte *Random Forest* metoden 10 timer lenger tid på å trene modellen enn ROI-LDA klassifiseringen.

Resultatene fra kNN klassifiseringsmetoden med tre nabovoksler viser at Kappa statistikken var signifikant høyere for kNN modellen enn for ROI-LDA modellen. I likhet med ROI-LDA klassifiseringen hadde kNN klassifiseringen en jevn balanse mellom sensitiviteten og spesifisiteten, og klassifiserte dermed vokslene godt i begge klasser. DSC verdiene mellom pasientene for kNN metoden varierte heller ikke like mye som *Random Forest* modellen, med DSC verdier i et smalere intervall (0,13-0,84). En ulempe med kNN klassifiseringsmetoden er at avstandsberegningene må gjøres for hver nye prøve, og

avstanden til hver av de nærmeste vokslene må undersøkes [72]. Dette gjør at kNN metoden brukte i overkant av 76 timer på å trene modellen i forhold til ROI-LDA modellen som brukte i overkant av 4 minutter ved bruk av *Leave-ten-patients-out* kryss-validering.

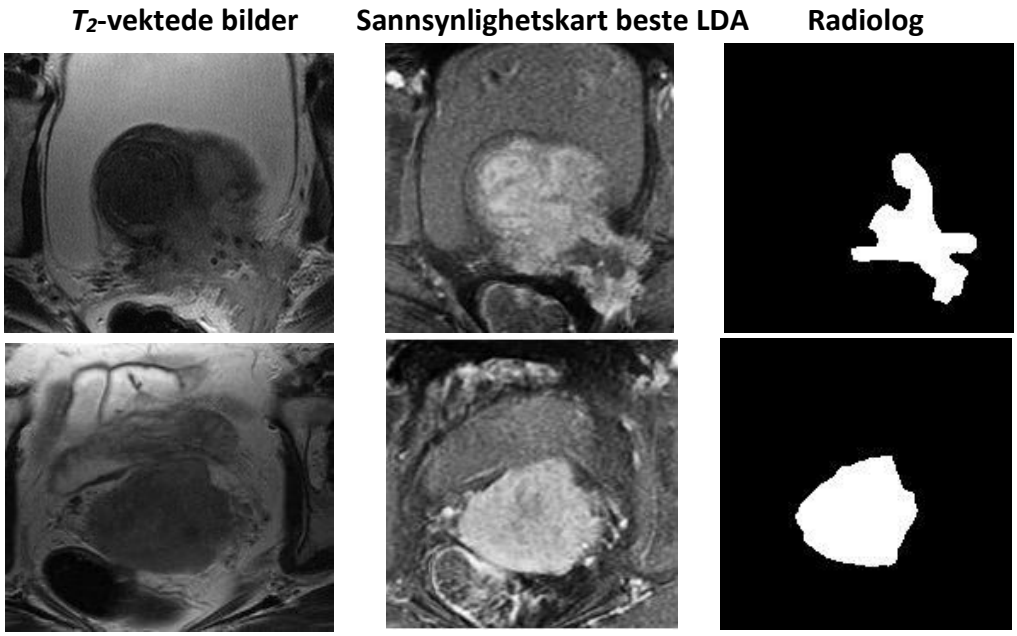
ROI-LDA klassifiseringsmetoden hadde generelt lave DSC verdier for små svulster (0,17-0,26) og hadde problemer med å klassifisere disse. *Random Forest* metoden derimot hadde høyere DSC verdier for de små svulstene (0,21-0,50), men dette var ikke signifikant forskjellig. kNN metoden hadde til sammenligning DSC verdier for de små svulstene mellom 0,13-0,32. ROI-LDA modellen hadde derimot høyere DSC verdier for de store svulstene (0,66-0,68), mens *Random Forest* metoden hadde DSC verdier mellom 0,30-0,33 for tilsvarende svulster. Til sammenligning hadde kNN metoden en DSC mellom 0,66-0,67 for de store svulstene. Dette indikerer at størrelsen på svulsten hadde innvirkning på hvilken klassifiseringsmetode som ga de høyeste DSC verdiene. Små svulster ligner mer på normalt vev, og dette kan være grunnen til at disse hadde lavere ytelsesverdier. En annen mulig grunn kan være at ROI masken ble tegnet for stor. Ved en mer passende ROI maske tegnet av erfarne radiologer, kan forholdet mellom ikke-svulst og svulst voksler for MR-bildene med små svulster reduseres, irrelevante voksler kan utelates og klassifiseringen kan forbedres.

7.4 Vurdering av svulstinntegningsprogrammet

Den beste ROI-LDA modellen hadde en gjennomsnittlig *K* verdi på 0,51 som er en rimelig overenstemmelse med radiologen, og alle modellene basert på DCE bildeserien hadde en gjennomsnittlig *K* verdi mellom 0,49-0,51. Disse verdiene er høyere enn verdier som er funnet i en studie om forventet overenstemmelse mellom radiologer av Hricak *et al.* [55]. Denne studien sammenlignet inntegninger gjort av fire radiologer på 152 pasienter med tidlig avansert livmorhalskreft, hvor radiologene ga bildene poeng som representerte svulstsynligheten. Det ble funnet en multirater Kappa verdi på 0,32 for svulstsynlighet, noe som indikerte uoverensstemmelser mellom radiologene. Dimopoulos *et al.* [75] sammenlignet inntegninger som ble gjort av to strålingsonkologer. De brukte andre parameter for å sammenligne overenstemmelsen, enn det som er gjort i denne oppgaven, men også de fant store variasjoner mellom strålingsonkologene. Derfor var enigheten mellom den beste ROI-LDA modellen og radiologene sammenlignbar med det som kan forventes mellom radiologer.

For noen av de beste klassifiserte svulstene for ROI-LDA modellen, vist i Figur 6-6, var det kun feilklassifisering i områdene rundt svulsten. Dette er områder hvor det er forventet høy usikkerhet, også blant radiologene [55]. Radiologene legger ofte på en ekstra sikkerhetsmargin for å kompensere for disse usikkerhetene [7]. ROI-LDA klassifiseringen av svulstene i Figur 6-6 var bedre enn forventet overenstemmelse mellom radiologer, presentert i Hricak *et al.* [55]. På den andre siden var det også svulster som hadde lave

ytelsesverdier med ROI-LDA klassifiseringen. Grunnen til dette kan være lav kontrast i MR-bildene. Hvis man tar eksempel i to T_2-vektede MR-bilder som vist i Figur 7-1 ser man tydelig konturer av svulsten i det T_2-vektede bildet i den nederste raden. Denne svulsten ble klassifisert godt i ROI-LDA analysen (DSC = 0,74). Svulsten i den øverste raden er vanskeligere å se på det T_2-vektede bildet, og ROI-LDA klassifiseringen ga en lavere DSC verdi (0,30). I det T_2-vektede bildet for den øverste svulsten er det vanskelig å finne en tydelig grense mellom vokselintensitetene. Denne lave kontrasten mellom svulst/ikke-svulst er også noe som kunne ført til stor usikkerhet i den manuelle inntegningen. Bildeklassifiseringen kan ikke bli mer nøyaktig enn radiologens inntegning, siden denne brukes i modelltreningen. Det betyr at der usikkerheten er høy for radiologene, er den også høy for svulstinntegningsprogrammet. Siden dette er gamle bilder fra 2001-2004 kan muligens nyere MR-bilder med bedre romlig oppløsning og kontrast føre til en bedre klassifisering.



Figur 7-1: Viser det T_2-vektede bildet, sannsynlighetskartet for den beste LDA modellen og radiologens maske for to svulster. Det T_2-vektede bildet i den andre raden har tydeligere intensitetsforskjeller mellom svulst og omkringliggende vev, enn det T_2-vektede bildet.

Leave-ten-patients-out kryss-validering ble brukt for å kompensere for overtilpasning, og for å kunne teste ytelsen til modellene. Her ble samme datasettet delt inn i forskjellige deler for trening og validering. Imidlertid hadde det vært fordelaktig å teste klassifiseringen på et uavhengig datasett [70]. Dette er også hensikten med automatiseringen av svulstinntegningen. Tanken er at programmet skal trenes en gang på et datasett, for så å kunne identifisere alle typer svulster som ikke er en del av dette treningssettet.

En annen fordel med det automatiske svulstinntegningsprogrammet er at det er fleksibelt og kan enkelt utvides til å inkludere flere andre bildetyper, for eksempel diffusjonsvektet MR [76], som nå brukes i undersøkelser av livmorhalskreft [77]. Uavhengig av

klassifiseringsalgoritme, bygde modellene med høyest ytelse på T_1-vektede og T_2-vektede bilder med 8 naboer og DCE bilder med 0 naboer, hvor alle bildene er autoskalert. Klassifiseringsmodeller med DCE, T_1-vektede og T_2-vektede bilder separat, hadde generelt lavere ytelsesparametere. Det vil si at det lå informasjon i alle bildetypene, samt i forholdet mellom nabovoksler som kan utnyttes i klassifiseringen. I kombinasjon med en passende registeringsalgoritme kan også andre typer bilder inkluderes i klassifiseringen, som vektlegger andre egenskaper til svulsten. Et eksempel er PET bilder som fanger metabolismen i svulstene [10, 11]. Ved bruk av flere typer bilder kan det beregnes svulstsannsynlighetskart som viser forskjellige typer egenskaper, for eksempel hypoksi, som knyttes til lav strålefølsomhet [78]. Dette har potensialet til å bli brukt til eksempelvis doseberegninger for strålebehandling, der hypoksiske områder i svulster kan bestråles med høyere dose [79].

Selv om svulstinntegningsprogrammet ikke skulle blitt implementert i sykehusene, kan det benyttes som en tilleggsobservatør. Det er ofte vanlig å la to radiologer tegne inn svulsten [9], og denne automatiske inntegningen kan brukes istedenfor en av radiologene slik at arbeidet halveres.

7.5 Sammenligning med andre inntegningsstudier

Tidligere studier av segmentering av kreftsvulster er gjort på ulike måter. Litjens *et al.* [9] beskriver en metode for prostatasvulstavgrensing hvor selve vokselklassifiseringen utføres med bakgrunn i symmetri, lokal kontrast og anatomiske egenskaper. I denne studien ble 347 pasienter undersøkt, hvorav 164 hadde prostatakreft. Metoden segmenterte først prostata, deretter ble bildene analysert for å undersøke sannsynligheten for at pasienten hadde prostatakreft. Til slutt ble kreftsvulsten automatisk inntegnet ved å trene klassifiseringsmodellen på masker fra erfarne radiologer. *Random Forest* med 300 trær og Gentleboost var vokselklassifiseringsmetodene som oppnådde de beste resultatene, med en AUC verdi på henholdsvis 0,89 og 0,87. Sensitiviteten per svulst var på 0,55. Resultatene fra *Random Forest* modellen i denne oppgaven lå på omtrent de samme verdiene, med en AUC på 0,91 og en gjennomsnittlig sensitivitet på 0,56. Forskjellene mellom studiene er åpenbart forskjellige kjønn på pasientene og segmentering av forskjellige type svulster. I tillegg ble prostata nøye segmentert ved bruk av atlas basert segmentering, mens i denne oppgaven ble en ROI maske implementert med bakgrunn på sannsynlighetskartene for LDA analysen. Likevel ga analysene omtrent like resultater, noe som kan indikere at metoden for automatisk svulstavgrensing kan generaliseres og brukes på flere typer kreft. For studien til Litjens *et al.* fungerte LDA klassifiseringen signifikant dårligere enn *Random Forest* metoden, noe som ikke var tilfelle i denne oppgaven.

En annen studie utført av Lu *et al.* [6] for automatisk avgrensing av livmorhalskreft, beskriver en registreringsmetode for svulstdeteksjon som skanner pasienten flere ganger i løpet av behandlingsforløpet. Dette for å kompensere for at svulsten forandrer og flytter på seg. Lu *et al*. [6] oppnådde en *Dice Similarity Coefficient* (DSC) på 0,78 ± 0,03 for deteksjon av GTV for seks pasienter gjennom en fem ukers periode, med 32 bildesnitt for T_2-vektede bilder. Denne automatiske inntegningen ble sammenlignet med en manuell inntegning gjort av en strålingsonkolog, og godkjent av en radiolog [6]. Blæren og livmorhalsen ble segmentert først, før svulsten ble klassifisert ved bruk av en *kernel density estimation (KDE)*.
Denne metoden var først og fremst basert på formen til svulstene og var i tillegg basert på tidligere inntegninger av den samme svulsten. Studien til Lu e*t al.* oppnådde en relativt høy DSC for segmentering av livmorhalskreftsvulsten i forhold til hva man oppnådde i denne oppgaven. Allikevel er det vanskelig å sammenligne resultatene siden vi i denne oppgaven brukte alle typer av MR-bildene, i tillegg til å ha et høyere pasientantall. I denne oppgaven tok man heller ikke utgangspunkt i en svulstinntegning, slik som i Lu *et al*. og svulsten ble klassifisert helt fra grunnen. Lu *et al.* konkluderte videre at denne metoden fungerte best for svulster som var runde i formen, noe som ikke alltid er tilfelle. Ved bruk av klassifiseringsmetodene i denne oppgaven kan man finne en mer robust algoritme for å klassifisere kreftsvulstene.

7.6 Forslag til videre arbeid

Det første skrittet i videre arbeid med denne oppgaven er å teste inntegningsprogrammet som er utviklet, på nyere bildedata av livmorhalskreft. Det er kun benyttet ett datasett i denne oppgaven, som er over 15 år gammelt. Ny teknologi og bedre bildekvalitet er utviklet for MR-avbildning, som kan gi nye og interessante funn. For eksempel brukes nå også diffusjonsvektet MR for å kartlegge livmorhalskreft og i noen tilfeller også PET [77]. Bedre bildekvalitet kan kanskje føre til en større forskjell på ytelsen til klassifiseringsmodellene og generelt øke klassifiseringsnøyaktigheten. Det hadde også vært interessant å teste inntegningsprogrammet på datasett som er uavhengig av treningssettet. Da får man en mer reell indikasjon på hvor godt den automatiske svulstinntegningen vil gjøre det på generell basis. I tillegg burde modellene trenes på enda flere pasienter, slik at man får en mer robust klassifisering. Det hadde også vært interessant å teste programmet for andre krefttyper, spesielt krefttyper i buken, som for eksempel prostatakreft.

Videre burde ROI masken testes for bruk sammen med radiologer. Om radiologene tegner inn ROI masken vil man få vurdert samspillet mellom radiologene og det automatiske svulstinntegningsprogrammet. Hovedpoenget her ligger i at radiologene kan tegne et grovt estimat på hvor de tror svulsten ligger, og la dataprogrammet tegne inn den nøyaktige avgrensingen. Dette vil spare radiologer for mye tid, i tillegg til at inntegningen ikke får

variasjoner mellom radiologene. Radiologer kan også bruke de T_2-vektede bildene for å tegne inn ROI masken, i tillegg til sannsynlighetskartet fra den beste LDA modellen.

Flere klassifiseringsmetoder, preprosesseringer og postprosesseringer kan testes. Hver av de nåværende klassifiseringsmetodene kan testes med andre innstillingsparametere enn det som er gjort i denne analysen. I tillegg finnes det flere preprosesseringsalgoritmer som kan kompensere for intensitetsforskjeller mellom bilder, men innvirkningen av preprosesseringen kan bli mindre betydningsfull med nyere og bedre bildekvalitet. Det er også mulig å teste klassifiseringsalgoritmene som ikke ble optimalt testet på grunn av tidsbegrensinger. Dette kan gjøres ved å optimalisere programkoden, teste på mindre datasett med færre pasienter eller bruke en kraftigere datamaskin. Slik vil treningstiden bli kortere, og man kan hente ut resultater fra disse klassifiseringsmetodene.

Resultatene fra de ulike klassifiseringsmodellene viser at ulike svulster (små/store) klassifiseres bedre for noen klassifiseringsmetoder. En mulighet er å implementere en algoritme som kan gi et godt estimat på hvilken klassifiseringsmetode som vil utføre den beste klassifiseringen for den enkelte pasienten. Om algoritmen klarer å finne sannsynligheten for at svulsten er liten/stor, kan den foreslå hvilken klassifiseringsmetode som mest sannsynlig vil gi den beste klassifiseringen. En annen mulighet er å legge inn et advarselssystem som advarer radiologene om den automatiske svulstinntegningen er høyst usannsynlig og ber dem teste andre klassifiseringsmetoder.

8 Konklusjon

Et automatisk svulstinntegningsprogram for livmorhalskreft ble videreutviklet og nye metoder for preprosessering, vokselklassifisering og postprosessering av MR-bilder ble testet. Disse metodene ble sammenlignet med tidligere lineære klassifiseringsmodeller.

Postprosessering ved å tegne inn en ROI maske som omringer et område med høy sannsynlighet for svulst, gitt i sannsynlighetskart predikert av den beste lineære modellen, er utført. Masken legges på originalbildene og svulstene klassifiseres på nytt ved bruk av lineære eller ikke-lineære klassifiseringsmetoder. Den lineære klassifiseringsmodellen med ROI masken (ROI-LDA) ga signifikant bedre klassifisering enn den tidligere lineære modellen.

Preprosesseringsmetoder for å kompensere for intensitetsvariasjoner mellom voksler fra forskjellige bildeopptak ble testet, for å undersøke om det forbedret den lineære vokselklassifiseringen. Ingen av preprosesseringsmetodene median filter, Savizky-Golay filter og CLAHE filter ga signifikant forbedring av klassifiseringen.

Ikke-lineære klassifiseringsmetoder basert på ROI maskede bilder ble implementert for å undersøke om disse ga en bedre klassifisering av svulsten enn den lineære ROI-LDA. Klassifiseringsmetodene *Random Forest* og kNN ga ikke en signifikant bedre svulstinntegning, og klassifiserte i stor grad på lik linje med ROI-LDA. *Random Forest* metoden ga større variasjon i inntegningen mellom pasientene enn ROI-LDA og kNN metodene. kNN metoden hadde noe, men ikke signifikant, høyere spesifisitet, sensitivitet og DSC, samt signifikant høyere Kappa enn ROI-LDA metoden. De ikke-lineære klassifiseringsmetodene brukte vesentlig lenger tid på modelltreningen enn ROI-LDA. SVM klassifiseringen ble ikke tilstrekkelig testet, grunnet for lange treningstider som ikke kunne fullføres. *AdaBoost* metoden klassifiserte alle vokslene som ikke-svulst.

Uavhengig av klassifiseringsmetode, bygde modellene med høyest ytelse på T_1-vektede og T_2-vektede bilder med 8 naboer og DCE bilder med 0 naboer. Dette betyr at informasjonen i alle bildetypene, samt forholdet mellom nabovoksler, er av betdyning ved klassifisering av svulstvokslene. Ved å inkludere flere bildetyper som vektlegger andre svulstegenskaper kan inntegningsprogrammet optimaliseres, og potensielt brukes til flere formål.

Klassifiseringsmetodene ROI-LDA og kNN utmerket seg ved å ha spesifisitet og sensitivitet over 0,90. I tillegg hadde begge modellene høye DSC og Kappa verdier på henholdsvis 0,53 og 0,51 for ROI-LDA og 0,58 og 0,56 for kNN. Disse Kappa verdiene er høyere enn forventet overensstemmelsen mellom radiologer. Således viser det automatiske inntegningsprogrammet for livmorhalskreftsvulster potensiale til å blir et nyttig verktøy for radiologer.

Kildeliste

1. International Agency for Research on Cancer, *Human papillomaviruses (IARC monographs on the evaluation of carcinogenic risks to humans, Volume 64).* Lyon, France: IARC, 1995.
2. Torre, L.A., et al., *Global cancer statistics, 2012.* CA: a cancer journal for clinicians, 2015. **65**(2): p. 87-108.
3. Kreftforeningen. *Livmorhalskreft*. 2015 [cited 2016 01.03]; Available from: https://kreftforeningen.no/om-kreft/kreftformer/livmorhalskreft/.
4. Parkin, D.M. and F. Bray, *The burden of HPV-related cancers.* Vaccine, 2006. **24**: p. S11-S25.
5. Ghose, S., et al., *A review of segmentation and deformable registration methods applied to adaptive cervical cancer radiation therapy treatment planning* Artificial Intelligence in Medicine, 2015(64): p. 75-87.
6. Lu, C., et al., *Simultaneous nonrigid registration, segmentation, and tumor detection in MRI guided cervical cancer radiation therapy.* Medical Imaging, IEEE Transactions on, 2012. **31**(6): p. 1213-1227.
7. Njeh, C., *Tumor delineation: The weakest link in the search for accuracy in radiotherapy.* Journal of Medical Physics, 2008. **33**(4): p. 136.
8. Haider, M.A., et al., *Correlations between dynamic contrast-enhanced magnetic resonance imaging–derived measures of tumor microvasculature and interstitial fluid pressure in patients with cervical cancer.* Journal of Magnetic Resonance Imaging, 2007. **25**(1): p. 153-159.
9. Litjens, G., et al., *Computer-aided detection of prostate cancer in MRI.* Medical Imaging, IEEE Transactions on, 2014. **33**(5): p. 1083-1092.
10. Kim, E.E., et al., *Clinical PET and PET/CT: principles and applications*. 2012: Springer Science & Business Media.
11. Granov, A., A. Stanzhevskiy, and T. Schwarz, *Positron emission tomography*. 2013: Springer Science & Business Media.
12. Delaney, G., et al., *The role of radiotherapy in cancer treatment.* Cancer, 2005. **104**(6): p. 1129-1137.
13. Barnett, G.C., et al., *Normal tissue reactions to radiotherapy: towards tailoring treatment dose by genotype.* Nature Reviews Cancer, 2009. **9**(2): p. 134-142.
14. Klein, S., et al., *Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information.* Medical physics, 2008. **35**(4): p. 1407-1417.
15. Ourselin, S., et al., *Reconstructing a 3D structure from serial histological sections.* Image and vision computing, 2001. **19**(1): p. 25-31.
16. Rueckert, D., et al., *Nonrigid registration using free-form deformations: application to breast MR images.* Medical Imaging, IEEE Transactions on, 1999. **18**(8): p. 712-721.
17. Chan, I., et al., *Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier.* Medical physics, 2003. **30**(9): p. 2390-2398.
18. Torheim , T.K.G., et al., *Autodelineation of cervical cancers using multiparametric MRI and linear discriminant analysis* In Multivariate Image Analysis Cancer

Treatment Planning and Evaluation, Norwegian University of Life Science, Ås. PhD Thesis, 2016:21.
19. Oncolex. *Stadier av livmorhalskreft*. 2013 [cited 2016 10.03]; Onkologisk oppslagsverk]. Available from: http://oncolex.no/GYN/Diagnoser/Livmorhals/Bakgrunn/Stadier.
20. IARC Monograph Working Group, *IARC monographs on the evaluation of carcinogenic risks to humans, vol 90: human papillomavirusses.* Lyon: International Agency for Research in Cancer, 2007.
21. Schiffman, M., et al., *Human papillomavirus and cervical cancer.* The Lancet, 2007. **370**(9590): p. 890-907.
22. Burchell, A.N., et al., *Epidemiology and transmission dynamics of genital HPV infection.* Vaccine, 2006. **24**: p. S52-S61.
23. Forman, D., et al., *Global burden of human papillomavirus and related diseases.* Vaccine, 2012. **30**: p. F12-F23.
24. Behtash, N. and N. Mehrdad, *Cervical cancer: screening and prevention.* Asian Pac J Cancer Prev, 2006. **7**(4): p. 683-6.
25. ICRU, *recording and reporting photon beam therapy (supplement to ICRU report 50).* ICRU report, 1999. **62**.
26. McRobbie, D.W., et al., *MRI: From picture to proton* Second ed. 2006: Cambridge University Press.
27. Westbrook, C., C.K. Roth, and J. Talbot, *MRI in practice*. fourth ed. 2011: Wiley-Blackwell
28. Bushong, S.C., *Magnetic Resonance Imaging: Physical and Biological Principles*. Second ed. 1996: Mosby.
29. Bitar, R., et al., *MR pulse sequences: What every radiologist wants to know but is afraid to ask* Radiographics, 2006. **26**(2): p. 513-537.
30. Lilley, J., *Nuclear physics: principles and applications*. 2013: John Wiley & Sons.
31. Hornak, J.P. *The Basics of MRI*. 1996-2014; Interactive Learning SoftWare]. Available from: http://www.cis.rit.edu/htbooks/mri/index.html.
32. Brown, M.A. and R.C. Semelka, *MRI: Basic principles and applications*. Third ed. 2003: John Wiley & Sons
33. Hashemi, R.H. and W.G. Bradley, *MRI: the basics*. 1997: Williams & Wilkins.
34. Brant, W.E. and E.E. de Lange, *Essentials of Body MRI*. 2012: Oxford University Press.
35. Yang, C., et al., *Comparison of quantitative parameters in cervix cancer measured by dynamic contrast–enhanced MRI and CT.* Magnetic resonance in medicine, 2010. **63**(6): p. 1601-1609.
36. Bjørnerud, A., *The Physics of Magnetic Resonance Imaging.* FYS-KJM-4740, University of Oslo 2008.
37. Yankeelov, T.E. and J.C. Gore, *Dynamic contrast enhanced magnetic resonance imaging in oncology: theory, data acquisition, analysis, and examples.* Current medical imaging reviews, 2009. **3**(2): p. 91.
38. Zahra, M.A., et al., *Dynamic contrast-enhanced MRI as a predictor of tumour response to radiotherapy.* The Lancet Oncology, 2007. **8**(1): p. 63-74.
39. Tofts, P.S., et al., *Estimating kinetic parameters from dynamic contrast-enhanced T 1-weighted MRI of a diffusable tracer: standardized quantities and symbols.* Journal of Magnetic Resonance Imaging, 1999. **10**(3): p. 223-232.
40. Torheim, T.K.G., *Multivariate analysis of DCE-MRI images of cancer tumours*. 2011, MSc thesis, Norwegian University of Life Science, Ås

41. Tofts, P.S., *Modeling tracer kinetics in dynamic Gd-DTPA MR imaging.* Journal of Magnetic Resonance Imaging, 1997. **7**(1): p. 91-101.
42. Andersen, E.K.F., *Dynamisk kontrastforsterket MRI av pasienter med livmorhalskreft. Korrelasjonsanalyse av bildeparametre mot langtidsoverlevelse etter stråleterapi.* 2009, MSc thesis, University of Oslo.
43. Andersen, E.K.F., et al., *Pharmacokinetic analysis and k-means clustering of DCEMR images for radiotherapy outcome prediction of advanced cervical cancers.* Acta Oncologica, 2011. **50**(6): p. 859-865.
44. Torheim, T.K.G., *Multivariate Image Analysis in Cancer Treatment Planning and Evaluation.* Norwegian University of Life Science, Ås. PhD thesis 2016:21.
45. Kruse, O.M.O., et al., *Pixel classification methods for identifying and quantifying leaf surface injury from digital images.* Computers and Electronics in Agriculture, 2014(108): p. 155-165.
46. Prats-Montalbán, J.M., A. de Juan, and A. Ferrer, *Multivariate image analysis: A review with applications.* Chemometr Intell Lab 2011(107): p. 1-23.
47. Fisher, R.A., *The use of multiple measurements in taxonomic problems.* Annals of Eugenics, 1936. **7**: p. 179-188.
48. Wang, S.J. and R.M. Summers, *Machine learning and radiology.* Medical Image Analysis, 2012. **16**: p. 933-951.
49. Korporaal, J.G., et al., *The use of probability maps to deal with the uncertainties in prostate cancer delineation.* Radiotherapy and Oncology, 2010. **94**: p. 168-172.
50. Gonzalez, R.C., R.E. Woods, and S.L. Eddins, *Digital Image Processing using MATLAB*. 2004: Pearson Education, Inc.
51. Olson, D.L. and D. Delen, *Performance evaluation for predictive modeling*. In Advanced data mining techniques. 2008: Springer.
52. Dice, L.R., *Measures of the Amount of Ecologic Association Between Species.* Ecology, 1945. **26**: p. 297-302.
53. García-Lorenzo, D., et al., *Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging.* Medical image analysis, 2013. **17**(1): p. 1-18.
54. Fleiss, J.L., B. Levin, and M.C. Paik, *The measurement of interrater agreement statistical methods for rates and proportions*. 2003: John Wiley & Sons, Inc.
55. Hricak, H., et al., *Early Invasive Cervical Cancer: CT and MR Imaging in Preoperative Evaluation—ACRIN/GOG Comparative Study of Diagnostic Performance and Interobserver Variability 1.* Radiology, 2007. **245**(2): p. 491-498.
56. Brown, C.D. and H.T. Davis, *Receiver operating characteristics curves and related decision measures: A tutorial.* Chemometrics and Intelligent Laboratory Systems, 2006. **80**(1): p. 24-38.
57. Burger, W., et al., *Principles of Digital Image Processing*. 2009: Springer.
58. Savitzky, A. and M.J. Golay, *Smoothing and differentiation of data by simplified least squares procedures.* Analytical chemistry, 1964. **36**(8): p. 1627-1639.
59. Krumm, J., *Savitzky-Golay filters for 2D images.* Microsoft Research, Microsoft Corporation, Redmond, WA, 2001. **98052**.
60. Ruffin, C. and R.L. King. *The analysis of hyperspectral data using Savitzky-Golay filtering-Theoretical basis (part 1)*. in *Geoscience and remote sensing symposium, Hamburg, Germany*. 1999.
61. Gander, W. and U. von Matt. *Ch 9. Smoothing Filters* 1997 [cited 2016 16.03]; Available from: http://webcache.googleusercontent.com/search?q=cache:vzbO375julMJ:www.sprin

ger.com/cda/content/document/cda_downloaddocument/9783540617938-c1.pdf%3FSGWID%3D0-0-45-79520-p175272572+&cd=6&hl=en&ct=clnk&gl=no.
62. Schafer, R.W., *What is a Savitzky-Golay filter? [lecture notes].* Signal Processing Magazine, IEEE, 2011. **28**(4): p. 111-117.
63. Solomon, C. and T. Breckon, *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. 2011: John Wiley & Sons.
64. Reza, A.M., *Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement.* Journal of VLSI signal processing systems for signal, image and video technology, 2004. **38**(1): p. 35-44.
65. Hummel, R., *Image enhancement by histogram transformation.* Computer graphics and image processing, 1977. **6**(2): p. 184-195.
66. Pizer, S.M., *An automatic intensity mapping for the display of CT scans and other images.* Proc. 7th Int. Mtg. on Information Process. In *Medical Imaging*, 1981: p. 276-309.
67. Ketcham, D.J. *Real-time image enhancement techniques*. in *Image processing*. 1976. International Society for Optics and Photonics.
68. Pizer, S.M., et al., *Adaptive histogram equalization and its variations.* Computer vision, graphics, and image processing, 1987. **39**(3): p. 355-368.
69. Zuiderveld, K. *Contrast limited adaptive histogram equalization*. in *Graphics gems IV*. 1994. Academic Press Professional, Inc.
70. James, G., et al., *An Introduction to Statistical Learning*. 2013: Springer.
71. Kuhn, M. and K. Johnson, *Classification Trees and Rule-Based Models, In Applied predictive modeling*. 2013: Springer.
72. Næs, T., et al., *A User-Friendly Guide to Multivariate Calibration and Classification*. 2002: NIR Publications.
73. Härdle, W., D.D. Prastyo, and C.H. Hafner, *Support Vector Machines with evolutionary feature Selection for Default Prediction. In Handbook of Applied Nonparametric and Semi-parametric Econometrics and Statistics*, ed. R.S. J. Racine, and Aman Ullah (eds) Oxford University Press, New York. 2014.
74. MathWorks. *Multiple Comparisons*. [cited 2016 27.04]; Available from: http://se.mathworks.com/help/stats/multiple-comparisons.html.
75. Dimopoulos JC, D.V.V., Berger D, *Inter-observer comparison of target delineation for MRI-assisted cervical cancer brachytherapy: application of the GYN GEC-ESTRO recommendations.* Radiother Oncol, 2009. **91**: p. 166-172.
76. Harry, V.N., *Novel imaging techniques as response biomarkers in cervical cancer.* Gynecologic oncology, 2010. **116**(2): p. 253-261.
77. Haldorsen, I.S., *MR bekken ved gynekologiske tilstander* Haukeland Universitetssykehus.
78. Rockwell, S., et al., *Hypoxia and radiation therapy: past history, ongoing research, and future promise.* Current molecular medicine, 2009. **9**(4): p. 442.
79. van der Heide, U.A., et al., *Functional MRI for radiotherapy dose painting.* Magnetic resonance imaging, 2012. **30**(9): p. 1216-1223.

Vedlegg

Vedlegg A: Vokselklassifisering med LDA

Images	Preprocessing	Neighbours	DSC	K	Sens	Spec	AUC
T2w	None	0	0.19	0.15	0.91	0.50	0.65
T2w	Autoscaled	0	0.19	0.15	0.91	0.52	0.73
T2w	None	8	0.20	0.15	0.90	0.51	0.66
T2w	Autoscaled	8	0.20	0.15	0.91	0.53	0.67
T1w	None	0	0.21	0.21	0.95	0.50	0.70
T1w	Autoscaled	0	0.20	0.17	0.95	0.52	0.73
T1w	None	8	0.22	0.19	0.95	0.51	0.71
T1w	Autoscaled	8	0.21	0.18	0.95	0.53	0.75
T2w + T1w	None	0	0.21	0.18	0.93	0.51	0.72
T2w + T1w	Autoscaled	0	0.21	0.17	0.95	0.52	0.75
T2w + T1w	None	8	0.22	0.18	0.93	0.52	0.73
T2w + T1w	Autoscaled	8	0.21	0.18	0.95	0.53	0.76
DCE	None	0	0.39	0.36	0.91	0.81	0.84
DCE	Autoscaled	0	0.40	0.38	0.93	0.81	0.85
T2w + DCE	None	0	0.38	0.35	0.93	0.80	0.85
T2w + DCE	Autoscaled	0	0.39	0.37	0.93	0.80	0.85
T2w + DCE	None	T2w: 8 DCE: 0	0.39	0.36	0.93	0.80	0.85
T2w + DCE	Autoscaled	T2w: 8 DCE: 0	0.39	0.37	0.93	0.80	0.85
T1w + DCE	None	0	0.41	0.38	0.93	0.82	0.86
T1w + DCE	Autoscaled	0	0.42	0.39	0.94	0.82	0.87
T1w + DCE	None	T1w: 8 DCE: 0	0.42	0.38	0.93	0.82	0.86
T1w + DCE	**Autoscaled**	**T1w: 8 DCE: 0**	**0.42**	**0.40**	**0.94**	**0.82**	**0.87**

T2w + T1w + DCE	None	0	0.40	0.38	0.93	0.82	0.86
T2w + T1w + DCE	Autoscaled	0	0.41	0.39	0.94	0.82	0.87
T2w + T1w + DCE	None	T2w, T1w: 8 DCE: 0	0.41	0.38	0.93	0.82	0.86
T2w + T1w + DCE	Autoscaled	T2w, T1w: 8 DCE: 0	0.42	0.39	0.94	0.82	0.87

Tabell A-1: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for LDA modellen uten ROI masken, basert på forskjellige kombinasjoner av egenskapsvektoren. T_2-vektede (T2w), T_1-vektede T1w, DCE-MRI eller en kombinasjon av disse bildetypene ble brukt. Bildene var enten autoskalert eller ikke pre-prosessert. Alle bildene er post-prosessert ved bruk av en morfologisk operasjon på de binære bildene for å fjerne små elementer mindre enn 10 voksler. Kolonnen "Neighbours" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene. Hentet fra Torheim et al. [18], modifisert med tillatelse.

Vedlegg B: Vokselklassifisering med ROI-LDA

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
T2w	Ingen	0	0,39	0,36	0,85	0,85	0,63
T2w	Ingen	8	0,40	0,37	0,85	0,86	0,64
T2w	Autoskalert	8	0,41	0,38	0,87	0,87	0,66
T2w	Autoskalert	0	0,40	0,38	0,88	0,86	0,65
T1w	Ingen	0	0,40	0,38	0,89	0,85	0,67
T1w	Ingen	8	0,42	0,40	0,88	0,86	0,68
T1w	Autoskalert	8	0,44	0,42	0,93	0,86	0,72
T1w	Autoskalert	0	0,42	0,40	0,94	0,85	0,70
DCE	Ingen	0	0,51	0,49	0,88	0,91	0,76
DCE	Autoskalert	0	0,51	0,49	0,88	0,91	0,77
DCE + T1w + T2w	Autoskalert	0	0,52	0,50	0,92	0,91	0,80
DCE + T1w + T2w	None	0	0,52	0,50	0,92	0,90	0,79
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,53	0,51	0,93	0,91	0,81
T1w + DCE	Autoskalert	T1w: 8 DCE: 0	0,53	0,51	0,93	0,91	0,81
T1w + DCE	Autoskalert	0	0,53	0,51	0,92	0,91	0,80

Tabell B-1: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for LDA modellen med ROI (region of interest) masken, basert på forskjellige kombinasjoner av egenskapsvektoren. T_2-vektede (T2w), T_1-vektede (T1w), DCE-MRI eller en kombinasjon av disse bildetypene ble brukt. Bildene var enten autoskalert eller ikke pre-prosessert. Alle bildene er post-prosessert ved bruk av en morfologisk operasjon på de binære bildene for å fjerne små elementer mindre enn 10 voksler. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

Vedlegg C: ROI-QDA klassifisering

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
T2w	Ingen	0	0,39	0,37	0,96	0,82	0,72
T2w	Ingen	8	0,45	0,43	0,96	0,86	0,80
T2w	Autoskalert	8	0,47	0,45	0,95	0,87	0,81
T2w	Autoskalert	0	0,40	0,38	0,96	0,83	0,75
T1w	Ingen	0	0,40	0,38	0,95	0,83	0,73
T1w	Ingen	8	0,48	0,47	0,97	0,88	0,82
T1w	Autoskalert	8	0,49	0,47	0,96	0,87	0,83
T1w	Autoskalert	0	0,42	0,40	0,96	0,85	0,77
DCE	Ingen	0	0,41	0,39	0,96	0,83	0,77
DCE	Autoskalert	0	0,42	0,40	0,94	0,84	0,77
DCE + T1w + T2w	Autoskalert	0	0,51	0,49	0,95	0,89	0,85
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,56	0,54	0,95	0,91	0,88
T1w + DCE	Autoskalert	T1w: 8 DCE: 0	0,53	0,51	0,95	0,90	0,86
T1w + DCE	Autoskalert	0	0,48	0,46	0,95	0,88	0,83

Tabell C-1: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for QDA modellen med ROI (region of interest) basert på forskjellige kombinasjoner av egenskapsvektoren. T_2-vektede (T2w), T_1-vektede (T1w), DCE-MRI eller en kombinasjon av disse bildetypene ble brukt. Bildene var enten autoskalert eller ikke pre-prosessert. Alle bildene er post-prosessert ved bruk av en morfologisk operasjon på de binære bildene for å fjerne små elementer mindre enn 10 voksler. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

Vedlegg D: ROI-kNN klassifisering

En nabovoksel:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
T2w + T1w	Autoskalert	0	0,31	0,29	0,25	0,98	0,59
DCE	Autoskalert	0	0,44	0,41	0,46	0,97	0,61
DCE + T1w + T2w	Autoskalert	0	0,55	0,53	0,62	0,97	0,66
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,62	0,60	0,79	0,96	0,74

To nabovoksler:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
T2w + T1w	Autoskalert	0	0,52	0,50	0,75	0,94	0,66
DCE	Autoskalert	0	0,51	0,49	0,82	0,92	0,67
DCE + T1w + T2w	Autoskalert	0	0,58	0,56	0,87	0,94	0,74
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,60	0,59	0,92	0,93	0,81

Tre nabovoksler:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
T2w + T1w	Autoskalert	0	0,52	0,50	0,89	0,91	0,70
DCE	Autoskalert	0	0,49	0,47	0,91	0,90	0,70
DCE + T1w + T2w	Autoskalert	0	0,56	0,54	0,93	0,92	0,77
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,58	0,56	0,95	0,92	0,83

Tabell D-1: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for kNN modellen med en, to og tre nabovoksler og ROI (region of interest) masken basert på forskjellige kombinasjoner av egenskapsvektoren. T_2-vektede (T2w), T_1-vektede (T1w), DCE-MRI eller en kombinasjon av disse bildetypene ble brukt. Bildene var enten autoskalert eller ikke pre-prosessert. Alle bildene er post-prosessert ved bruk av en morfologisk operasjon på de binære bildene for å fjerne små elementer mindre enn 10 voksler. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

Vedlegg E: ROI-*Random Forest* klassifisering

10 trær:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
DCE	Autoskalert	0	0,26	0,25	0,20	0,99	0,76
DCE + T1w + T2w	Autoskalert	0	0,42	0,41	0,36	0,99	0,84
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,55	0,53	0,51	0,99	0,88
T1w + T2w	Autoskalert	0	0,21	0,20	0,15	0,99	0,73

50 trær:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
DCE	Autoskalert	0	0,29	0,28	0,24	0,99	0,80
DCE + T1w + T2w	Autoskalert	0	0,46	0,44	0,42	0,99	0,87
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,58	0,56	0,55	0,99	0,91
T1w + T2w	Autoskalert	0	0,20	0,19	0,14	0,99	0,76

100 trær:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
DCE	Autoskalert	0	0,30	0,29	0,24	0,99	0,81
DCE + T1w + T2w	Autoskalert	0	0,46	0,44	0,42	0,99	0,87
DCE + T1w + T2w	None	0	0,49	0,47	0,44	0,99	0,87
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,58	0,56	0,56	0,99	0,91
T1w + DCE	Autoskalert	T1w: 8 DCE: 0	0,47	0,46	0,45	0,99	0,88
T1w + DCE	Autoskalert	0	0,41	0,39	0,37	0,98	0,85

200 trær:

Bilder	Preprosessering	Naboskap	DSC	K	Sens	Spes	AUC
DCE	Autoskalert	0	0,30	0,29	0,24	0,99	0,81
DCE + T1w + T2w	Autoskalert	0	0,46	0,44	0,42	0,99	0,87
DCE + T1w + T2w	Autoskalert	T1w, Tw2: 8 DCE:0	0,62	0,60	0,79	0,96	0,74
T1w + T2w	Autoskalert	0	0,20	0,19	0,14	0,99	0,77

Tabell E-1: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for Random Forest modellen med 10, 50, 100 og 200 beslutningstrær og ROI (region of interest) masken basert på forskjellige kombinasjoner av egenskapsvektoren. T_2-vektede (T2w), T_1-vektede (T1w), DCE-MRI eller en kombinasjon av disse bildetypene ble brukt. Bildene var enten autoskalert eller ikke pre-prosessert. Alle bildene er post-prosessert ved bruk av en morfologisk operasjon på de binære bildene for å fjerne små elementer mindre enn 10 voksler. Kolonnen "Naboskap" indikerer hvor mange nabovoksler som ble brukt på de respektive bildene, her enten null eller åtte. Verdiene i tabellen viser et gjennomsnitt over alle bildesnittene fra alle pasientene.

Vedlegg F: ROI-LDA klassifisering med Median Filter preprosessering

Bilder	Naboskap	DSC	K	Sens	Spes	AUC
T2w	8	0,40	0,37	0,85	0,86	0,64
T1w	8	0,42	0,39	0,88	0,86	0,68
DCE	0	0,51	0,49	0,88	0,91	0,76
DCE + T1w + T2w	T1w, Tw2: 8 DCE:0	0,53	0,51	0,91	0,91	0,79

Tabell F-1: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for ROI-LDA klassifisering med Median filter preprosessering med 3x3 nabovoksler.

Vedlegg G: ROI-LDA klassifisering med Savitzky-Golay filter preprosessering

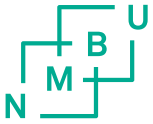
Bilder	Parameter R = rammestørrelse P = polynom	DSC	K	Sens	Spes	AUC
T2w	R = 5 P = 2	0,41	0,38	0,80	0,88	0,64
T2w	R = 9 P = 2	0,43	0,40	0,78	0,90	0,64
T1w	R = 5 P = 2	0,44	0,41	0,85	0,88	0,69
T1w	R = 9 P = 2	0,46	0,44	0,84	0,90	0,69
DCE	R = 5 P = 2	0,50	0,48	0,87	0,91	0,74
DCE	R = 7 P = 2	0,51	0,49	0,87	0,91	0,74
DCE	R = 9 P = 2	0,51	0,49	0,85	0,92	0,74
DCE	R = 9 P = 3	0,51	0,49	0,85	0,92	0,74

Tabell G-1: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for LDA klassifiseringsmetode med Savitzky-Golay glattefilter preprosessering med ulike verdier for rammestørrelse R og polynom P. Alle DCE verdiene har 0 nabovoksler, mens T_1-vektede og T_2-vektede bilder har 8 nabovoksler.

Vedlegg H: ROI-LDA klassifisering med CLAHE preprosessering

Bilder	Parameter [m,n] antall voksler k: konstrastgrense b: antall bins d: distribusjon	DSC	K	Sens	Spes	AUC
T2w	[8,8], k = 0,01, b = 256, d = rayleigh	0,30	0,28	0,96	0,73	0,80
T2w	[8,8], k = 0,005, b = 256, d = rayleigh	0,30	0,28	0,97	0,73	0,79
T2w	[32,32], k = 0,01, b = 256, d = rayleigh	0,30	0,27	0,97	0,72	0,77
T1w	[8,8], k = 0,01, b = 256, d = rayleigh	0,29	0,26	0,94	0,73	0,77
T1w	[8,8], k = 0,005, b = 256, d = rayleigh	0,30	0,27	0,97	0,72	0,78
T1w	[8,8], k = 0,02, b = 256, d = rayleigh	0,28	0,25	0,90	0,73	0,76
T1w	[32,32], k = 0,01, b = 256, d = rayleigh	0,28	0,25	0,92	0,72	0,77
T1w	[32,32], k = 0,005, b = 256, d = rayleigh	0,28	0,25	0,92	0,72	0,77
T1w	[32,32], k = 0,02, b = 256, d = rayleigh	0,24	0,20	0,77	0,73	0,71
DCE	[8,8], k = 0,01, b = 256, d = rayleigh	0,38	0,36	0,96	0,81	0,86
DCE	[8,8], k = 0,005, b = 256, d = rayleigh	0,41	0,38	0,95	0,84	0,88
DCE	[8,8], k = 0,02, b = 256, d = rayleigh	0,36	0,34	0,96	0,79	0,85
DCE	[32,32], k = 0,01, b = 256, d = rayleigh	0,33	0,31	0,98	0,76	0,86
DCE	[32,32], k = 0,005, b = 256, d = rayleigh	0,33	0,31	0,98	0,76	0,86
DCE	[32,32], k = 0,02, b = 256, d = rayleigh	0,31	0,29	0,98	0,74	0,84
DCE	[64,64], k = 0,01, b = 256, d = rayleigh	0,31	0,29	0,99	0,74	0,85
DCE	[8,8], k = 0,01, b = 64, d = rayleigh	0,42	0,40	0,96	0,85	0,88
DCE	[8,8], k = 0,005, b = 64, d = rayleigh	0,43	0,40	0,94	0,86	0,89
DCE	[8,8], k = 0,01, b = 32, d = rayleigh	0,43	0,41	0,94	0,86	0,88
DCE	[8,8], k = 0,01, b = 64, d = uniform	0,44	0,42	0,93	0,87	0,87
DCE	[8,8], k = 0,005, b = 32, d = rayleigh	0,43	0,41	0,94	0,86	0,88

Tabell H-1: DSC (Dice similarity coefficient), K (Kappa statistikk), Sens (sensitivitet), Spes (spesifisitet) og AUC (arealet under kurven) for LDA klassifiseringsmetode med CLAHE (Constrast Limited Adaptive Histogram Equalization) preprosessering med ulike verdier for antall voksler [m,n], kontrastgrense k, antall bins i histogrammet b og histogrammets distribusjon d. Alle DCE verdiene har 0 nabovoksler, mens T_1-vektede og T_2-vektede bilder har 8 nabovoksler.



Norges miljø- og biovitenskapelig universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway