Norges miljø- og
biovitenskapelige
universitet

Masteroppgave 2023 30 stp
Realtek

Automatisk deteksjon av abnormalitet i hundealbuer

Automatic detection of abnormalities in dog elbows

Sunniva Elisabeth Daae Steiro
Miljøfysikk og fornybar energi

Forord

Denne masteroppgaven markerer slutten på seks år som student på Miljøfysikk og fornybar energi ved Norges miljø- og biovitenskapelige universitet. Mye har skjedd på disse årene, og jeg hadde ikke trodd at jeg skulle skrive masteroppgave i hverken fysikk eller maskinlæring da jeg startet høsten 2017. Denne våren har jeg fordypet meg i både fysikk og maskinlæring, i tillegg til hundeanatomi, og jeg har lært mye både faglig og personlig. Det har vært som en gulrot på slutten av studiet å kunne fordype seg i ett tema, med oppturene og nedturene det innebærer.

Jeg hadde ikke kommet meg gjennom denne våren med masterarbeid uten hjelp og støtte fra flere hold. Spesielt vil jeg takke Ph.d-stipendiat Bao Ngoc Huynh for å legge til rette, hjelpe og diskutere med og for meg. Hovedveilederen min Cecilia Marie Futsæther har gitt meg uunnværlig hjelp og tilbakemelding, i tillegg til gode samtaler og støtte. Biveilederen min Oliver Tomic har også vært en god støttespiller både i og utenom møter i forbindelse med masteroppgaven.

Jeg vil også takke spesialistkandidat i radiologi ved NMBU Veterinærhøgskolen Mari Nyborg Hauback for mye hjelp og gjennomgang av alt som har med veterinærmedisin å gjøre, og førsteamanuensis og internasjonal spesialist i veterinærradiologi ved NMBU Veterinærhøgskolen Hege Kippenes Skogmo for diskusjoner og gjennomgang av resultater og datagrunnlag. Denne masteroppgaven hadde ikke vært mulig uten Mari og Heges arbeid med røntgenbilder, og engasjement for prosjektet. Takk til Norsk Kennel Klub som kilde til datamateriale.

Masteroppgaven min ble også gjennomlest av Kjersti, mamma og momma, og uten deres innspill hadde den ikke nådd opp til samme nivå. Til slutt vil jeg takke gjengen på lesesalen, og ikke minst alle foreninger, grupper og kollektiv jeg har vært en del av gjennom årene mine på Ås.

> *Die reinste Form des Wahnsinns ist es,*
> *alles beim Alten zu lassen und gleichzeitig zu hoffen,*
> *dass sich etwas ändert.*
>
> - Albert Einstein

Ås, 14. mai 2023
Sunniva Elisabeth Daae Steiro

Sammendrag

Albueleddsdysplasi er en genetisk videreførbar utviklingssykdom som omfatter flere albueabnormaliteter, og kan føre til halthet hos hunder. I dag gjennomføres screening av hunder som er utsatt for albueleddsdysplasi, og dette fører til manuell analyse av i røntgenbilder til i snitt 5000 hunder per år fordelt på to veterinærradiologer i Norge. Bruk av konvolusjonelle nevrale nettverk (CNN) har vist seg å være effektivt ved bildegjenkjenning, men det finnes ikke rapportert forskning på bruk av CNNer ved diagnostisering av albueleddsdysplasi. Denne masteroppgaven utforsker derfor potensialet for bruk av CNN ved diagnostisering av albueleddsdysplasi ved screening.

Totalt ble 4617 ulike bilder av hundealbuer fra perioden 2018-2021 brukt til analyser i denne masteroppgaven. Ulike modeller i EfficientNet-familien ble prøvd ut på klassifisering av normale og abnormale prøver. For å evaluere modellenes generaliserte ytelse ble eksterne evalueringer utført på modellene med høyest helhetlig ytelse. Høyeste oppnådde nøyaktighet ved klassifisering av normale og abnormale prøver var på over 0.95, med en Matthews korrelasjonskoeffisient-score (MCC) på 0.91.

Modellen med høyest ytelse ved klassifisering av normale og abnormale prøver viste høy grad av predikerbarhet for røntgenbildene. Men videre forskning på feilprediksjoner, og ulike kombinasjoner av bilder og metainformasjon bør gjennomføres for å gjøre modellene oppnådd i denne masteroppgaven mer robuste.

I tillegg til klassifisering av normale og abnormale prøver ble andre binære samt flerklassemodeller utprøvd. Disse modellene omfattet klassifisering av prøver med sykdomsgradering 1 mot andre prøver med sykdom; prøver med artrose mot andre prøver med sykdom; prøver med sykdomsgradering nivå 1, nivå 2 og nivå 3; prøver med alle syv diagnosegrupper ved albueleddsdysplasi. I tillegg ble det prøvd ut å klassifisere tidligere feilklassifiserte og riktig klassifiserte prøver, for å se om et nevralt nettverk fant en sammenheng blant feilklassifiserte prøver.

Ingen andre modeller oppnådde like høy ytelse som normal/abnormal-modeller, men det er spesielt relevant å utforske flerklassemodeller til å skille nivå 1, 2 og 3 som verktøy ved screening av albueleddsdysplasi. Modellen med høyest ytelse ved klassifisering av nivå 1, 2 og 3 oppnådde en nøyaktighet på 0.67, og MCC-score på 0.50.

Abstract

Elbow Dysplasia is a genetically inheritable developmental disease including several elbow abnormalities, that commonly leads to disability in dogs. Today in Norway, screening of dogs prone to Elbow Dysplasia is done, which leads to manual analysis of radiographs from roughly 5000 dogs every year, done by two veterinary radiologists. The use of Convolutional Neural Networks (CNNs) is shown to be efficient in image recognition, but there are no reported studies done on the use of CNNs on radigraphic diagnostics of Elbow Dysplasia. Therefore, in this master's thesis, the potential of CNNs used for automatically diagnosing Elbow Dysplasia during screening is explored.

A total of 4617 radiographs from the period of 2018-2021 were used for analyses in this master's thesis. Different models of the EfficientNet family were tried on classification of normal and abnormal elbow samples (radiographs). To evaluate the models' generalized performance, external evaluations were executed on the models with the highest performances. The highest accuracy achieved in classifying normal and abnormal samples was 0.95, with an MCC score of 0.91.

The model that achieved highest performance in classifying normal and abnormal samples showed high predictability for radiographs. However, more research on wrongly predicted samples, and different combinations of images and meta information should be done in order to make the models achieved in this thesis more robust.

In addition to classifying normal and abnormal samples, other binary and multiclass models were examined. These models included classifying samples with Elbow Dysplasia grade 1 versus other abnormalities; samples with Arthrosis versus other abnormalities; samples with Elbow Dysplasia grade 1, grade 2 and grade 3; samples with all seven subgroups of Elbow Dysplasia. Moreover, a CNN classifying wrongly and correctly classified samples from the external evaluation of the normal/abnormal model was experimented with, in order to see if the CNN could find a correlation in misclassified samples.

No other model achieved performances as high as the normal/abnormal models, but specifically using a multiclass model classifying grade 1, 2 and 3 as a tool in screening of Elbow Dysplasia is relevant to explore further. The highest achieved performance in classifying grade 1, 2 and 3 gave an accuracy of 0.67, and an MCC-score of 0.50.

Innhold

Forkortelser

Forkortelse	Betydning
CNN	Convolutional neural network
AD	Albueleddsdysplasi
NKK	Norsk Kennel Klub
OCD	Osteokondrose dissecans
MCD	Medial Coronoid Disease
UAP	Ununited Anconeal Process
FP	Falske positive prediksjoner
FN	Falske negative prediksjoner
SP	Sanne positive prediksjoner
SN	Sanne negative prediksjoner
F1	Harmonisk gjennomsnitt av gjenkalling og presisjon
MCC	Matthews korrelasjonskoeffisient
B0-B4	EfficientNet-modellers kompleksiteter

Kapittel 1

Introduksjon

Albueleddsdysplasi (AD) er en smertefull utviklingssykdom som kan føre til halthet hos hunder [1]. Sykdommen er genetisk videreførbar, og det er derfor vanlig å screene hunder i forbindelse med avl, for å unngå spredning av gener for AD [1]. AD diagnostiseres ved hjelp av røntgenbilder, og dette arbeidet kan være tidkrevende [1]. Konvolusjonelle nevrale nettverk har vist seg å være svært effektive ved bildeanalyse [2], og i denne masteroppgaven undersøkes potensialet for bruk av CNN ved diagnostistisering av AD.

1.1 Motivasjon

I §25 i dyrevelferdsloven om avl [3] står det blant annet at det ikke skal drives avl som endrer eller viderefører dyrs gener slik at det påvirker dyrenes fysiske eller mentale funksjoner negativt. I tillegg skal det ikke avles på en måte som "reduserer dyrs mulighet til å utøve naturlig atferd, eller vekker allmenne etiske reaksjoner" [3].

Albueleddsdysplasi (AD) er en av sykdommene Norsk Kennel Klub (NKK) har særskilte regler på når det gjelder avlsforbud på registrerte hunder. Generelt arbeider NKK for at avl skal skje i ønsket retning hva gjelder blant annet rasens sunnhet [4]. I den forbindelse har NKK regler for registrering av hunder til avl, og de har utarbeidet en liste over raser med krav om kjent AD-status for at medlemmer skal kunne drive avl på disse rasene. Disse hundene omfatter berner sennen, hvit gjeterhund, labrador retriever, newfoundland og sankt bernhards, som alle er kjente raser til kjæledyrshold i Norge. Spesielt labrador retriever er populær, og var rasen med 5. flest nyregistreringer hos NKK i 2022 [5].

Hundehold er populært i Norge, og i 2022 var det siden 2010 registrert totalt 372 451 hunder hos NKK. Ikke alle norske hunder er registrert hos NKK, og det finnes ingen sikre kilder på antall uregistrerte hunder. Antall registreringer gjorde et hopp i 2021, med omtrent 10% økning i koronaåret 2020 fra 2019, og omtrent 30% økning i 2021 sammenlignet med 2019. Dette kan tyde på at mange ønsket å skaffe seg hund under pandemien med mye tid hjemme, noe som også ble fanget opp av media [6, 7]. Når etterspørselen etter hund skyter i været, prøver tilbudet å henge med, og dette kan ha resultert i mer ukritisk avl [8]. Førsteamanuensis i kirurgi ved NMBU Veterinærhøgskolen, Elena Regine Moldal, sier i en artikkel på nrk.no at de har sett flere hunder med skjelett- og leddplager etter pandemien [8].

Per dags dato finnes det ingen forskrift om avl av hund, annet enn den generelle dyrevelferdsloven nevnt over [8]. Men alle medlemmer av NKK må følge klubbens retningslinjer ved avl,

som for eksempel screening for AD hos avlshunder for spesifiserte raser. Denne prosessen koster penger og tid, og dersom hunden viser seg å ha albueleddsdysplasi kan eieren miste inntekter fra avl dersom dette er et mål. Screeningprosessen omfatter besøk hos veterinær for å ta røntgenbilder. Disse bildene sendes videre til en av to sertifiserte AD-spesialister i Norge, som er veterinærradiologer som arbeider på oppdrag av NKK [9]. Spesialisten graderer sykdom i albueleddene fra 0 til 3, hvor 0 er normal albue uten tegn til AD, og skalaen for sykdom går fra 1 til 3, der 3 er sterk grad av AD [10].

Forskjellene mellom normale albueledd og grad 1 er små, og noen ganger svært vanskelig å skille mellom, også for veterinærradiologene. Siden AD er ansett som en genetisk videreførbar sykdom, kan selv en registrert grad 1 AD føre til tap av muligheter for avl på hunden. Hundeeiere er dermed svært opptatt av å ikke få et "ufortjent stempel", og det kan oppstå situasjoner med saksøking dersom det viser seg at hunden er registrert med AD ved en feilavlesning [11]. Dersom feilavlesningen slår ut andre veien, kan dette resultere i avl på hunder som viderefører en sykdom som reduserer livskvaliteten til hunden.

Det vil alltid forekomme inter- og intravariabilitet i bildediagnostiske vurderinger, og dette kan også forekomme i forbindelse med albueleddsdysplasi. Intravariabilitet er variasjoner i ytelse for en enkelt person på samme oppgave [12], i dette tilfellet veterinærradiologen, mens intervariabilitet er fenomenet der ulike radiologer kan ha variasjoner i diagnostisering av samme bilder.

En annen utfordring med screeningprosessen er tiden det tar for AD-spesialistene å gå gjennom alle henvendelser, som ligger på mellom 4500 og 5500 i året. Hver hund tar anslagsvis 5 minutter å screene i gjennomsnitt [11], som summeres opp til omtrent 417 timer for 5000 hunder. Dette betyr at over 11 37,5-timers arbeidsuker kreves i gjennomsnitt hvert år, bare til analyse av røntgenbilder i forbindelse med AD-screening. Disse tallene er grove anslag, og det tas forbehold om at anslagene kan avvike.

1.2 Dyp læring i human- og veterinærmedisin

Et kjapt søk på google's søkemotor google.com, med søkeordet “deep learning” gir per 27. mars 2023 122 000 000 treff for publiseringer fra kun 2023. Med eksempler som setningsfullføring, oversetting, Apples Siri og Amazons Alexa, og ikke minst ChatGPT, er det tydelig at verden har fått øynene opp for bruk av kunstig intelligens. Dette gjelder også i akademia, og samme søk på googles søkemotor for akademisk litteratur, google scholar[1], ga 17 200 resultater.

Selv om det finnes forskning på bruk av dyp læring innenfor humanmedisin og veterinærmedisin, spesielt innen radiologi [13], er det ikke kjent at det finnes forskning på dette innenfor screening og diagnostisering av AD. Flere eksempler på forskning på maskinlæring innenfor bildediagnostisk veterinærmedisin er sammenfattet i et litteraturstudie av Hennessey et al. fra 2022 [14]. Eksempler på slike studier er klassifisering av lungeabnormaliteter, og å skille mellom normale og forstørrede hjerter (kardiomegali) [14].

En av studiene gjort i 2021 av McEvoy et al. [15] på prediksjon av tilstedeværelse av hofteleddsdysplasi (HD) ga en presisjonsscore på 0.91, og en sensitivitetsscore på 0.53. Studien konkluderte med at konvolusjonelle nevrale nettverk (CNN) er anvendelige på veterinære avbildninger, og at modellene oppnådd i studien har potensiale til å være et verktøy ved screening av HD. Hofteleddsdysplasi er i likhet med AD en utviklingslidelse som NKK jobber aktivt mot

[1] https://scholar.google.com/

å spre genetisk [16], og med stadig utvikling i maskinlæringsteknologi, er det grunn til å utforske potensialet til dyp læring som verktøy også ved screening av AD i hunder.

Også innen humanmedisin finnes det studier som både bruker objektdeteksjon og klassifisering av abnormaliteter i ben fra røntgenbilder. Eksempler på dette er deteksjon av lesjoner (skade/-endring) i ben i lemmer og hofter av Regnard et al. fra 2022 [17], og klassifisering av effusjon i albueledd som følge av brudd, gjennomført i 2022 av Huhtanen et al. [18]. Disse to studiene trente flere modeller, og beste oppnådde sensitivitet var hhv. 0.98 og 0.89. Et annet studie fra 2019 av Rayan et al. [19] brukte flere bilder av samme albue til å predikere abnormalitet (brudd), for å etterligne radiologers tilgang på flere bilder av samme pediatriske pasient ved diagnostisering. Dette studiet oppnådde nøyaktighet på 0.88, og sensitivitet og spesifisitet på hhv. 0.91 og 0.84 på prediksjonene [19].

Studier på bruk av maskinlæring ved bildediagnostikk viser optimistiske resultater, og med dagens utvikling av maskinlæring kan det forventes fremskritt på bruk av maskinlæringsteknologi som verktøy i diagnostisering i både human- og veterinærmedisin. Det er altså vist at høy ytelse oppnås med dyp læring, og muligheten er dermed til stede for å effektivisere flere deler av helsevesenet [14].

1.3 Formål

Formålet med denne masteroppgaven er å undersøke potensialet til konvolusjonelle nevrale nettverk (Convolutional Neural Network, CNN) til binær klassifisering av albueleddsdysplasi (abnormaliteter) fra røntgenbilder av hundealbuer. Modellene brukt i denne oppgaven var EfficientNet med forhåndstrente vekter [20]. EfficientNet er en familie av CNNer som oppnår state-of-the-art nøyaktigheter, med høyere effektivitet enn sammenlignbare modeller, ved hjelp av systematisk, helhetlig oppskalering av en referanse-CNN [20].

I tillegg til den binære normal/abnormal-modellen ble flere andre modeller undersøkt, blant annet klassifisering av AD-grad 1 mot resten, og å skille mellom nivå 1, 2 og 3. Resultatene til noen av modellene i denne masteroppgaven ble analysert inngående, og modellen med høyest helhetlig ytelse ble testet på et større usett datasett, og trent videre for å se etter forbedringspotensialer.

Datasettene brukt til trening, validering, testing og evaluering av EfficientNet-modeller til klassifisering av normale og abnormale albuer i denne masteroppgaven besto til sammen av 4617 røntgenbilder av hundealbuer fra NKKs medlemmer. 1090 tilfeldig valgte bilder ble brukt til opptrening, validering og testing av normal/abnormal-modellen, med forholdet 2:1:1 for trenings-:validerings-:testsett. Resterende 3527 røntgenbilder ble brukt som eksternt datasett.

1.4 Oppsett av masteroppgave

I denne masteroppgaven presenteres først relevant teori om røntgenavbildning, albueleddsdysplasi og maskinlæring i kapittel 2. Kapittel 3 introduserer datasettet brukt i oppgaven, i tillegg til arbeidsmetoder. Resultater knyttet til alle utprøvde problemstillinger for klassifisering av albueleddsdysplasi presenteres i kapittel 4, mens analysen diskuteres i kapittel 5. I kapittel 6 trekkes en konklusjon fra analysen, og forslag til videre arbeid presenteres. Siste del av masteren inneholder referanser og to vedlegg.

Kapittel 2

Teori

2.1 Røntgen

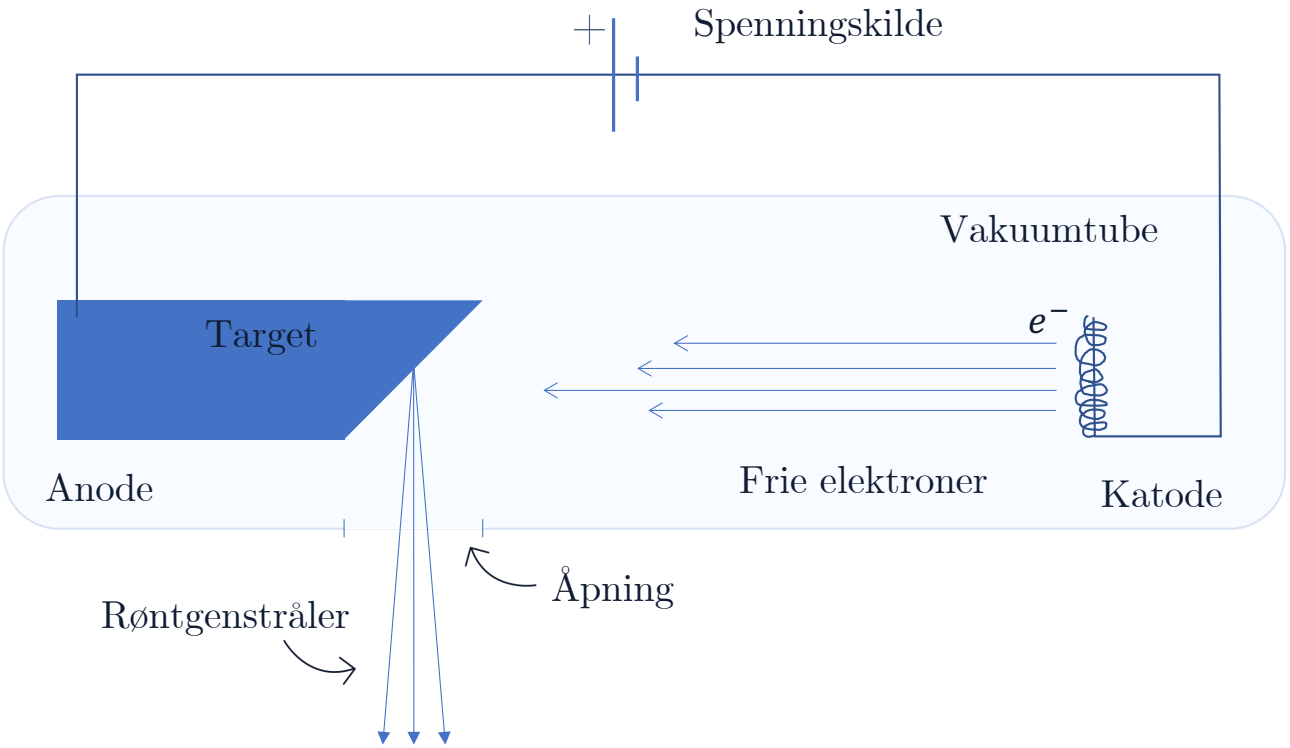
Røntgenstråler ble oppdaget av Wilhelm Röntgen i 1895, og trolig først brukt i klinisk sammenheng i 1896 [21]. I dag brukes røntgen fortsatt mye i klinisk sammenheng, men bruken strekker seg også ut til andre områder, som for eksempel bagasjegjennomgang eller for å detektere fremmedlegemer i matvarer [22].

Røntgenstråler er, i likhet med synlig lys, elektromagnetisk stråling. Synlig lys har bølgelengder tilsvarende energier på mellom 1.5 og 3 eV, mens røntgenstråling har energi grovt sett mellom 1 og 100 keV [22]. Røntgenstråling har så høye energier at den kan endre et atoms netto ladning, og regnes derfor som *ioniserende stråling*. Andre typer ioniserende stråling er partikkelstråling og gammastråling (γ-stråling), og all stråling over omtrent 10 keV regnes som ioniserende [23].

I likhet med røntgenstråler er også gammastråler elektromagnetisk stråling, men forskjellen ligger i hvordan strålingen oppstår og energinivået på strålingen ($\sim$100 keV - 1 MeV for γ-stråling). Gammastråling oppstår som energi frigitt fra ustabile atom*kjerner* gjennom desintegrasjonsprosesser, mens røntgen er energi frigitt fra *elektronorbitalen* til, og det elektriske feltet rundt atomer [23, 24].

2.1.1 Røntgenkilde

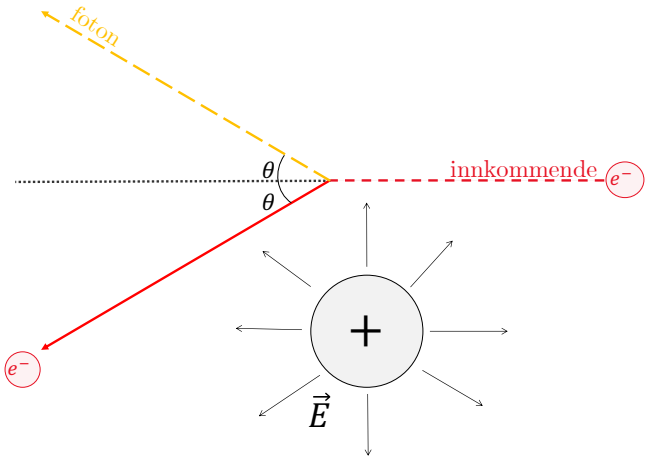
Generering av røntgenstråler gjøres i en vakuumtube lik skissen i figur 2.1. Tuben inneholder en krets med påsatt DC-spenning, en oppvarmet kveil (katoden) og et target (anoden), som er materialet som legger til rette for å generere røntgenstråling. Ved å varme opp katoden frigjøres elektronene, og på grunn av spenningsforskjellen mellom anoden og katoden, slynges de gjennom det evakuerte mellomrommet mellom anoden og katoden ved hjelp av det resulterende elektriske feltet [23]. For konvensjonelle røntgenkilder ligger denne spenningen mellom 25 og 150 keV [23]. De termiske elektronene vekselvirker med anoden (target), og mesteparten av energien til elektronene overføres til target som varme [24]. En måte å avlede varmen i target på, er å la target rotere for å spre varmelasten fra den innstrålte elektronstrålen [23].



Figur 2.1: *Skisse på vakuumtube brukt til generering av røntgenstråler. Tegnet med inspirasjon fra figur i Fundamental Physics for Probing and Imaging av Wade Allison (2006) [23].*

I tillegg til avsatt varme i target, genereres også røntgenstråler, og disse kan oppstå i to former: bremsestråling og karakteristisk stråling (*fluorescence x-rays*) [22].

Bremsestråling oppstår når et elektron møter et atom, og bremses ned når banen til elektronet avbøyes på grunn av det elektriske feltet rundt kjernen til atomet (Coulombvekselvirkning, se figur 2.2). Elektronets retning har altså endret seg, og for at energien skal være konservert, sendes det ut et foton som kompenserer for endringen i bevegelsesmengden til elektronet. Jo nærmere elektronet kommer kjernen når det vekselvirker, desto høyere blir akselerasjonen, og dermed også energien som gis til fotonet. Spekteret til avgitt røntgenstråling i form av bremsestråling er kontinuerlig, og den høyest mulige energien til røntgenstrålene er lik den høyest mulige energien til elektronet [22]. Det vil si at maksimal energi på røntgenstrålen er gitt av tubens spenning [24].

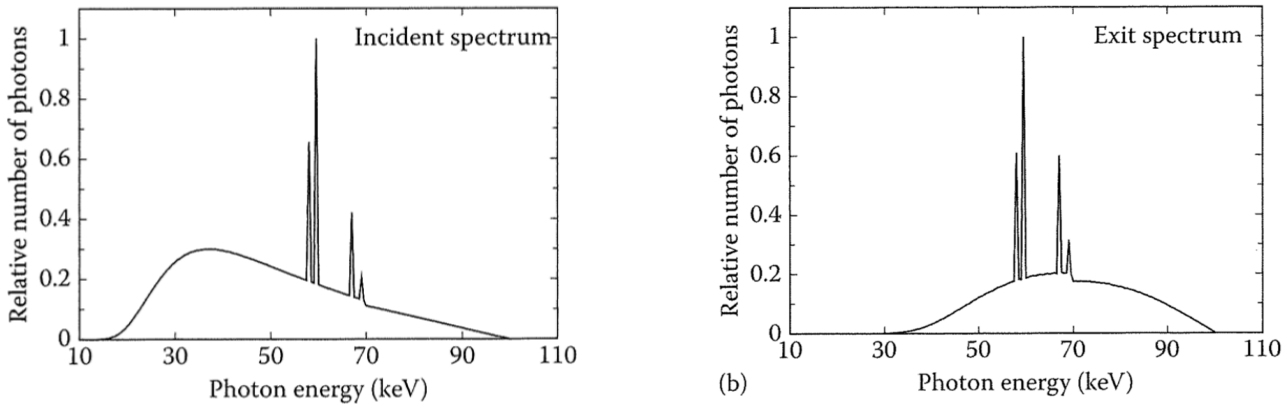


Figur 2.2: *Når et elektrons bane passerer et elektrisk felt vil banen bøyes av (akselereres) på grunn av Coulombkraften. På grunn av bevaring av energi, vil energien i systemet være lik før og etter vekselvirkningen, og dermed sendes det ut energi i form av et foton i samme vinkel, motsatt vei, fra elektronets opprinnelige bane.*

Karakteristisk stråling oppstår når et høyenergetisk elektron treffer et elektron i de innerste skallene til et atom, slik at atomet blir ionisert [23]. Ett av elektronene i de øvre skallene deeksiterer da til den ledige plassen i innerste skall, og dermed avgis en mengde energi i form av et foton. Mengden energi som frigis vil være unik for alle atomer med ulike strukturer, og derfor kalles denne typen stråling *karakteristisk* stråling [22].

Figur 2.3 viser energispekteret til røntgenstråling fra wolfram, som er det vanligste target-materialet i klinisk røntgenutstyr [22]. Grunnstoffet er spesielt egnet som target på grunn av det høye smeltepunktet, i tillegg til den høye Z-verdien på 74. Sannsynligheten for vekselvirkning mellom elektroner og atomkjerner er sterkt avhengig av atomkjernens størrelse [25], som kan vises av ligning 2.1 for lineær energiavsetning i form av bremsestråling [24]. I ligningen står N for target-materiets atomtetthet, E er elektronets kinetiske energi, m_e er massen til elektronet, og c er lysets hastighet. Z er antall protoner i targetkjernene, og funksjonen $F(E, Z)$ er avhengig av Z og sterkt avhengig av E, men varierer lite for energier opp til 1 MeV [24].

$$(-\frac{dE}{ds})_{brems} = N(E + m_e c^2)Z^2 F(E, Z) \tag{2.1}$$



Figur 2.3: *Røntgenspektrum for en røntgentube med Wolframtarget; 100 kV spenning med 2.5 millimeters aluminiumfilter for å absorbere stråler med lavere energier før de treffer pasienten. Spekteret er vist (a) før og (b) etter energiavsetning i 18.5 cm mykt vev i tillegg til 1.5 cm bein. Figurene er hentet med tillatelse fra Webb's Physics of Medical Imaging av M. Flower (2012) [21].*

I figur 2.3a er det tydelig at intensiteten til stråling fra den lavere halvdelen av energispekteret er høyest før energiavsetning, og dette kommer av to fenomen. Elektronstrålen fra anoden består av elektroner med et spekter av energier opp til tubens maksimale spenning. I tillegg vil elektroner vekselvirke flere ganger i target, med litt lavere energi etter hver vekselvirkning. Dermed vil et spekter av energier tilhørende røntgenstrålene oppstå, med flere stråler med lavere energi.

For røntgenspekteret etter energiavsetning i vev er energier lavere enn 30 keV så og si ikke til stede, som vist i figur 2.3b. Dette viser at elektroner med energier på det lavere intervallet har større sannsynlighet for å vekselvirke med vevet enn elektroner med høyere energi.

Intensiteten, $I(x)$, til røntgenstrålen er summen av den totale mengden energi fra alle enkeltstråler, og gitt i ligning 2.2 [24]. $I(x)$ vil avta langs strålingens bane, og er derfor en funksjon av x, strekningen tilbakelagt av strålen. μ er den lineære absorpsjonskoeffisienten, et mål på sannsynligheten for vekselvirkning mellom fotoner og materie langs strekningen x [21].

$$I(x) = I_0 e^{-x\mu} \tag{2.2}$$

Fotoner med lav energi har større sannsynlighet for å bli absorbert i materier på grunn av den fotoelektriske effekten [23]. Dette er et fenomen der fotoner i nærheten av atomkjerner kolliderer med elektroner, og all energien til fotonet overføres til elektronet [26]. Sannsynligheten for fotoelektrisk effekt er proporsjonal med $Z^5/E^{3.5}$, der Z er atomnummeret til kjernen, og E er fotonets energi [26]. Dette viser at sannsynligheten er sterkt avhengig av både Z og E, og dermed vil lavenergetiske røntgenstråler absorberes av pasienten, og ikke bidra til bildeformeringen. Derfor filtreres stråler med lavere energi før de treffer pasienten, for eksempel med aluminium [21]. Figur 2.3 viser resultatspekteret etter at det meste av fotonene med energier under 20 keV er filtrert bort for wolfram-targets, mens ved mammografi filtreres stråler over 20 keV bort [24].

Selv om røntgenstråling er ioniserende stråling, bidrar ikke røntgenbilder av ledd til farlige doser på pasienten. En gjennomsnittlig røntgenundersøkelse på lemmer og ledd gir en dose på

omtrent 0.001 mSv, mens til sammenligning eksponeres en person fra Storbritannia i snitt for omtrent 0.007 mSv hver dag bare fra bakgrunnsstråling fra naturen [23].

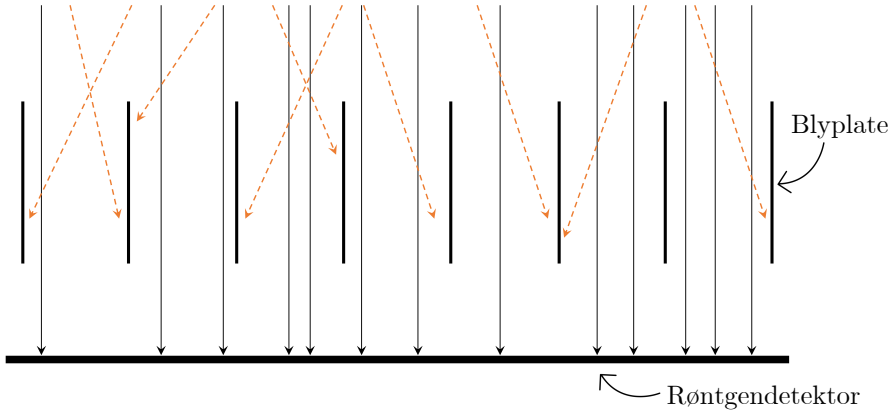
Generelt for røntgenbilder oppstår både bremsestråling og karakteristisk stråling når et grunnstoff med moderat Z-verdi bombarderes med en DC-strøm av elektroner, se figur 2.3. Røntgenstrålebehandling mot kreftsvulster og røntgenavbildninger ved mammografi krever hhv. høyere (opp til 50 MV) og lavere spenninger (30 kV) på strømkilden [23, 24].

2.1.2 Røntgenbilder

Røntgenbilder dannes ved å sende røntgenfotoner gjennom en pasient, for så å måle gjennomstrålingen med en fotondetektor [21]. Den målte intensiteten på fotondetektoren I, er et resultat av vekselvirkning på vei gjennom pasienten, slik ligning 2.2 viser. Fotonene som går gjennom pasienten kan enten være primære eller sekundære, som betyr fotoner henholdsvis uten og med interaksjon med det bestrålte legemet [21].

Det er hovedsakelig primærfotoner som gir informasjon på detektoren, fordi det kun er disse som når fram til detektoren i en kollimert stråle som står perpendikulært på detektoren [21]. Siden sekundære fotoner har vekselvirket med materiet vil de ikke følge den opprinnelige banen til røntgenstrålen, og dermed når de detektoren med en vinkel. Disse vil bidra til høyere signal på områder som ikke tilsvarer fotonets originale bane, altså bidrar disse strålene som tilfeldig støy på bildet for øvrig [21].

Siden røntgenbilder analyseres med fokus på kontraster, brukes et anti-spredning-raster mellom pasienten og detektoren, slik at sekundærfotoner ikke når detektoren i det hele tatt [24]. Rasteret kan for eksempel bestå av parallelle remser av bly, som vil absorbere de uønskede strålene [21]. Remsene i rasteret plasseres perpendikulært på detektoren, slik at parallelle stråler slipper gjennom, mens stråler med vinkler treffer platene, se figur 2.4 [24].



Figur 2.4: *Et raster av blyplater perpendikulært på røntgendetektoren absorberer sekundærfotoner (oransje, stiplede) fra å nå detektoren, mens primærfotoner (svarte, heltrukne) har fri bane.*

Signalet til røntgenstrålen på detektoren avgjør opasiteten på røntgenbildet [21]. Det vil si at ved lav opasitet, eller høy gjennomsiktighet, vil mange stråler gå gjennom legemet uten vekselvirkning, og dermed nå helt fram til detektoren. Sannsynligheten for vekselvirkning vil avhenge av summen av vekselvirkning mellom fotonet og hver av vevstypene det går gjennom [21]. Sannsynligheten for vekselvirkning mellom røntgenstråler og et spesifikt materie er

gitt ved den lineære absorpsjonskoeffisienten μ for materiet. Absorpsjonskoeffisienten omfatter absorpsjon ved fotoelektrisk effekt, comptonspredning og pardannelse, og summen av alle absorpsjonskoeffisientene til ulike vev langs fotonets bane avgjør opasiteten på røntgenbildet [21].

For eksempel for benvev er det stor sannsynlighet for at fotonene vekselvirker med atomene i benet, og dermed når nesten ingen røntgenstråler detektoren. Ben består av mye kalsium, som har atomnummer 20, og en mye høyere absorpsjonskoeffisient enn andre vanlige grunnstoff i vev som karbon og oksygen. Derfor vil benvev vises som veldig hvitt på røntgenbilder i energispekteret 15-150 keV, mens karbon gir mørke signaler siden det trenges lettere gjennom av fotonene, slik som figur 2.5 viser [23].

Figur 2.5: *Eksempel på et røntgenbilde. Dette er et røntgenbilde av en normal hundealbue. Legg merke til at vev med høy lineær absorpsjonskoeffisient, slik som ben, er hvitere på bildet, altså har det høyere opasitet. I kantene har røntgenstrålene ikke gått gjennom hunden, og dermed er så å si all stråling fra kilden i dette området fanget opp av detektoren.*

2.2 Hundealbue

Alle hunder i Norge kan registrere seg hos Norsk Kennel Klub (NKK). NKK er en interesseorganisasjon for hundeeiere og -avlere i Norge, og dyrevelferd står i fokus i reglene til klubben. Et svært viktig tema for NKK er derfor sunn hundeavl, og organisasjonen har fokus på å spre informasjon om blant annet anatomien til hunden [27]. Medlemmer av NKK kan registrere hunden sin i databasen til NKK, og for noen raser skal utvalgte sykdommer registreres som til stede eller ikke [28]. Dersom hundeeiere oppgir hundens arvelige og ikke-arvelige sykdommer, gjør det det enklere for avlsinteresserte å velge hunder uten genetisk arvelige belastninger til avlsarbeid.

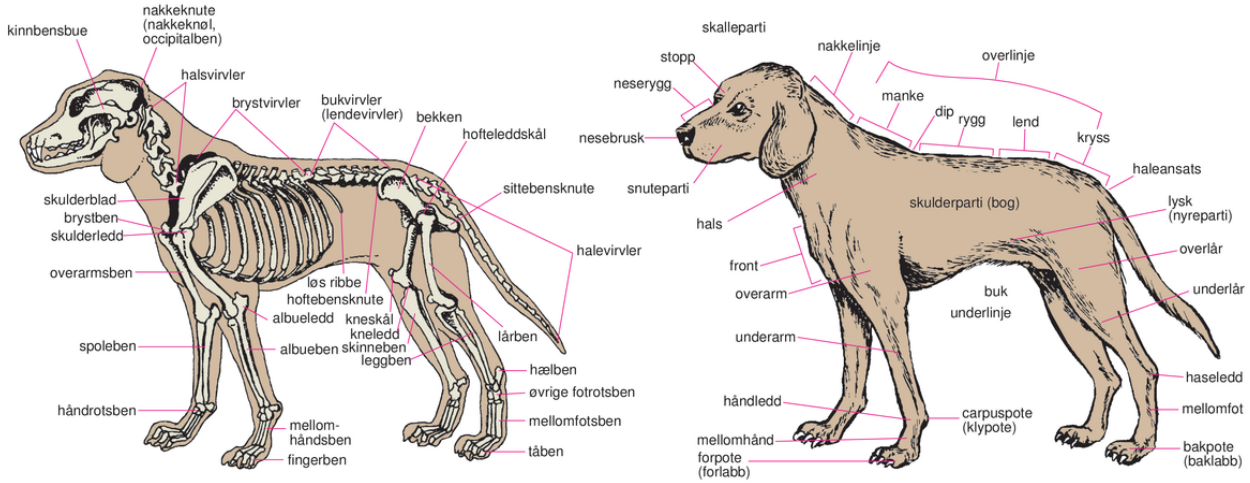
Ett av fokusene ved avlsarbeid i samarbeid med NKK er albueleddsdysplasi (AD). AD er en fellesbetegnelse for utviklingslidelser som ofte leder til artrose, smerter og halthet hos hunder [1]. Etter at AD ble anerkjent som en genetisk videreførbar lidelse av International Elbow

Working Group i 1993, ble det innført screening av predisponerte raser for AD [29, 1].

Screening for AD utføres ofte med røntgenundersøkelser, selv om det også kan påvises med andre metoder, blant annet CT [30, 31]. I NKK kan kun sertifiserte veterinærer utføre screening for AD, og det finnes bare to slike spesialister i Norge per dags dato. Ved screening av albueleddsdysplasi i Norge følges standardprotokollen fra 2017 satt av International Elbow Working Group[1] [11]. Screening er i følge NKK en "undersøkelse av et stort antall dyr i en rase uavhengig av kliniske symptomer", og utføres derfor utenom vanlig klinisk behandling av sykdom hos enkelthunder [32].

2.2.1 Albueleddsdysplasi

Albueleddet er et av de største leddene i hundens framben, se figur 2.6.



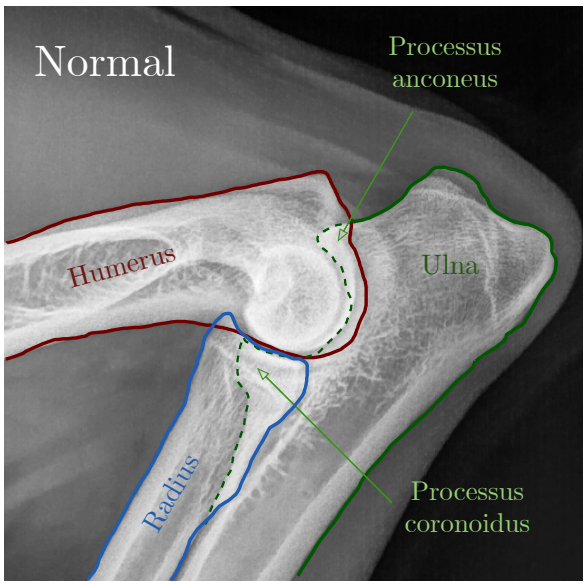
Figur 2.6: *Hundens anatomi. Albueleddene til hunden befinner seg på begge frambena. Gjengitt med tillatelse fra Store norske leksikon [33]*

Hundens albueledd består av tre hoveddeler, nemlig ulna, humerus og radius, som til sammen utgjør albueleddet, se figur 2.7a. Humerus har to parallelle, runde "hoder" på enden, som møter radius og ulna som ligger parallellt med hverandre i underarmen. Disse hodene kalles *kondyler*. Ulna er den lengste av de to benene som utgjør underamen.

Blant abnormalitetene som inngår under paraplybetegnelsen albueleddsdysplasi, er følgende utviklingslidelser inkludert: Ununited Anconeal Process (UAP), Osteokondrose Dissecans (OCD) og Medial Coronoid Disease (MCD) [30]. Disse forstyrrelsene fører ofte til artrose og/eller sklerose, og det er derfor vanlig med en kombinasjon av AD og sklerose og/eller artrose [29, 34]. AD er registrert videreført hos mellom 0 % og 55% av hundepopulasjonen, avhengig av rase, screeningmetoder og hvilken hundepopulasjon man ser på [1]. Eksempler er 17% AD hos labradorer i Storbritannia, og 70% hos berner sennen-hunder i Nederland [35].

Det kan være vanskelig å oppdage AD ved hjelp av røntgen, da noen av sykdommene kan ligne naturlige variasjoner i benbygningen. Derfor sammenlignes ofte høyre og venstre albue for å oppdage sykdom [11]. Det er imidlertid vanlig at begge albuene er rammet av lesjon (sykdom), og i tillegg kan flere av lidelsene i AD-kategorien opptre samtidig i hundens albueledd [30].

[1]Se vet-iewg.org/ for oversikt over protokoller.


Normal
Processus anconeus
Humerus
Ulna
Radius
Processus coronoidus

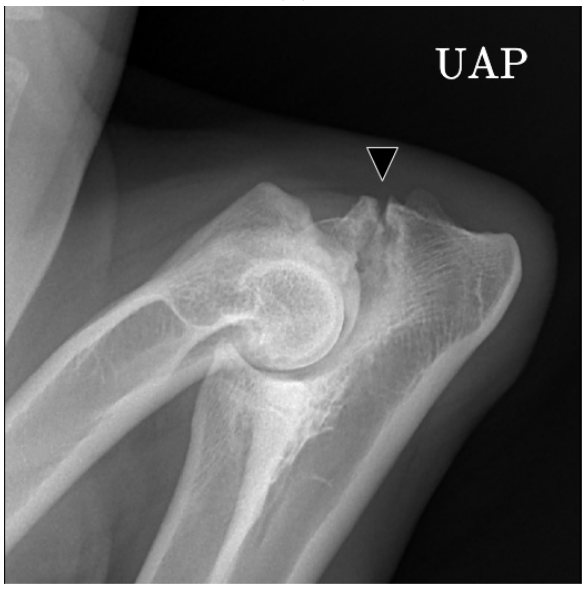
(a)


MCD

(b)


OCD

(c)


UAP

(d)


Artrose

(e)


Sklerose

(f)

Figur 2.7: *Bilde på forrige side. Karakteristiske tegn til sykdom er markert med pilhoder. (a) En hundealbue uten sykdom. Albuen består av tre ben: ulna (grønn), radius (blå) og humerus (rød). Processus anconeus og prossesus coronoideus på ulna er markert med piler, og ligger plassert med overlapp fra hhv. humerus og radius. (b) MCD ses ved overlappen mellom ulna og radius, hvor processus coronoideus vanligvis har en karakteristisk trekantform på friske hundealbuer. (c) OCD kan ses som en liten del langs kondylen med lavere opasitet enn kondylen ellers. (d) UAP ses på enden av ulna som en linje med lavere opasitet enn omkringliggende vev. Dette kommer av at benet ikke er grodd sammen. (e) Artrosen ses her som påleiringer i gropa mellom ulna og humerus, samt påleiringer langs oversiden av humerus. (f) Sklerose i hundealbuen vises som høyere opasitet enn normalt vev, slik at man ikke ser trabeklene som skal være til stede i normalt benvev. Legg merke til den høye opasiteten hvor ulna møter kondylen fra undersiden.*

Røntgenbilder av normale hundealbuer viser en tydelig trekantformet kontur der ulna overlapper radius, og dette kalles processus coronoideus, se figur 2.7a [31]. Hunder med **Medial Coronoid Disease** viser en utydelig, flatere, avrundet, fragmentert eller spredt linje på samme sted på røntgenbilder [31]. Et eksempel på en hundealbue med tegn til MCD er figur 2.7b.

MCD rammer for det meste store hunder, men kan også oppstå i små hunder. Dette er den hyppigst forekommende av de tre utviklingssykdommene, med forekomst i over 96% av tilfellene diagnostisert med AD [1]. Gjennomsnittlig alder for diagnostisering av hunder med MCD er 13 måneder, men kliniske tegn til lidelsen kan oppstå så tidlig som i en alder av 4 måneder.

Osteokondrose dissecans i albuen er en sykdom som fører til svikt i benformeringsprosessen i albueleddet, forårsaket av nekrotisk (død) brusk [36]. OCD vises på røntgenbilder som en liten del av kanten på kondylen med lavere opasitet en omkringliggende vev, se figur 2.7c. I likhet med UAP, forekommer OCD i store hunderaser, men kliniske tegn på lidelsen oppstår vanligvis ved mellom 6 og 9 måneder gamle hunder [30]. Av alle hunder med AD, forekommer OCD i mellom 3 % og 25 % av tilfellene [1].

En bit av ulna i albueleddet kalles processus anconeus (figur 2.7a), og hos noen hunder vokser denne biten separat fra resten av ulna når hunden er valp [30]. Dersom processus anconeus ikke har vokst sammen med ulna innen hunden er ca 5 måneder, har hunden tilstanden **Ununited Anconeal Process** [1]. UAP kan vises på røntgenbilder som en linje med lav opasitet mellom processus anconeus og tuppen av ulna, som følge av at det ikke er ben der (se figur 2.7d) [31].

UAP rammer store hunder, og hannhunder er rammet omtrent dobbelt så ofte som tisper [37]. Raser med høyere risiko for UAP er berner sennen, mastiffer, rottweilere og sankt bernhards, men det forekommer også i andre raser [30].

En vanlig sekundærlidelse ved albueleddsdysplasi er **artrose**, altså opptrer ofte artrose sammen med en eller flere av utviklingslidelsene nevnt over [1]. Artrose er irreversibel, og oppstår i de fleste tilfeller av AD. Opp mot 20 % av hundepopulasjonen over ett år ble estimert rammet av artrose i en studie i 1996 [38].

Artrose ses radiografisk som påleiringer av benvev rundt ledd, se figur 2.7e. Medfølgende plager er stivhet, smerte og innskrenket bevegelighet, og følgelig er tidlig artrose en uønsket sykdom. Observerbare tegn til artroseplager varierer fra milde og periodiske, til alvorlige og vedvarende plager [1].

En annen typisk sekundærlidelse ved AD er **sklerose**. Sklerose er en betegnelse på fortykkelse av vev i organer grunnet forkalkning, i dette tilfellet vil dette si fortykkelse av bein [39]. Dette vises som høyere opasitet enn normalt/omkringliggende vev på røntgenbilde av hundealbuen,

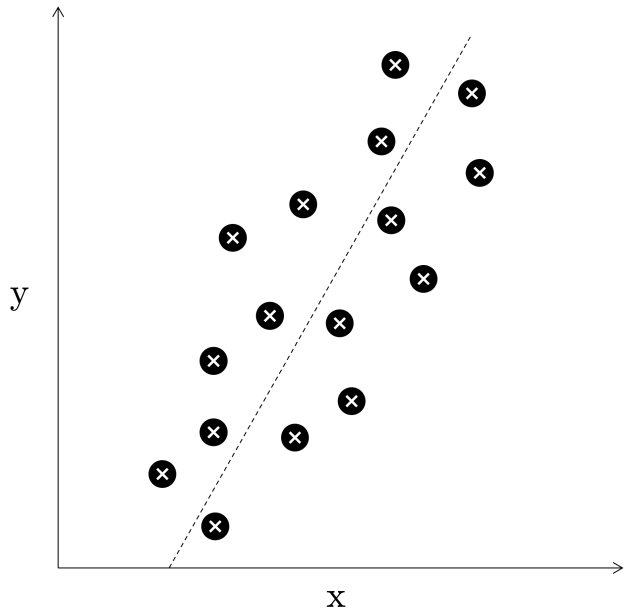
se figur 2.7f. Friskt vev vil vanligvis ha et karakteristisk kryssmønster kalt *trabekler* synlig på store deler av benet. Ved artrose forsvinner dette mønsteret i opasiteten, siden området med artrose er tykkere og mer kompakt.

2.3 Maskinlæring

Kunstig intelligens oppsto på 1950-tallet, og senere en underklasse av kunstig intelligens kalt *maskinlæring* [2]. Med økende ytelse på datamaskiner øker også mulighetene innenfor blant annet maskinlæring. Maskinlæring er et felt som omhandler å bruke algoritmer til å hente ut informasjon fra data [40]. Blant annet kan ukjent, eller "gjemt", informasjon hentes ut av data ved hjelp av ikke-veiledet (unsupervised) læring [40]. De to andre undergruppene av maskinlæring kalles veiledet (supervised) og forsterkende (reinforcement) læring. Veiledet læring bruker kjent informasjon om tidligere hendelser til å predikere utfall på senere hendelser, mens forsterkende læring bruker resultat etter interaksjoner til å forbedre egen ytelse, som for eksempel ved å vinne eller tape i sjakk [40].

2.3.1 Veiledet maskinlæring

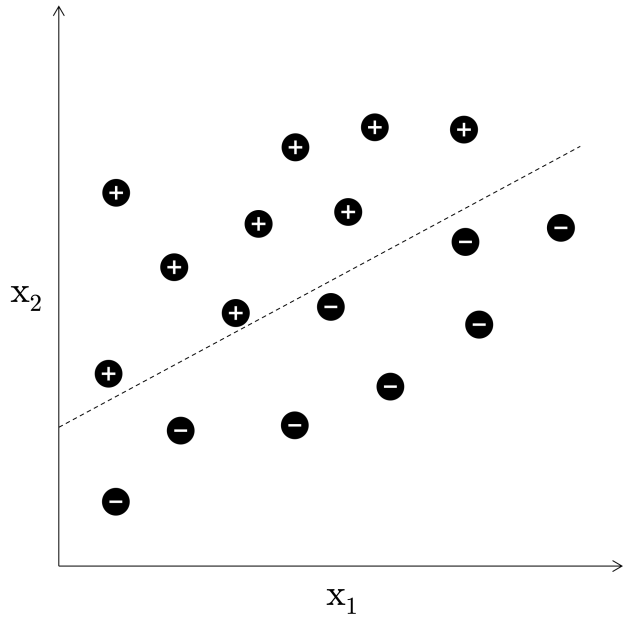
Innen veiledet maskinlæring finnes to hovedkategorier på problemstillinger innen prediksjoner: regresjon og klassifisering [40]. Innen regresjon er prediksjonen et tall på en kontinuerlig skala, for eksempel prisen på en bolig ved gitte omstendigheter (antall kvm, etasjer, byggår, osv). En vanlig modell som brukes i regresjonsproblemer er lineær regression, se figur 2.8.



Figur 2.8: *Et eksempel på et regresjonsproblem som predikerer nye prøvers y-verdi ut ifra gitt x-verdi. Her er et eksempel med en kontinuerlig variabel (x-verdi) og den kontinuerlige predikerte verdien (y-verdi). Basert på Raschka og Mirjalili [40].*

Ved klassifisering er derimot et begrenset antall klasser definert, og modellen skal plassere hver prøve i datasettet i riktig klasse, også kalt *målklasse*. Det mest grunnleggende eksemplet på klassifisering er binære problem der man skiller mellom klasse 1 (positiv) og klasse 0 (negativ).

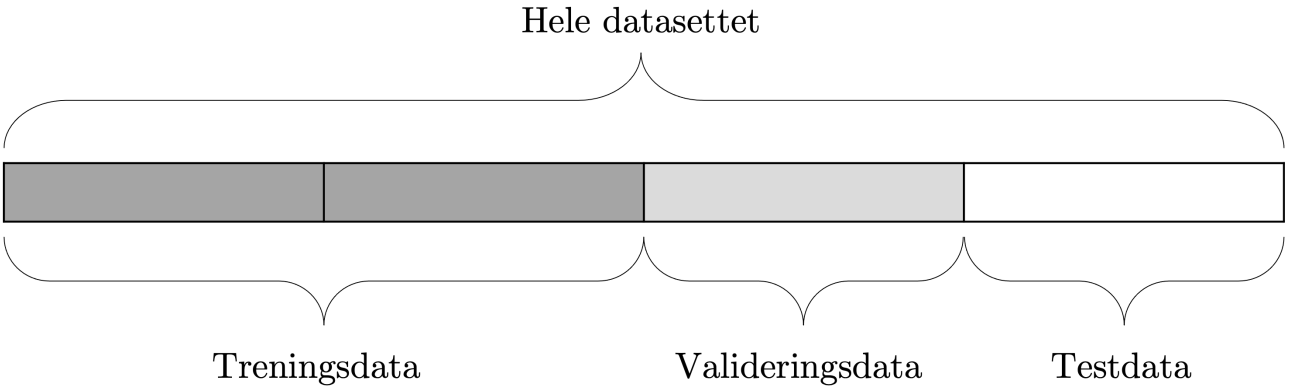
Et klassisk eksempel er å klassifisere en prøve som kvinne eller mann basert på for eksempel høyde og vekt, se figur 2.9 for skisse på hvordan klassifiseringen kunne sett ut.



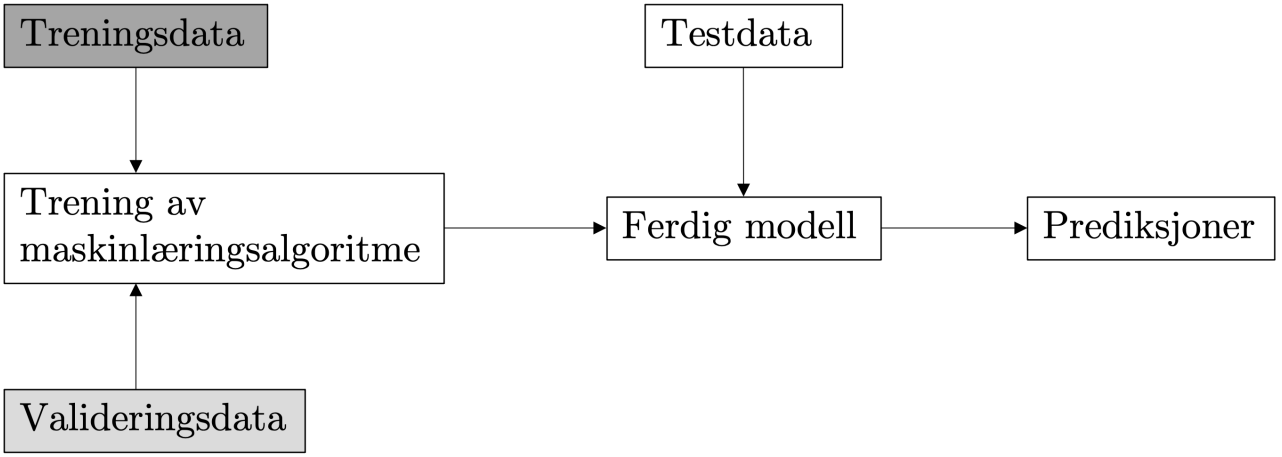
Figur 2.9: *Et eksempel på klassifiseringsproblem som separerer positiv og negativ klasse fra hverandre basert på egenskap x_i. I dette tilfellet vises to egenskaper, x_1 og x_2. Alle prøvene til høyre for hyperplanet blir predikert å tilhøre negativ klasse, mens prøver til venstre blir predikert å tilhøre positiv klasse. Basert på Raschka og Mirjalili [40].*

Det er også mulig å lage en modell som grupperer mellom flere enn en klasse, dette kalles et flerklasseproblem. Et eksempel er å predikere hvilket dyr hver prøve tilsvarer blant målklassene *hund, katt og hest*. Maskinlæringsmodeller kan bare håndtere tall, ikke ord, og dermed ville målklassene bli oversatt til klasse 0, 1, og 2 i dette tilfellet.

Måten modellen kan skille alle prøvene fra hverandre, uansett hvilket type problem det gjelder, er å gi modellen informasjon om faste *egenskaper* (features) til hver av prøvene [40]. Hver prøve har altså samme sett med egenskaper, men med ulike verdier for hver egenskap, som ofte er karakteristisk for målklassen prøven tilhører. Modellen trenes på et gitt antall prøver kalt *treningsdata*, med alle egenskapene i tillegg til de riktige målklassene [40]. Etter trening testes modellen på nye data på samme form som treningsdataene, kalt *testdata*. Testdata har ikke målklassene inkludert i datasettet, men disse brukes til å teste ytelsen til modellen etter trening. I tillegg kan man ha *valideringsdata*, som har likt oppsett som testdata, men som brukes for å validere modellen underveis i treningen. Figur 2.10a viser oppdeling av datasett, mens figur 2.10b viser flyten i trening, testing og validering av en maskinlæringsmodell.



(a) *Et helt datasett deles ofte opp i 3 deler, hvor for eksempel* 2/4 *er treningsdata og valideringsdata og testdata kan utgjøre* 1/4 *hver av det totale datasettet.*



(b) *Treningsdata og valideringsdata brukes ved treningen og valideringen av en maskinlæringsalgoritme. Dette resulterer i en prediksjonsmodell som tar inn ukjent testdata og predikerer målklassen til alle prøvene i testdatasettet.*

Figur 2.10

Den konvensjonelle maskinlæringsmodellen tar inn tabelldata hvor hver rad tilsvarer en prøve ($\boldsymbol{x}^{(i)}$), og hver kolonne tilsvarer en egenskap (x_j). I eksemplet med hund, katt og hest, kan dette være egenskaper som vekt og høyde, som i tabell 2.1.

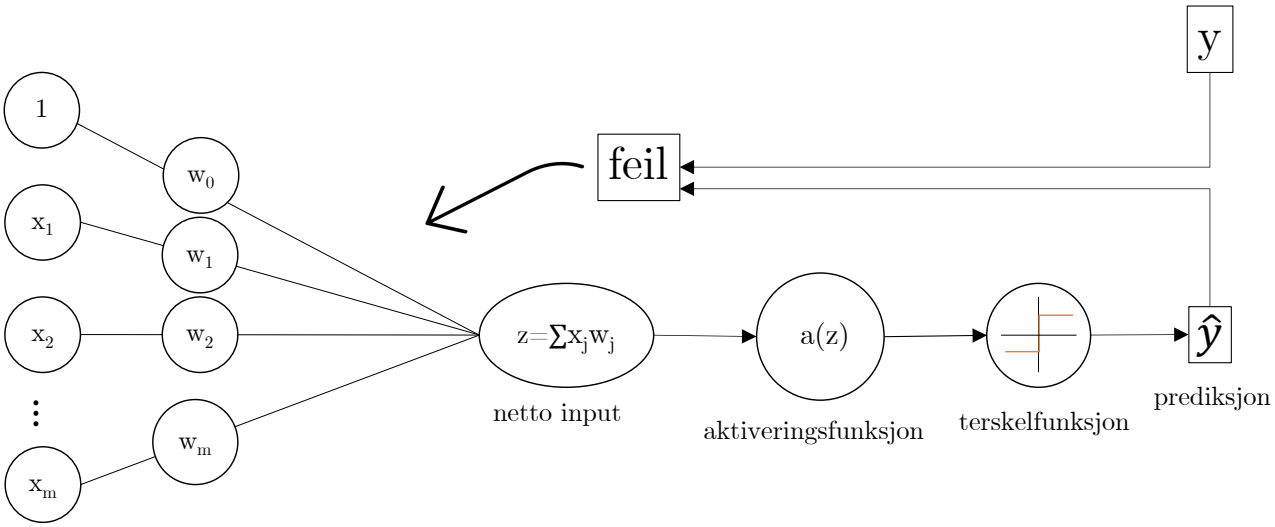
Tabell 2.1: *Et eksempel på tabelldata hvor de tre målklassene katt, hund og hest er representert. Det er n prøver i tabellen, og 2 egenskaper til hver prøve i tillegg til målklassen. Målklassen er her representert med både navn og tall, hvor tallet brukes som målklasse i dataen når den brukes i modellen.*

Prøve	Vekt (kg)	Høyde (m)	Målklasse
1	10	0.2	katt (1)
2	63	1.1	hund (0)
3	50	0.7	hund (0)
4	13	0.25	katt (1)
⋮	⋮	⋮	⋮
$x^{(n)}$	200	2	hest (2)

Eksemplene over på regresjon og klassifisering bruker samme grunnprinsipp for å predikere målklassen til hver av prøvene. Prinsippet tar utganspunkt i den enkleste modellen innen maskinlæring, nemlig det kunstige nevronet (artificial neuron).

2.3.2 Nevroner

Et kunstige nevron tar utgangspunkt i binære problem hvor man skal skiller positiv og negativ klasse. Modellen tar inn en vektor med m egenskaper, hvor $\boldsymbol{x}^{(i)}$ korresponderer til prøve nummer i, mens x_j korresponderer til egenskap nummer j [40]. Et eksempel på et vanlig kunstig nevron er Adaline [40], vist som en skisse i figur 2.11.



Figur 2.11: *Illustrasjon av Adaline. Skalarproduktet til vektorene som inneholder egenskaper* (x_i) *og vekter* (w_i) *sendes gjennom en aktiveringsfunksjon* ($a(z)$). *Feilen i prediksjonene til modellen baseres på forskjellen i prediksjon* ($\hat{y}$) *og grunnsannhet* (y), *og vektene oppdateres deretter. Illustrasjon laget etter inspirasjon fra Raschka og Mirjalili [40].*

Tilhørende hver egenskap gis et unikt *vekttall* ved oppstart av modellen, w_j, hvor j tilsvarer samme indeks som tilhørende egenskap [40]. Denne vekten er et mål på hvor mye hver egenskap har å si for å kunne skille en klasse fra andre klasser, og kan ta alle tall på den reelle aksen. Vektene er like for alle prøvene, men oppdateres etter hvert som modellen trenes.

Prosessen ved prediksjon starter med at alle egenskap-vekttall-produktene til den enkelte prøven summeres. Dette kalles netto input-funksjonen, z:

$$z = \sum_{j=0}^{m} x_j w_j + b. \tag{2.3}$$

I ligning 2.3 er det totalt m egenskaper per prøve, i tillegg til bias, b. Bias sørger for at hyperplanet definert av skalarproduktet til $\boldsymbol{x}$ og $\boldsymbol{w}$ ikke nødvendigvis går gjennom origo [40].

Netto input-funksjonen sendes gjennom en ny funksjon kalt aktiveringsfunksjonen, som for det kunstige nevronet Adaline bare er identitetsfunksjonen, $a(z) = z$. Identitetsfunksjonen til Adaline kan også ta alle verdier på den reelle aksen, siden den bare er summen av tall fra den reelle aksen. For en regresjonsmodell er den predikerte verdien til prøven gitt som $a(z)$.

Ved klassifisering sendes a gjennom en terskelverdifunksjon, gitt i ligning 2.4.

$$\hat{y} = \begin{cases} 1, & a(z) \geq 0 \\ 0, & a(z) < 0 \end{cases} \tag{2.4}$$

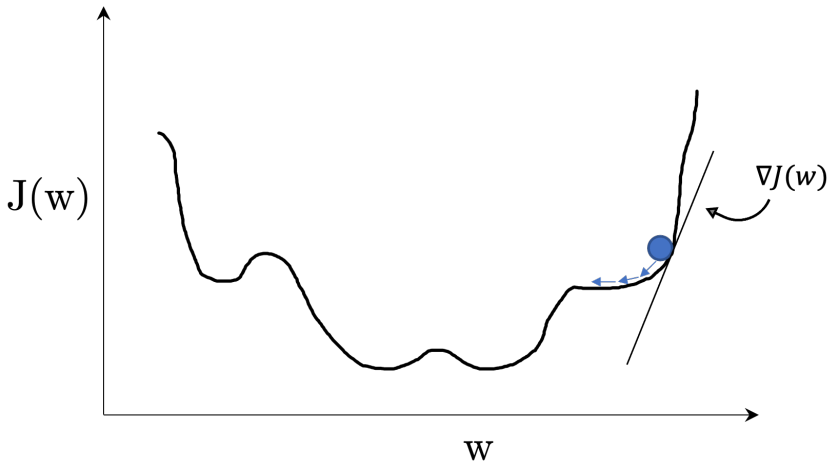
Når modellen har kommet med et forslag $\hat{y}^{(i)}$ på hvilken klasse prøve i tilhører, evalueres resultatet ut ifra hvilken klasse prøven faktisk tilhører, altså grunnsannheten $y^{(i)}$. Dersom $\hat{y}^{(i)} \neq y^{(i)}$ får modellen en feil som brukes til å oppdatere vektene for å tilpasse modellen videre. Prinsippet bak trening av Adaline er å minimere den totale feilen til hele datasettet, og dette gjøres ved hjelp av *gradientnedstigning* (gradient descent) [40].

Gradient descent tar utgangspunkt i *tapsfunksjonen* $J(\boldsymbol{w})$, som er gitt i ligning 2.5 [40]. Vektene til modellen oppdateres som gitt i ligning 2.6.

$$J(\mathbf{w}) = \frac{1}{2}\sum_{i=0}^{m}(y^{(i)} - \hat{y}^{(i)})^2 \tag{2.5}$$

$$\mathbf{w} := \mathbf{w} + \Delta\mathbf{w}, \Delta\mathbf{w} = -\eta\nabla J(\mathbf{w}) \tag{2.6}$$

Ved vektoppdateringen i ligning 2.6 multipliseres tapsfunksjonen $J(\boldsymbol{w})$ med en verdi η som typisk varierer mellom 0.0 og 1.0, men kan ta hvilken verdi som helst verdi [40]. Denne verdien kalles *læringsrate*, og den avgjør hvor mye vektene skal dyttes i den ene eller andre retningen. Med høy læringsrate kan vektene hoppe mye fram og tilbake, og ha vansker med å konvergere til optimal verdi, mens veldig små læringsrater gjør det vanskelig for modellen å nå dette optimumet [40]. Figur 2.12 illustrerer gradientnedstigning visuelt, der en blå ball illustrerer verdien til $J(\boldsymbol{w})$, og vektoppdatering er illustrert av små blå piler. Tanken er å følge kurven til $J(\boldsymbol{w})$ i motsatt retning av gradienten, som er illustrert som en tangentiell linje på kurven.

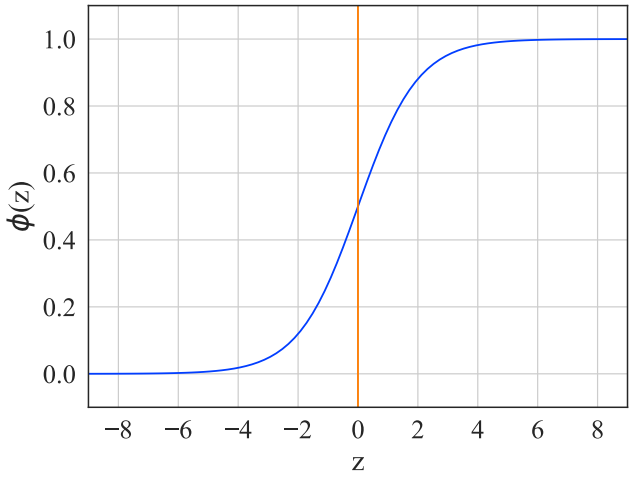


Figur 2.12: *Konseptet gradient descent går ut på å flytte verdien til det totale tapet (representert som blå ball) mot minste mulige verdi. Dette gjøres ved å flytte ballen stegvis i motsatt retning av gradienten til tapsfunksjonen, J(**w**). Illustrasjon laget etter inspirasjon fra Raschka og Mirjalili [40].*

Siden Adaline har en lineær aktiveringsfunksjon, fungerer modellen godt på lineære problem, men for å utvide bruksområdet er det vanlig å sende z gjennom en ikke-lineær aktiveringsfunksjon, som for eksempel sigmoid-funksjonen $\varphi(z)$:

$$\varphi(z) = \frac{1}{1 + e^{-z}}. \tag{2.7}$$

Sigmoid-funksjonen transformerer alle z-verdier til tall mellom 0 og 1. Dermed blir terskelverdien i ligning 2.4 endret fra 0 til 0.5 for klassifisering av binære problem. En visualisering av aktiveringfunksjonen er gitt i figur 2.13.



Figur 2.13: *Sigmoid-funksjonen $\varphi(z)$ (blå kurve), som transformerer alle z-verdier til tall mellom 0 og 1. Rød linje er trukket opp for z=0, og krysser blå kurve ved $\varphi(z) = 0.5$.*

Når φ er en verdi mellom 0 og 1 kan man se på verdien som en sannsynlighet for at en prøve tilhører positiv klasse [40]. Dersom z er et veldig høyt tall vil $\varphi(z)$ være nær 1.0, som betyr at nevronet er veldig sikker på at prøven tilhører positiv klasse.

Ved flerklassetilfeller kan man også få ut sannsynligheten for at prøven tilhører hver enkelt klasse, og dette kalles *softmax*-funksjonen [40]. Softmax-funksjonen $p(z_i)$ er gitt i ligning 2.8, der $p(z_i)$ indikerer sannsynligheten for at en prøve tilhører målklasse i, og det totalt finnes M mulige målklasser [40]. z_i er netto input for målklasse i.

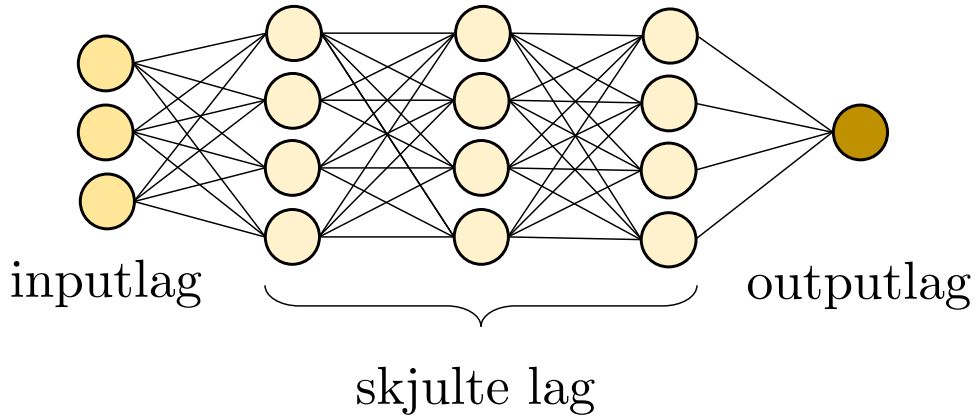
$$p(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{M} e^{z_j}} \tag{2.8}$$

Dersom en prøve av en hest skal klassifiseres som hund, katt eller hest, kan sannsynlighetene gis som en liste av for eksempel sannsynlighetene [0.1, 0.2, 0.7]. Her står sannsynligheten for at prøven tilhører klasse 0, 1 eller 2 på de respektive plasseringene i lista. Summen av hver av sannsynlighetene for at en prøve tilhører målklassene er alltid 1.

2.3.3 Nevralt nettverk

Et nevralt nettverk er et nettverk av nevroner som jobber parallelt og på rekke etter hverandre, se figur 2.14. Hvert nevron tar i mot impulser fra alle nevroner i raden før den, og sender ut

igjen én impuls, som vektlegges individuelt av hvert nevron i laget etter [40]. Det siste laget, *output-laget*, består av ett nevron i binære problemer, eller antall nevron tilsvarende antall klasser i et flerklasseproblem. Input-laget må bestå av like mange nevroner som egenskaper på prøvene. Utenom disse lagene kan nettverket bestå av et vilkårlig antall lag og vilkårlig antall nevroner i hvert lag.



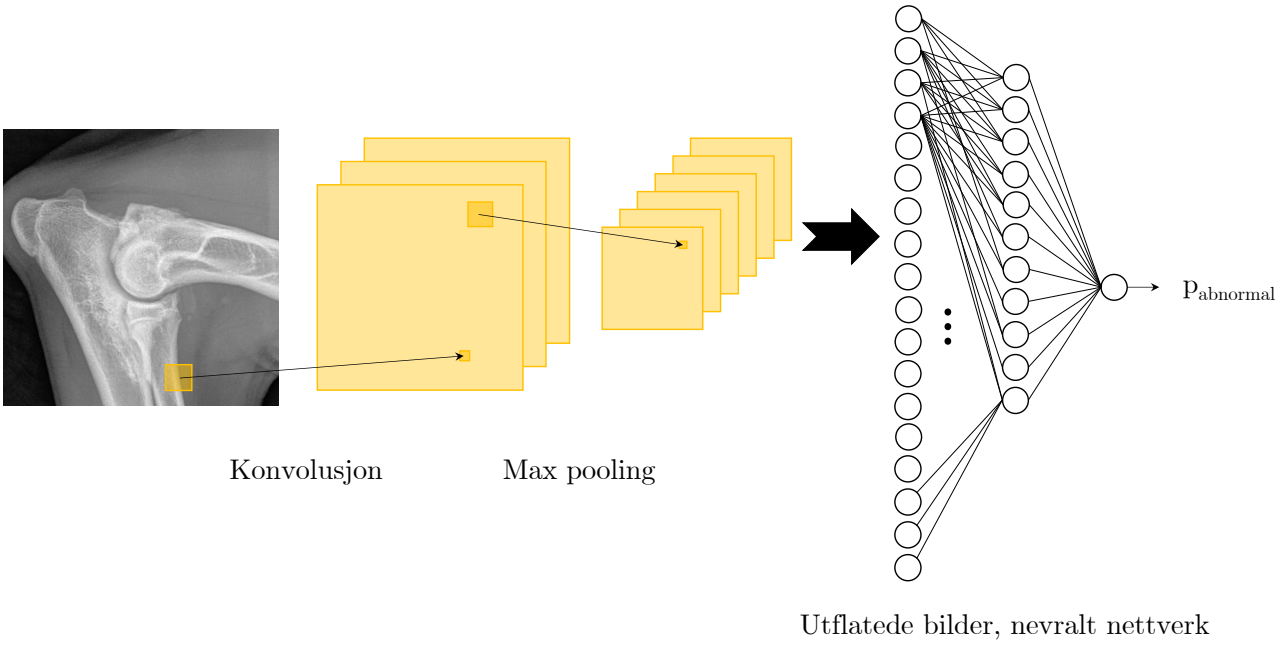
Figur 2.14: *Eksempel på et nevralt nettverk for en binær modell, med fullt koblede lag. Modellen tar inn tre egenskaper til hver prøve i inputlaget, og outputlaget består av én node som avgjør om prøven tilhører positiv eller negativ klasse. Modellen har også tre skjulte lag med like mange noder, men antall noder i de skjulte lagene har ingen påvirkning på problemstillingen eller inputformat.*

Selve læringen til modellen foregår ved at predikert verdi blir sammenlignet med grunnsannheten, og dersom disse to ikke er like, sendes *feilen* tilbake gjennom nettverket. Dette kalles *tilbakepropagering* (backpropagation), og er en av de vanligste algoritmene i trening av nevrale nettverk [40].

Hver gang modellen har sendt alle prøvene en gang fram og tilbake, altså predikert alle prøver, og så rettet på feilen, kalles en *epoke* [40]. Innenfor en epoke sendes vanligvis alle prøvene inn parti for parti (batches) i stedet for alle på én gang. Dette gjør at modellen oppdateres flere ganger per epoke. Vanligvis må modellen kjøre flere epoker før feilen i prediksjonene konvergerer til den minste verdien modellen klarer å oppnå.

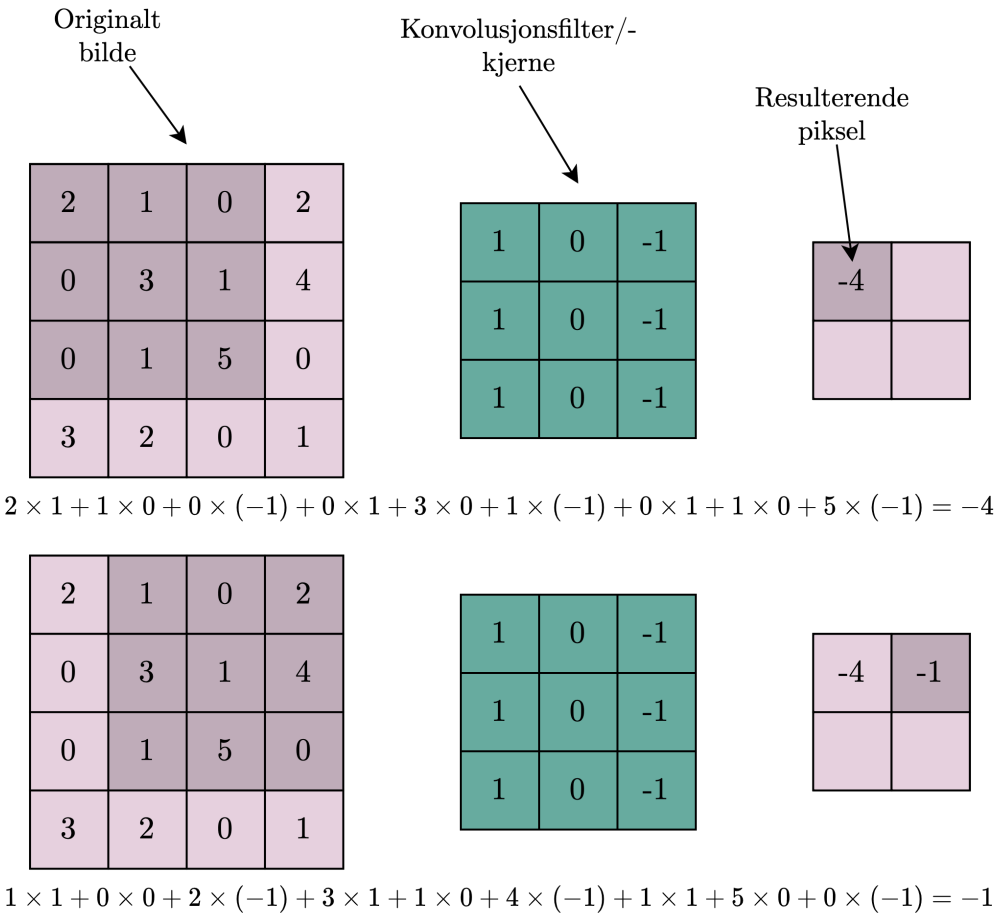
2.3.4 Konvolusjonelle nevrale nettverk

Konvolusjonelle nevrale nettverk, eller CNNer, er nevrale nettverk som tar inn bilder i stedet for lister av egenskaper for hver prøve[40]. Et eksempel på oppbyggingen av en CNN er gitt i figur 2.15. CNNer er basert på hvordan den menneskelige hjernen prosesserer objekter oppfattet med øynene [40]. Disse nettverkene er bygd opp av blant annet konvolusjonslag og samlingslag i tillegg til lag med nevroner slik som forklart i kapittel 2.3.3 om nevrale nettverk, og vist i figur 2.15.



Figur 2.15: *En skisse på en CNN (Convolutional Neural Network). Inputbildet helt til venstre behandles med konvolusjon og maxpooling for å hente ut viktige egenskaper i bildet. Til slutt flates alle pikslene ut til en endimensjonal liste som sendes gjennom et nevralt nettverk. Siste nevron bruker sigmoid aktiveringsfunksjon for å predikere sannsynligheten ($p_{abnormal}$) for at bildet tilhører en abnormal prøve (positiv klasse).*

En konvolusjon utføres ved hjelp av en *kjerneoperator/konvolusjonsfilter* som består av et lite vindu av for eksempel 3×3 vekter [40]. Disse brukes til å vektlegge alle piksler i bildet for å hente ut informasjon om mønstre i bildet, se figur 2.16.



Figur 2.16: *Illustrasjon av 2D konvolusjon. I dette eksemplet utføres konvolusjonen på et originalbilde (venstre) med* 5×5 *piksler. Konvolusjonsfilteret (midten) består av* 3×3 *vekter, og det resulterende bildet (høyre) består av* 2×2 *piksler. For hvert steg filteret tar over det originale bildet summeres de ni pikslene (mørk rosa) med de ni overlappende vektene (grønne piksler) i konvolusjonsfilteret. Her er de to første stegene i konvolusjonen vist, med tilhørende multiplikasjon og summasjon under.*

Når man gjennomfører flere konvolusjoner etter hverandre kan modellen oppdage komplekse detaljer som skiller en klasse fra en annen. Ved tilbakepropagering oppdateres vektene i konvolusjonsfilteret på samme måte som andre vekter i nevrale nettverk.

Mellom konvolusjonslagene brukes gjerne et *samlingslag (pooling)* for å redusere dimensjonen på bildene, illustrert som *Max pooling* i figur 2.15 [2]. Dette ligner konvolusjoner, men i stedet for å multiplisere verdier, velges kun én av verdiene på pikslene innenfor samlingslagets vindu (som regel 2×2). Denne verdien brukes i det resulterende bildet, mens andre piksler ikke går videre i nettverket. Den vanligste metoden er å bruke maksverdien i piksel-gruppen (max pooling), da dette ofte tilsvarer en viktig piksel i et mønster på bildet.

Etter flere lag med konvolusjoner og samling flates bildene ut, det vil si at alle pikslene samles etter hverandre i én lang liste (1-dimensjonal). I figur 2.15 skjer dette i siste steg av modellen, før prediksjon i det nevrale nettverket. Da oppfattes hver piksel som en slags egenskap til prøven, akkurat som alder eller høyde ville vært en egenskap i et datasett hvor man skulle predikere vekt på individer. Egenskapene sendes videre i det nevrale nettverket og klassetilhørigheten til

prøven predikeres på vanlig måte.

2.3.5 Overført læring

Det er mulig å ta vare på alle vektene til en modell trent på et annet datasett, for så å bruke disse vektene på problemstillingen man har for hånden. Dette kalles *overført læring* (transfer learning) [41]. Hvis man lagrer vektene til modeller som har høy nøyaktighet på prediksjoner av for eksempel katt og hund, kan man gjerne bruke disse vektene som utgangspunkt, og så for eksempel legge på flere lag i nettverket for å prøve å skille mellom hund, katt *og* hest.

Overført læring kan brukes som en modell man bygger på med flere lag, der den importerte modellen fryses med vektene slik de ble trent fra før. Da oppdateres kun vektene i lagene som blir lagt på rundt den importerte modellen. Ellers kan man bruke den importerte modellen som den er, og heller la den trene mer med finjustering (fine tuning) på det datasettet man bruker i problemstillingen for hånden [2]. I tillegg til å bruke vektene til importerte modeller, kan man også bruke modellarkitekturen uten ferdigtrente vekter [42].

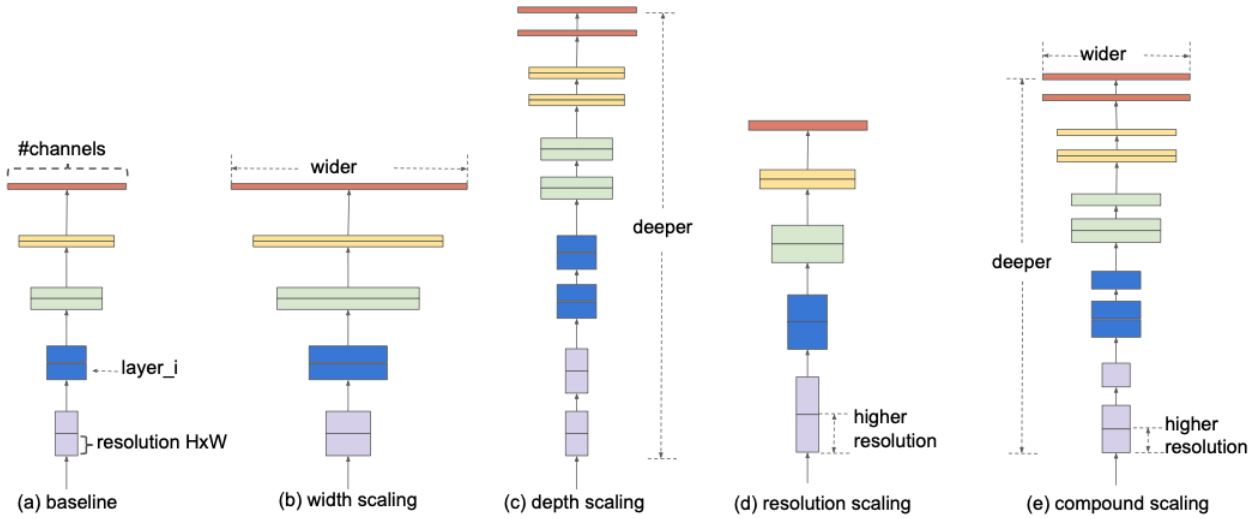
2.3.6 EfficientNet

En ferdigtrent modell med bedre nøyaktighet og effektivitet enn mange andre CNNer per dags dato er EfficientNet [20]. EfficientNet oppnår gode resultater med en ny teknikk på utvidelse av kompleksitet på referansemodellen B0[2]. Kompleksiteten til en referansemodell kan økes ved å øke en eller flere av de tre dimensjonene *vidde*, *dypde* og *oppløsning*. Vidden til en CNN refererer til antall filter eller nevroner i et lag, dybden refererer til antall lag, og oppløsning refererer til antall piksler i input-bildene. EfficientNet skiller seg fra andre modeller ved at alle tre dimensjonene økes samtidig, med betingelsen gitt i ligning 2.9 [20].

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \tag{2.9}$$

der dybde er gitt som $d = \alpha^\phi$, vidde som $w = \beta^\phi$ og oppløsning som $r = \gamma^\phi$. ϕ er gitt som økningen i dataressurser som brukes i modellen [20]. En illustrasjon på effekten av utvidelse i hver av dimensjonene er gitt i figur 2.17.

[2]EfficientNet-B0 er brukt i EfficientNet-familien, men utvidelsesmetoden kan brukes på andre modeller brukt som referanse [20].



Figur 2.17: *Illustrasjon på skalering av modeller. (a) er et eksempel på en referansemodell. (b)-(d) er eksempler på typer oppskalering av referansemodellen. (e) er en helhetlig oppskalering, som kombinerer alle skaleringsmetodene illustrert i (b)-(d). Figur gjengitt med tillatelse fra Tan og Le [20].*

For den enkleste modellen, som har kompleksitet B0, er hver variabel gitt som:

$$\alpha = 1.2$$

$$\beta = 1.1$$

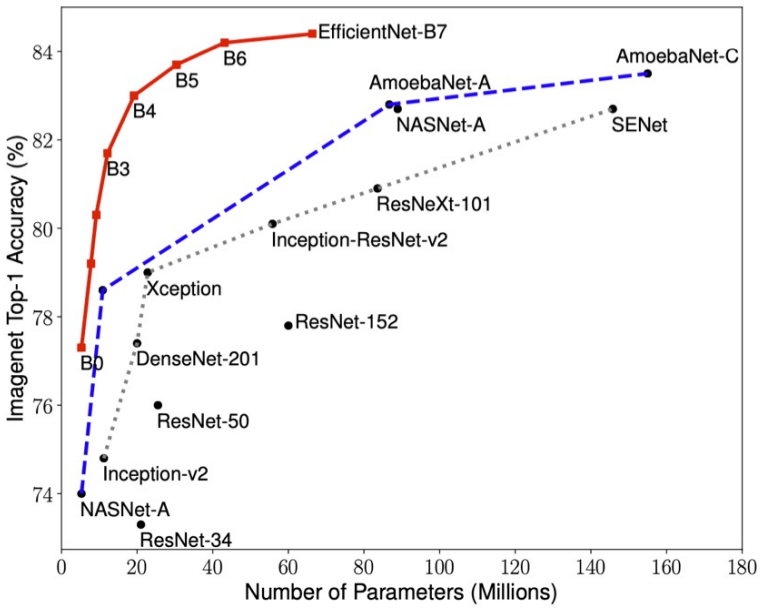
$$\gamma = 1.15$$

$$\phi = 1 \ .$$

Kompleksitetene til EfficientNet spenner fra B0 til B7, med økende dimensjoner, og dermed også antall parametre, gitt i tabell 2.2. EfficientNet når foreløpig state-of-the-art-nøyaktigheter, men med færre parametre enn sammenlignbare modeller, se figur 2.18.

Tabell 2.2: *Antall parametre tilhørende hver modell i EfficientNet-familien [20].*

Modell	**Antall parametre**
EfficientNet-B0	5.3M
EfficientNet-B1	7.8M
EfficientNet-B2	9.2M
EfficientNet-B3	12M
EfficientNet-B4	19M
EfficientNet-B5	30M
EfficientNet-B6	43M
EfficientNet-B7	66M



Figur 2.18: *Sammenlignig av top-1 nøyaktighet*[3] *og antall parametre i ulike CNNer. EfficientNet-familien scorer bedre på ImageNet*[4] *enn andre modeller med like mange parametre. Figur gjengitt med tillatelse fra Tan og Le [20].*

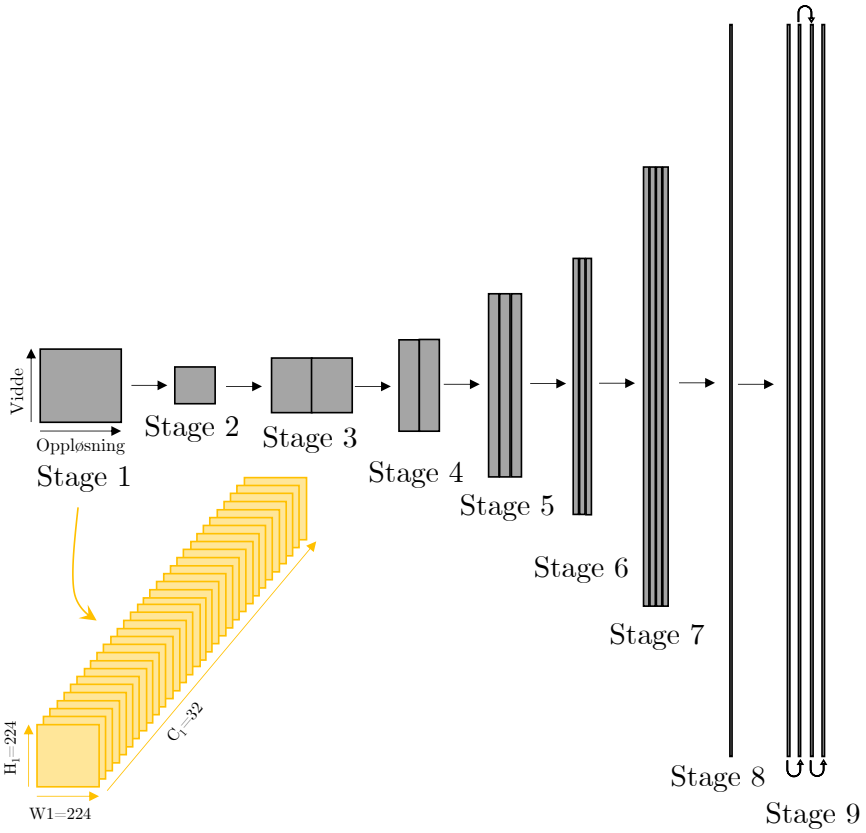
Som nevnt tidligere brukes EfficientNet-B0-modellen som referansemodell ved oppskalering i EfficientNet-familien. EfficientNet-B0 er illustrert i figur 2.19 basert på detaljer i tabell 2.3. En referansemodell (baseline model) er en enkel modell som tjener som en referanse i maskinlæringsprosjekt [44]. Hva som regnes som en enkel modell er individuelt for hvilken problemstilling man har for hånden, men målet er uansett å oppnå høyere ytelse ved å trene mer komplekse modeller enn referansemodellen. En enkel referansemodell kan være å plassere alle prøver i målklassen som er høyest representert i datasettet [44]. I et datasett med 90 % prøver i kategorien "friske pasienter" og 10 % i kategorien "syke pasienter", vil en slik referansemodell gi riktig prediksjon til 90 % av alle prøvene. Dette kan virke som en høy ytelse ved første øyekast. Imidlertid kan ikke modellen brukes i det hele tatt, siden den ikke kan finne en eneste syk pasient i datasettet. I kapittel 2.3.7 blir det gitt nærmere forklaring på ulike ytelsesmål i maskinlæring.

[3]top-1 nøyaktighet er nøyaktighet basert på den høyeste sannsynligheten, i motsetning til f.eks. top-3 nøyaktighet, der prediksjonen regnes som korrekt dersom en av de tre høyeste sannsynlighetene er den sanne målklassen [43]. Se kapittel 2.3.7 for definisjon av nøyaktighet.

[4]https://www.image-net.org/challenges/LSVRC/

Tabell 2.3: *Referansemodellen B0 i EfficientNetfamilien. Hver "stage" (steg) i tabellen er illustrert i figur 2.19. For hvert steg i CNNen brukes tilhørende operator, med input-oppløsning (Resolution) og antall filtre (Channels). Antall lag i hvert sted er også oppgitt. Tabell gjengitt med tillatelse fra [20].*

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

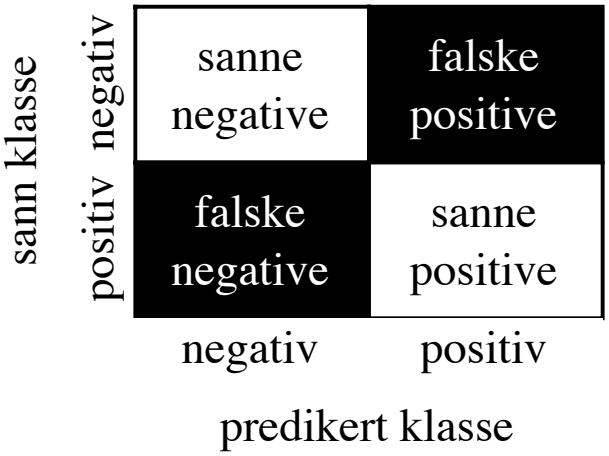


Figur 2.19: *Visualisering av referansemodellen til EfficientNet, B0, basert på tabell 2.3. I stage 9 er vidden delt opp i fjerdeldeler, dvs. at stage 9-vidden er fire ganger så stor som den i stage 8. H og W står for høyde og bredde, og angir sammen oppløsningen. C står for antall filtre, som angir vidden til modellen, mens dybden er gitt av antall stages og lag i hver stage.*

2.3.7 Ytelsesmål

Ved klassifisering mellom to klasser er fire ulike utfall mulig. Disse tar utgangspunkt i om man klarer å klassifisere gruppe 1 (positive prøver) fra gruppe 0 (negative prøver). Korrekt

klassifisering av gruppe 1 og 0 og kalles henholdsvis sanne positive (SP) og negative (SN). Prøver feilklassifisert som gruppe 1 og 0 kalles henholdsvis falske positive (FP) og falske negative (FN). Mengden prøver som havner i hver av de fire kategoriene presenteres ofte i en *forvirringsmatrise* (confusion matrix), som vist i figur 2.20 [40]. Målet i alle maskinlæringsoppgaver er å klassifisere alle prøver i den sanne gruppa prøven tilhører.



Figur 2.20: *Forvirringsmatrise. Mengden av prediksjoner som faller under hver kategori presenteres i en slik matrise, blant annet for å visualisere hvilken sann gruppe som blir mest feilpredikert.*

Resultatene på klassifiseringen måles i ulike ytelsesmål. Den vanligste og enkleste kalles nøyaktighet (accuracy), og teller hvor mange prøver som ble riktig predikert [45]. Dette gis som andelen riktig predikerte prøver i forhold det totale antallet prøver i datasettet, se ligning 2.10.

$$accuracy = \frac{\sum(\hat{y}^{(i)} = y^{(i)})}{\sum y^{(i)}} = \frac{SP + SN}{SP + SN + FP + FN} \tag{2.10}$$

For eksempel vil nøyaktigheten på lista med predikerte prøveverdier $\hat{\mathbf{y}}$=[0, 0, 1, 1] være 0.75 for en grunnsannhet lik $\mathbf{y}$=[0, 1, 1, 1]. Hver av verdiene i prediksjonen $\hat{\mathbf{y}}$ avgjøres av terskelverdien satt i aktiveringsfunksjonen for binære problem. Når det gjelder flerklasseproblem brukes softmax funksjonen (se ligning 2.8) til å gi sannsynligheten for at en prøve tilhører hver av de mulige klassene i problemstillingen [40]. Da brukes ikke lenger en satt terskelverdi, men høyeste sannsynlighet avgjør hvilken klasse prøven blir tildelt. Dersom et problem inneholder tre ulike målklasser, er altså laveste mulige verdi for å være mest sannsynlig >0.33.

Målet med klassifisering er altså å legge alle prøvene i rett klasse, som også vil si å unngå feilprediksjoner. Noen vanlige scoreberegninger som gir en indikator på grad av sann- og feilprediksjon er *sensitivitet/gjenkalling* (sensitivity/recall), *spesifisitet* (specificity) og *presisjon* (precision), gitt i ligning 2.11, 2.12 og 2.13 [46].

$$sensitivitet/gjenkalling = \frac{SP}{SP + FN} \tag{2.11}$$

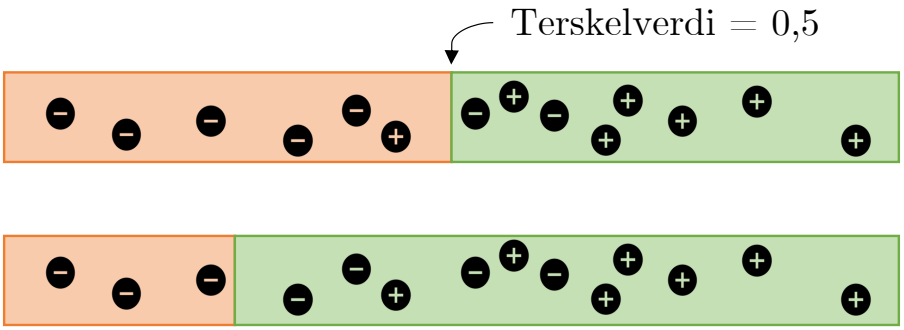
$$spesifisitet = \frac{SN}{FP + SN} \tag{2.12}$$

$$presisjon = \frac{SP}{SP + FP} \tag{2.13}$$

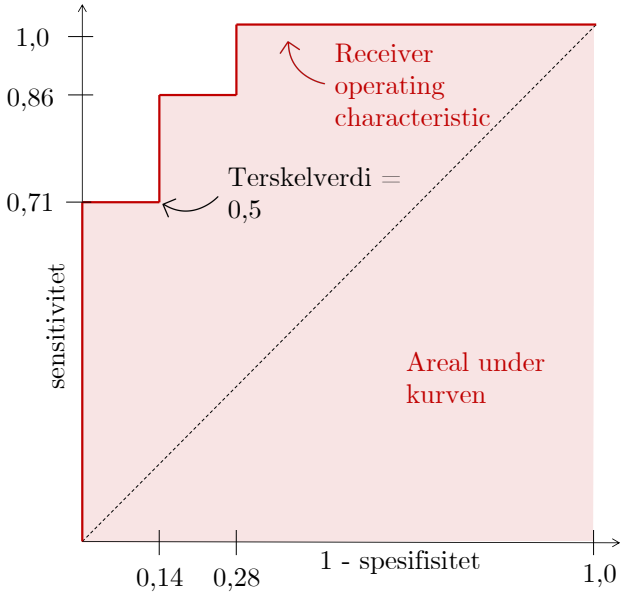
Sensitivitet gir et mål på andelen sanne positive av alle prøver som faktisk tilhører postitiv klasse. Denne maksimeres dersom antall falske negative ønskes minimert, og sanne positive ønskes maksimert. Et eksempel på en slik situasjon er ved prediksjon av kreft hos mennesker. Da er det viktig at ingen syke mennesker får beskjed om at de er friske (dvs. falsk negativ), siden det i verste fall kan medføre død på grunn av ubehandlet sykdom.

Spesifisitet gir andel sanne negative blant alle prøver som faktisk tilhører negativ klasse. Denne ønsker man å maksimere i tilfeller der det er viktig at prøver som er negative også predikeres som negativ. Dette kan gjelde ved for eksempel screening av hunder, siden falske positive prøver kan medføre økonomisk tap for eieren.

Et mål som kombinerer spesifisitet og sensitivitet er AUC (Area Under the receiver operating characteristic Curve), se figur 2.21 [40]. AUC (noen ganger referert til som AUC ROC) spenner fra 0 til 1, hvor 0.5 regnes som tilfeldig gjetning, 1 er en perfekt separasjon av sanne positive og negative, mens 0 betyr perfekt feilklassifisert separasjon av sanne positive og negative prøver. Dette ytelsesmålet representerer den totale ytelsen til den binære modellen ved å måle arealet under kurven som representerer antall korrekt klassifiserte positive prøver mot antall feilklassifiserte positive prøver. Hvert punkt på kurven er beregnet ved å flytte terskelverdien i sigmoid-aktiveringsfunksjonen, se figur 2.13, fra 0 til 1 .



(a) *ROC*



(b) *ROC*

Figur 2.21: *Hvert punkt på en Receiver operating characteristic-kurve er beregnet ut ifra sensitiviteten og spesifisiteten til en modell der terskelverdien for positiv klasse endres. I (a) illustrerer grønt område alle prøver klassifisert som positive. Aktiveringsverdiene til hver prøve i (a) er gitt av posisjonen til prøven, der grenseverdien går fra 0 (venstre) til 1 (høyre). (b) viser kurven som resultat av å flytte terskelverdien fra 0 til 1, og området under kurven (lyserødt) er arealet under ROC-kurven.*

Ulempene med presisjon og gjenkalling er at de enten er et mål på hvor mange av alle positive prøver som blir fanget opp *eller* hvor mange positive prediksjoner som faktisk er positive (sanne positive) [40]. En mye brukt score som tar hensyn til begge sider kalles F1-scoren [40], gitt i ligning 2.14 [46].

$$F1 = 2\frac{presisjon \times gjenkalling}{presisjon + gjenkalling} = \frac{2SP}{2SP + FP + FN} \tag{2.14}$$

F1 er også kjent som harmonisk gjennomsnitt av presisjon og gjenkalling, og tar verdier på intervallet fra 0 til 1 [46]. Denne verdien gir en score på antall sanne positive også tatt hensyn til falske positive og falske negative, i motsetning til gjenkalling og presisjon. Dette ytelsesmålet brukes ofte dersom det er ønskelig å optimere en modell for positiv klasse, og dersom datasettet ikke er balansert. Dersom man ønsker å optimere for negativ klasse, kan man bytte om alle

positive og negative komponenter i ligning 2.14, slik at F1-verdien blir en score på hvor godt modellen predikerer negative prøver. En ulempe med F1, er at ytelsesmålet ikke tar hensyn til både positiv og negativ klasse samtidig.

En scoreberegning som tar hensyn til alle fire utfall i et binært klassifiseringsproblem er MCC-score (Matthews Coefficient Correlation) [46]. Denne beregningen spenner fra -1 til 1, hvor 1 er en perfekt klassifisering, 0 tilsvarer "tilfeldig gjetning", og -1 er perfekt feilklassifisering. Utregningen til MCC er gitt i ligning 2.15 [46].

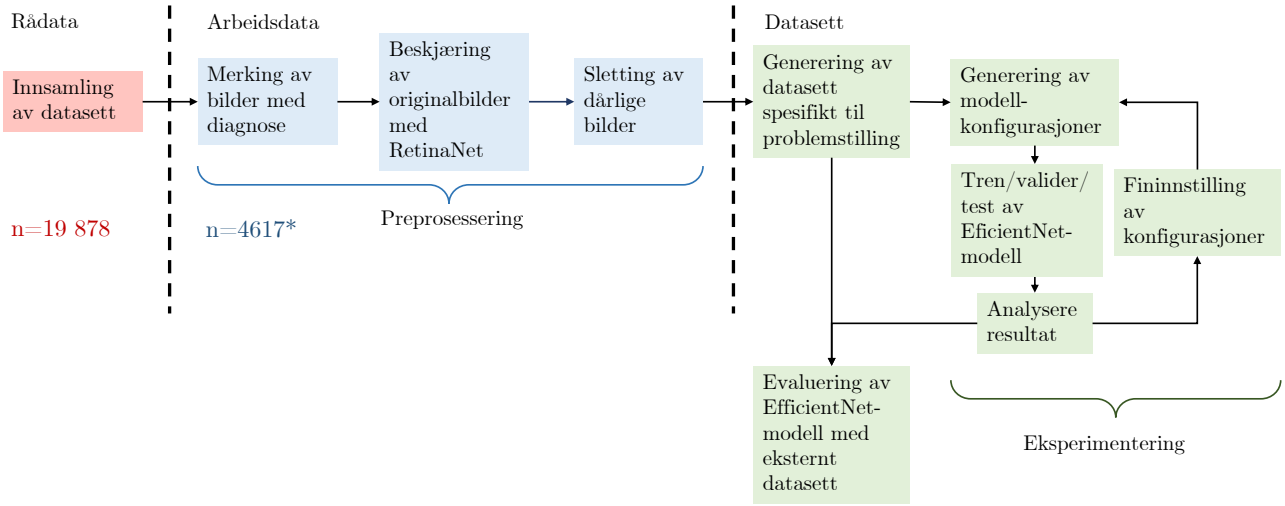
$$MCC = \frac{SP \times SN - FP \times FN}{\sqrt{(SP + FP)(SP + FN)(SN + FN)(SN + FP)}} \tag{2.15}$$

Kapittel 3

Material og metode

Som forklart i kapittel 2.3.1 kan maskinlæringsmodeller brukes til både binære og flerklasse-problemer. EfficientNet er en familie av CNNer som forklart i kapittel 2.3.6, og det er modeller med ulike kompleksiteter i EfficientNet-familien som brukes til alle problemstillingene i denne masteroppgaven. EfficientNet er valgt på grunn av den høye ytelsen til modellene.

I dette kapitlet presenteres materialene brukt i masterarbeidet, fra innsamling av data til analyse av resultater etter kjøring av CNN-modeller. Arbeidsflyten ved behandling av data er presentert i figur 3.1, og hvert steg forklares nærmere i de ulike delkapitlene. I denne masteroppgaven blir "prøver" brukt om enkeltbilder som brukes til trening og evaluering av maskinlæringsmodeller. Eksperimentering blir brukt om utprøving av ulike konfigurasjoner ved trening, validering og testing av modeller, se figur 3.1, men også ved ekstern evaluering.



Figur 3.1: *Arbeidsflyten i prosessen. 19 878 røntgenbilder, kalt rådata (i rød), ble overlevert fra NKK til veterinærer ved NMBU Veterinærhøgskolen. Et utvalg av røntgenbildene ble markert med diagnose av veterinærene, før de ble preprosessert videre med beskjæring ved hjelp av RetinaNet, og uegnede bilder ble slettet fra arbeidsdataen (blå). *Arbeidsdataen besto av 4617 bilder etter sletting av uegnede bilder. Før eksperimentering for hver problemstilling ble et passende utvalg av arbeidsdataen plukket ut ved generering av datasett, og brukt i trening, validering og testing av maskinlæringsmodeller (grønn). Evaluering ble gjort på noen modeller etter trening, validering og testing, og egne eksterne datasett ble generert til eksterne evalueringer. Prøver inkludert i ekstern evaluering ble ikke brukt ved trening, validering og testing.*

3.1 Rådata

Rådataen, eller rådatasettet, med bilder på DICOM-format ble overlevert fra Norsk Kennel Klub til veterinærer ved NMBU Veterinærhøgskolen i september 2022. Det ble også hentet røntgenbilder tatt i 2022 i april 2023. Dette er første steg (markert i rødt) i arbeidsflyten skissert i figur 3.1. I tillegg til bilder ga NKK et tabulært datasett med metainformasjon tilhørende alle hunder som gjennomgikk AD-screening via NKK i perioden 2012-2022. Dette omfattet blant annet diagnosegraden (nivå 1, 2 eller 3) på hvert framben, alderen, kjønnet og rasen til hver hund. Dette metadatasettet regnes som en god representasjon på AD-forekomst i den totale populasjonen av større hunderaser i Norge [11]. Videre utforsking av rådataen er gjort på hunder screenet i perioden 2018-2021.

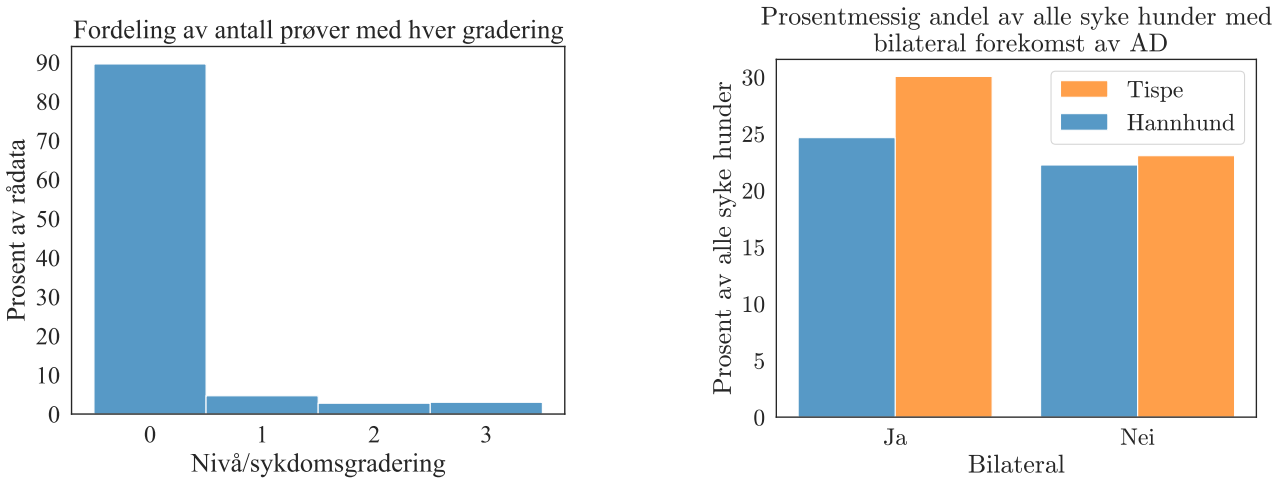
Alle bildene i rådatasettet er tatt ved norske veterinærklinikker i forbindelse med screening, både i tilfeller hvor hunden skal avles på selv og når dette ikke gjelder. Screening for AD blir for det meste gjort på hunderaser hvor AD forekommer (se kapittel 2.2.1), og for det meste med renrasede hunder [11]. Bildene er registrert hos NKK uavhengig av diagnose.

Røntgenbildene tatt i perioden 2018 til 2021 tilhører mellom 4500 og 5500 hunder for hvert år, som totalt blir 19 878 hunder i arbeidsdataen. I bildesettet er det flere bilder av samme hund, men ikke alle bildene eller hundene er med videre i datasettet brukt i denne masteroppgaven. Alderen på hundene spenner fra 6 måneder til 12 år, med et gjennomsnitt på 21 måneder.

Prosjektgruppa ved NMBU har hatt tilgang til bilder og annen informasjon i et lukket filsystem som bare spesifiserte brukere har tilgang til. På denne måten bevares personvernet til hundene og hundeeierne. Når bildene sendes gjennom tredje steg i arbeidsflyten (figur 3.1), *beskjæring*, separeres de fra metadataen, og kan ikke spores tilbake til identiteten uten spesiell tilgang.

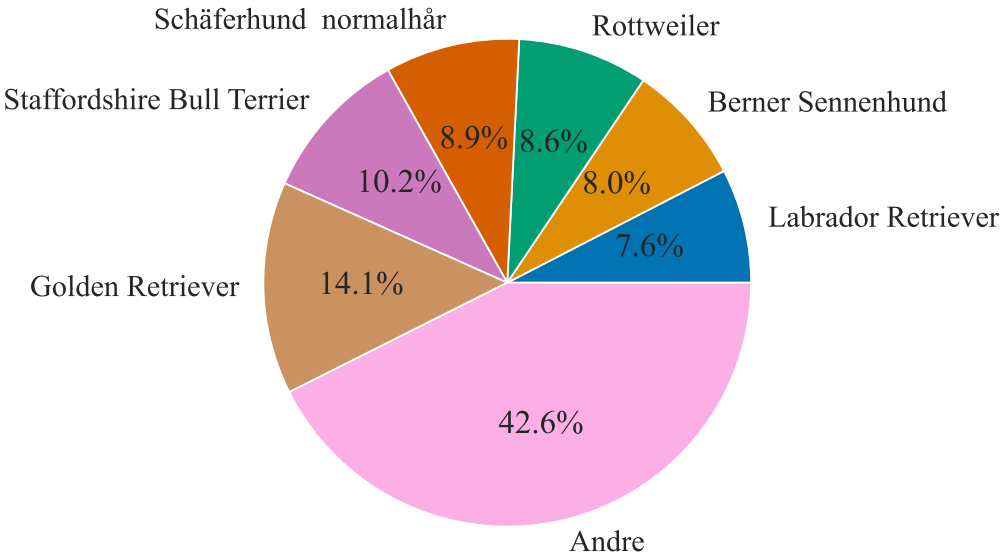
3.2 Analysering av rådata

Rådataen, altså hele datasettet før preprosessering, inneholder informasjon om totalt 19 878 hunder. Blant disse hundene er 2073 (10,4%) hunder registrert som syke (se figur 3.2a), hvorav omtrent 55% har registrert AD i begge albueledd (bilateral forekomst), se figur 3.2b.



Figur 3.2: *(a) Blant alle screenede hunder i perioden 2018-2021 ble nesten 90% registrert som uten tegn til AD, kalt normale. De resterende drøye 10% var jevnt fordelt på sykdomsgrad 1, 2 og 3. Her regnes den høyeste graderingen av venstre og høyre albue som registrert diagnose. (b) Blant alle hunder registrert med albueleddsdysplasi hos NKK mellom 2018 og 2021 hadde omtrent 55% av hundene AD i begge albueledd (bilateral forekomst). Dette gjelder uavhengig av gradering av AD i hver albue, så lenge AD er til stede i albueleddene.*

Blant de 2073 hundene med registrert AD utgjorde seks raser omtrent 57% av alle hundene. De resterende $\sim$ 43% av sykdomstilfellene kom fra 116 andre raser, se figur 3.3. Av alle hunder som gjennomgikk screening for AD i samme periode (2018-2021) var det totalt 209 ulike raser, altså var omtrent 42% av alle rasene i screeningprogrammet uten forekomst av AD.



Figur 3.3: *Blant totalt 2073 tilfeller av AD i perioden 2018-2021 utgjorde seks raser omtrent 57 % av tilfellene. Alle de andre 116 rasene med AD utgjorde resterende 42.6 % av det totale antallet AD-rammede hunder. Hver rase i gruppa "andre" hunder utgjorde mindre enn 5 % hver.*

3.3 Programvare

All videre behandling av data ble gjort med programmering i Python versjon 3.7.15. Rammeverket *deoxys* [47] ble brukt til oppsettet av pipeline til maskinlæringsmodellen. Deoxys har åpen kilde, og er utviklet av Bao Ngoc Huynh for å hjelpe radiologer med dyplæring på medisinske bilder [47]. Til preprosessering og analyse av resultat ble programfiler presentert i tabell 3.1 brukt, og alle filene ligger tilgjengelig på github[1]. Alle filer nevnt senere i dette kapitlet ligger i tabellen.

Alle eksperimenter ble kjørt på NMBUs Orion High Performance Computing (Orion) [48], som er en cluster av CPUer og GPUer tilgjengelig for studenter og ansatte ved NMBU. Orion bruker Slurm[2] til cluster management, altså til å sette opp køer for jobber på clusteren [48].

Tabell 3.1: *Oversikt over filer brukt i denne masteroppgaven, med filplassering. Filnavnene er eksempler der det finnes flere lignende filer brukt til ulike eksperimenter. Alle filer kan bli funnet på github* https://github.com/huynhngoc/cubiai.

Plassering	**Filnavn**	**Bruk**	**Utvikler**
/cubiai/	autocrop.py	Beskjæring av bilder ved hjelp av RetinaNet	Bao Ngoc Huynh
/cubiai/notebook/	0_normal.ipynb	Gjennomgang og sletting av dårlige bilder	Bao Ngoc Huynh og Sunniva E. D. Steiro
/cubiai/dataset_gen/	gen_normal_abnormal_2.py	Generering av datasett fra bilder i arbeidsdataen	Bao Ngoc Huynh og Sunniva E. D. Steiro
/cubiai/config_gen/	/cubiai/gen_config.py	Generering av konfigurasjoner, e.g. kombinasjon av blant annet modellkompleksitet og læringsrate til et eksperiment	Bao Ngoc Huynh og Sunniva E. D. Steiro
/cubiai/	/cubiai/experiment_binary.py	Kjøring av eksperiment. Pipeline med trening, validering og testing av EfficientNet-modell	Bao Ngoc Huynh
/cubiai/	/cubiai/slurm_pretrain_binary.sh	Slurm-script til kjøring av EfficientNet-modeller som slurm jobs på Orion.	Bao Ngoc Huynh
/cubiai/sunniva/	Elbow_Experiments.xlsx	Registrering av eksperimenter	Sunniva E. D. Steiro

3.4 Preprosessering av arbeidsdata

Arbeidsdataen brukt til trening, validering og testing etter preprosessering er et undersett av rådataen, som ble merket med diagnoser i tillegg til sykdomsgradering i steg 2 i figur 3.1. Alle stegene for preprosessering av arbeidsdata er skissert i arbeidsflyten i figur 3.1. Bildene ble merket med spesifikke diagnoser med gradering av radiologer ved NMBU Veterinærhøgskolen. De aktuelle diagnosene var artrose, sklerose, MCD, OCD og UAP, og graderingene spant fra nivå 1 til 3. Se tabell 3.2 for komplett oversikt over hvor mange prøver (røntgenbilder) som var med i hvilken kategori totalt i arbeidsdataen. Bilder av friske hunder ble kategorisert som "nivå 0 normal". Alle kategoriene ble også tildelt et nummer kalt *encoding*, se tabell 3.2.

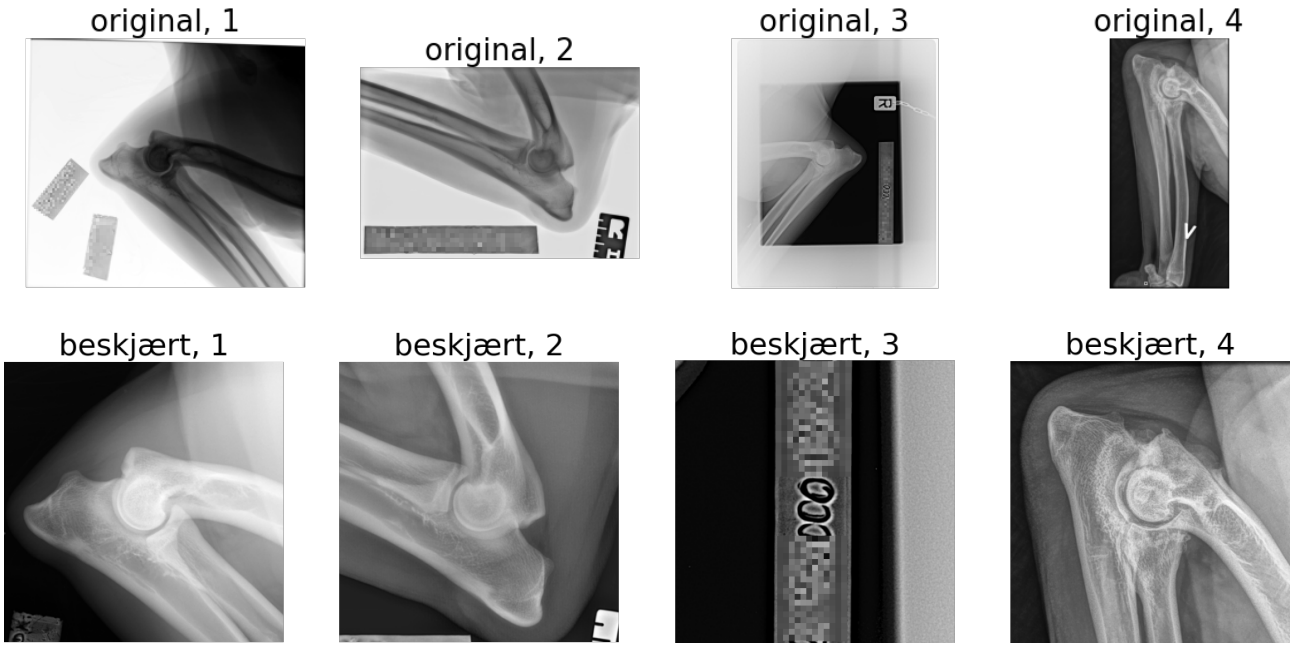
[1] https://github.com/huynhngoc/cubiai

[2] https://slurm.schedmd.com/overview.html

Tabell 3.2: *Oversikt over antall prøver (røntgenbilder) i hver kategori i arbeidsdataen, med tilhørende encoding brukt i stedet for navn på diagnose.*

Diagnose	**Antall**	**Encoding**
Nivå 0 normal	2273	0
Nivå 1 artrose og/eller sklerose	1027	1
Nivå 2 artrose	496	2
Nivå 2 MCD	212	3
Nivå 3 artrose	216	4
Nivå 3 MCD	330	5
Nivå 3 OCD	16	6
Nivå 3 UAP	49	7
Sum	4617	

Siden røntgenbildene er tatt av forskjellige veterinærer ved ulike klinikker og av ulike hunderaser, er det variasjoner i bildene som for eksempel posisjonen til hundens albue, hvilke andre komponenter som er med på bildene, og hvor åpen vinkelen på albuen er. Se figur 3.4 for eksempler på variasjoner i bilder.

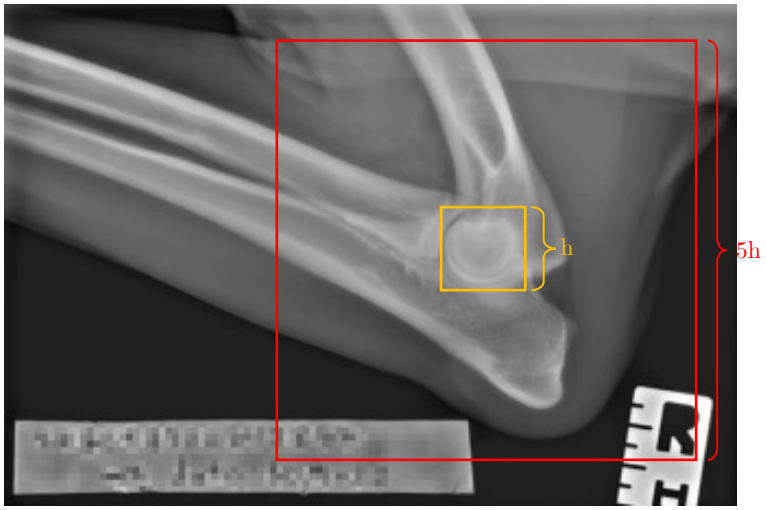


Figur 3.4: *Bilder med samme nummerering korresponderer til hverandre før (øverste rad) og etter (nederste rad) preprosessering. Originalbildene er tatt av et større område enn det som er interessant i forbindelse med AD. Beskjærte bilder har omtrent like dimensjoner, og inneholder kun aktuelle områder for predikering av albueleddsdysplasi. I noen tilfeller detekteres feil objekt, som i "beskjært, 3", der et null-tall ble detektert i stedet for kondylen. Bilde 1 og 2 er normale, mens bilde 3 har artrose nivå 1 og bilde 4 artrose nivå 3. Det er tydelig variasjon i både originale og beskjærte bilder, som for eksempel plassering av albue og hvilke komponenter som er med på bildene.*

Etter at bildene i arbeidsdataen var kategorisert basert på diagnose, ble bildene prosessert videre for å kunne brukes av EfficientNet-modellene. De originale bildene hadde forskjellige dimensjoner, avhengig av hvor stor del av hundenes framben som var tatt med på røntgenbildene. Siden det bare er området rundt kondylen som inneholder tegn til albueleddsdysplasi på røntgenbilder, ble alle originalbildene beskjært rundt dette området. Figur 3.4 viser originalbilder med tilhørende bilder før og etter beskjæring.

For å beskjære bildene automatisk i stedet for ved manuell gjennomgang ble det brukt en CNN kalt RetinaNet[3]. RetinaNet er en effektiv gjenstandsdetektor-modell som bruker *Focal Loss*-tapsfunksjon [49]. Focal Loss tar hensyn til ubalansen i klasser mellom "interessante" objekter i et bilde og "bakgrunnsobjekter" [49]. RetinaNet ble brukt i programfila autocrop.py (se tabell 3.1), som leser inn hvert bilde og beskjærer bildene i området rundt den sirkulære formen på kondylen. Originalbildene er røntgenbilder på DICOM-format[4], og disse ble lest med Python-biblioteket simpleITK[5]. Bildene ble omformatert til 8-bits .npy-matriser[6] (numpy array-filer) for bruk i EfficientNet-modellene. De originale DICOM-filene varierte mellom 16- og 32-bits piksler.

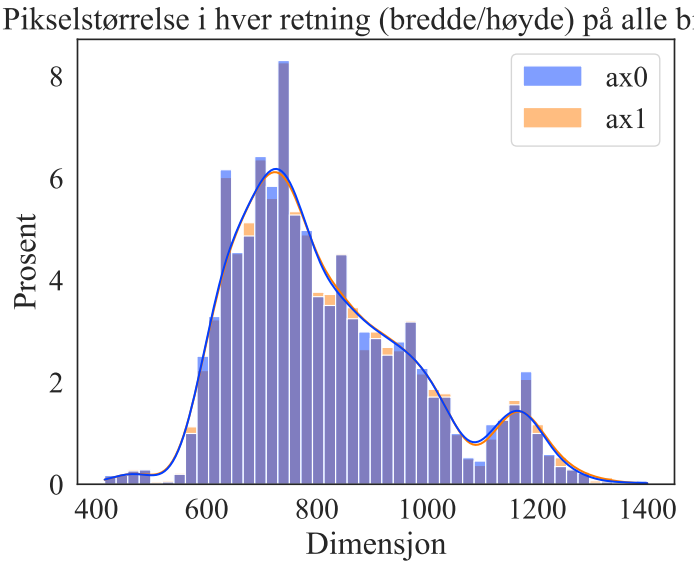
Bildene ble beskjært i ulik størrelse, som avhenger av størrelsen på pikselmatrisa kondylen ble predikert å være i. Størrelsen er et resultat av at bildene ble beskjært slik at høyden og bredden på det beskjærte bildet ble fem ganger så store som høyden og bredden i pikselmatrisa som inneholder kondylen. En illustrasjon på beskjæring av bilder er vist i figur 3.5.



Figur 3.5: *Illustrasjon på beskjæring av bilder. Kondylen ble detektert ved hjelp av RetinaNet, som forklart i kapittel 3.4. Rektangelet som omringer kondylen (gult) ble forstørret fem ganger i hver retning, der rektangelets side er gitt som h i figuren. Det røde kvadratet representerer området som ble tatt med i det beskjærte bildet. Her er rektangelet illustrert som kvadrat, men i praksis vil ulikheter i høyde og bredde finne sted. Illustrasjonen er en skisse.*

Resultatet av denne beskjæringen var bilder på mellom 400 og 1400 piksler i horisontal og vertikal retning, med et gjennomsnitt på omtrent 800 piksler i begge retninger, se figur 3.6.

[3] https://github.com/facebookresearch/Detectron
[4] https://www.dicomstandard.org/
[5] https://simpleitk.org/
[6] https://numpy.org/devdocs/reference/generated/numpy.lib.format.html



Figur 3.6: *Bildestørrelsene på de automatisk beskjærte røntgenbildene varierte fra ca 400 til 1400 piksler i hver retning (ax0 og ax1), med et gjennomsnitt på omtrent 800 piksler i hver retning. ax0 tilsvarer antall piksler i vertikal retning, mens ax1 tilsvarer antall piksler i horisontal retning.*

Som figur 3.4 viser, ble noen av bildene beskjært feil, som resulterte i uegnede bilder. Dette kom av at RetinaNet detekterte en sirkulær form som ikke var kondylen. På bilde 3 i figur 3.4 har modellen valgt ett null-siffer på identifikasjonslappen som senter. Bilder som ikke ble lest inn riktig ble slettet fra datasettet etter beskjæring. Totalt ble 77 av 4694 bilder slettet fra arbeidsdataen, slik at antall bilder brukt til trening, validering, testing og evaluering av modeller i denne masteroppgaven ble 4617. For å se over alle beskjærte bilder og å fjerne de uegnede bildene fra det preprosesserte datasettet ble programfilene i mappa `cubiai/notebook` brukt.

3.5 Generering av datasett for klassifisering

Etter preprosessering av arbeidsdataen ble de 4617 prøvene delt opp i mindre datasett som eksperimenter skulle bli utført på, dette er tredje del (i grønt) av arbeidsflyten skissert i figur 3.1. For hvert eksperiment ble en ny modell med en unik konfigurasjon trent, se kapittel 3.6 for detaljer om konfigurasjoner. De to hovedgruppene av modeller, altså hovedgruppene av problemstillinger, i denne masteroppgaven er *binære* og *flerklasse*modeller, og alle typer problemstillinger utprøvd er listet som følger:

Binære eksperiment

- Normale vs. abnormale albuer

  Utvidet analyse: Klassifiserbare vs. ikke klassifiserbare prøver

- Nivå 1 vs. øvrige nivåer (“resten”)
- Artrose nivå 1, 2 og 3 vs. øvrige diagnoser (“resten”)

Flerklasseeksperiment

- Skille mellom alle 3 nivåer
- Skille mellom alle diagnoser (encoding 1 - 7 i tabell 3.2)

For hver problemstilling ble et individuelt datasett til trening, validering og testing generert. Noen av modellene ble også evaluert med et eksternt datasett, som ble generert i steg 5 i figur 3.1. En ekstern evaluering er en kjøring av modellen på et datasett bestående av prøver som ikke har vært brukt til trening, validering eller testing av modellen tidligere (eksterne prøver). I tillegg til eksterne evalueringer ble noen av modellene trent om igjen med eksterne prøver. Omtrening betyr å trene en allerede trent modell videre med et nytt treningssett. Det vil si at modellen initialiseres med ferdigtrente vekter, og vektene oppdateres videre ved trening på de hittil usette prøvene.

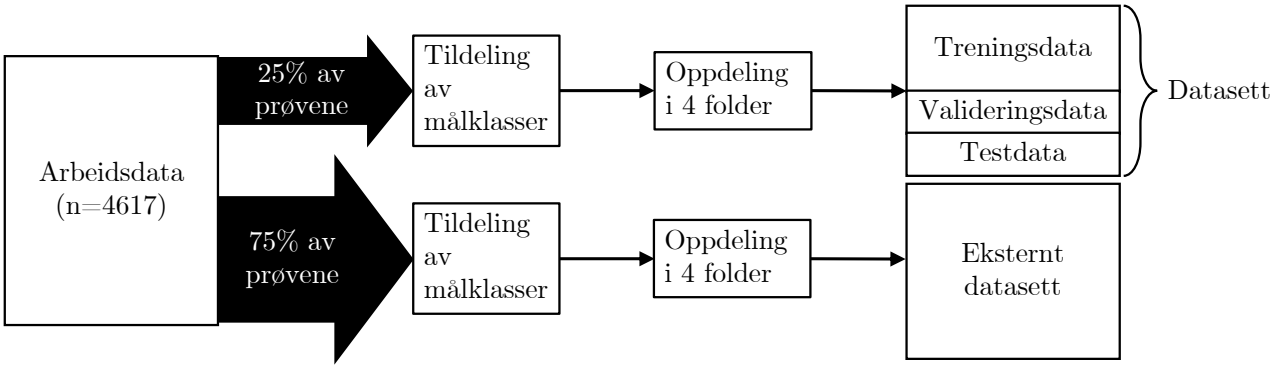
I kapittel 3.5.1 og 3.5.2 presenteres de mest brukte datasettene brukt under trening og evaluering av modeller til ulike problemstillinger i denne masteroppgaven. Datasettene kan ha innvirkning på ytelsen til modeller, og fordelingen av diagnosegrupper i hver målklassegruppe i ulike datasett diskuteres i kapittel 5. En fullstendig oversikt over alle datasett brukt i denne masteroppgaven er presentert i vedlegg A og i fila Elbow_Experiments.xlsx (se tabell 3.1).

Ved generering av datasett ble også dimensjonene til input-bildene til EfficientNet-modellene satt. Det vil si at de beskjærte bildene i ulike dimensjoner (se figur 3.6) ble omformet til en standard størrelse. Der bildestørrelsen var mindre enn standard input-størrelse ble bildene utvidet med svarte piksler (0-verdier). Standardstørrelsen ble variert mellom 640, 800 og 1280 piksler for ulike eksperimenter.

3.5.1 Binære datasett

Datasett til normal/abnormal-modeller

For binære eksperimenter var hovedmålet å skille abnormale fra normale prøver, og genereringen av datasettet brukt i de fleste eksperimentene med denne problemstillinga ble gjort med programfila gen_normal_abnormal.py (se tabell 3.1). Prosessen for generering av datasett slik det gjøres i blant annet gen_normal_abnormal.py er skissert i figur 3.7. For å ha et stort sett med ekstern data ble 25% av prøvene fra hver sykdomskategori i arbeidsdataen tilfeldig plukket ut til datasettet brukt til trening av normal/abnormal-modeller, se tabell 3.3. Siden det kun var 16 prøver i diagnosegruppa nivå 3 OCD ble det tatt ut 8 tilfeldige prøver, slik at det skulle være minst 2 mulige prøver i hvert av trenings-, valideringss- og testsettet. Dette ble gjort for å sikre at alle diagnosegruppene var representert i trenings-, valideringss- og testsett. Det ble inkludert omtrent like mange normale som abnormale prøver i datasettene, se tabell 3.3. Alle normale prøver forble klasse 0, mens alle andre prøver fikk tildelt klasse 1.

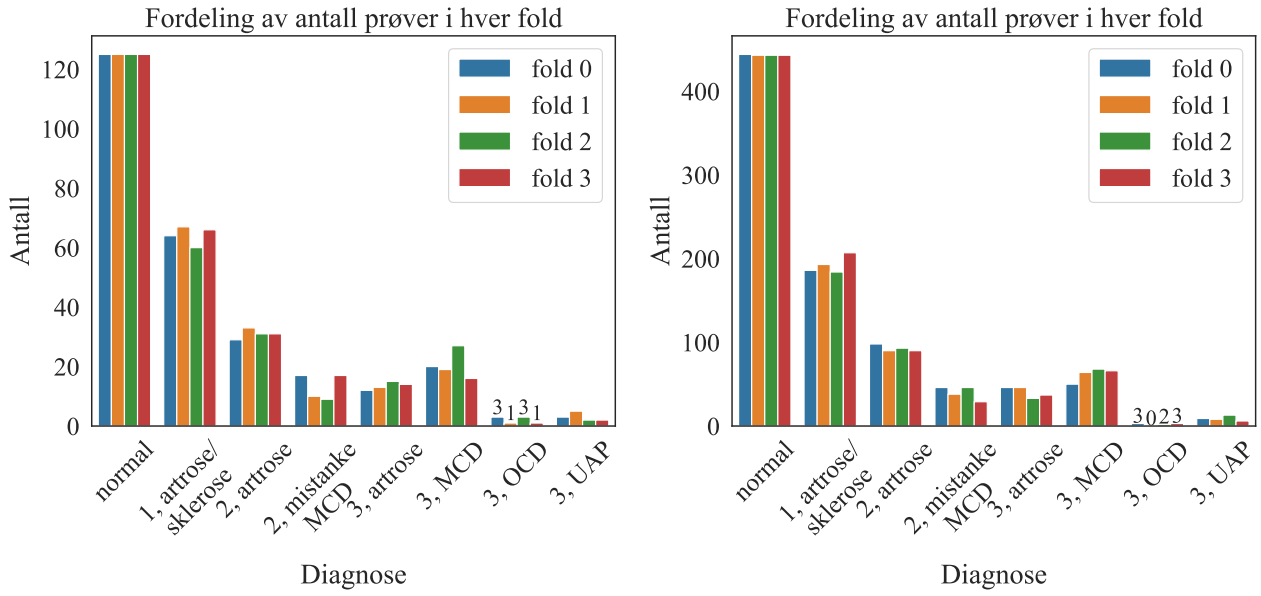


Figur 3.7: *Prosessen for generering av datasett med* gen_dataset.py*. Omtrent 25% av arbeidsdataen blir brukt i datasettet til eksperimentering med problemstillingen normal vs. abnormal. Resterende 75 % av arbeidsdataen ble brukt til eksternt datasett. Andre prosentandeler gjelder for andre datasett.*

Tabell 3.3: *Tabell med oversikt over antall bilder fra hver diagnosegruppe totalt i arbeidsdataen, og antall prøver plukket ut til datasettet brukt til trening, validering og testing (tren/val/test) av normal/abnormal-modeller. Siste kolonne i tabellen representerer antallet eksterne prøver fra hver diagnosegruppe. Eksterne prøver forble usett for modellene inntil ekstern evaluering.*

Diagnose	**Totalt**	**Tren/val/test**	**Eksterne**
Nivå 0 normal	2273	500	1773
Nivå 1 artrose	1027	257	770
Nivå 2 artrose	496	124	371
Nivå 2 MCD	212	53	159
Nivå 3 artrose	216	54	162
Nivå 3 MCD	330	82	248
Nivå 3 OCD	16	8	8
Nivå 3 UAP	49	12	36
Sum	4617	1090	3527

Etter tildeling av klasser ble datasettet delt opp i trenings-, validerings- og testsett som neste steg i Pythonfila gen_normal_abnormal.py. Dette ble gjort ved å tilfeldig splitte opp hele datasettet i 4 folder, der fold 0 og fold 1 ble brukt til trening, mens fold 2 ble brukt til validering, og fold 3 til testing, se figur 3.8a. Denne oppdelingen av datasett ble generelt brukt for all ordinær trening, validering og testing i denne masteroppgaven. Fold 3, dvs. testsettet, er prøver som forblir usett under trening av modellen, som gir en mer generalisert representasjon av ytelsen til modellen etter trening enn resultatet på valideringssettet.



(a) *Datasett brukt til trening, validering og testing*

(b) *Eksternt datasett*

Figur 3.8: *Antall prøver fra hver diagnosegruppe per fold i datasett brukt til trening og evaluering av CNNer til klassifisering av normale og abnormale prøver. (a) viser fordelingen til datasettet brukt til trening og evaluering, mens (b) viser fordelingen til det eksterne datasettet brukt til blant annet ekstern testing. Antall prøver med diagnose 3, OCD i hver fold er markert med antall over hver stolpe. Disse datasettene ble brukt til klassifisering av normale og abnormale prøver.*

I tillegg til testsett brukt for å sjekke ytelsen etter hvert eksperiment, ble det laget et eksternt datasett som besto av alle resterende prøver i arbeidsdataen, etter at prøvene brukt til trening, validering og testing var plukket ut. Prøvene i det eksterne datasettet forble usett for alle modeller inntil ekstern evaluering ble gjennomført. Tabell 3.3 gir en oversikt over antall prøver fra hver diagnosegruppe som ble brukt i hvilket datasett. Fordelingen av antall prøver fra hver diagnosegruppe i det eksterne datasettet er også vist som stolpediagram i figur 3.8b.

Under oppdeling av datasett i folder ble hver fold bestående av omtrent halvparten klasse 0 og halvparten klasse 1. Ved evaluering av modeller med eksterne datasett ble alle foldene til det eksterne datasettet brukt til evaluering. Det samme eksterne datasettet ble brukt ved omtrening av utvalgte normal/abnormal-modeller, der fold 0 og 1 ble brukt til omtrening, og fold 2 og 3 til evaluering av den omtrente modellen.

Som en utvidet analyse av feilprediksjoner av normal/abnormal-modellen med høyest ytelse ble det trent en modell som skulle klassifisere såkalte *klassifiserbare* og *ikke klassifiserbare* prøver (røntgenbilder). Datasettet til klassifiserbare vs. ikke klassifiserbare prøver inneholdt alle feilklassifiserte prøver ved ekstern evaluering av normal/abnormal-modellen med høyest ytelse. Disse feilklassifiserte prøvene ble regnet som *ikke klassifiserbare*, og var positiv klasse i denne problemstillingen. Et likt antall, tilfeldig utplukket, prøver fra eksternt datasett som ble riktig klassifisert ved evaluering av normal/abnormal-modellen med høyest ytelse ble brukt som negativ klasse, kalt *klassifiserbare* prøver.

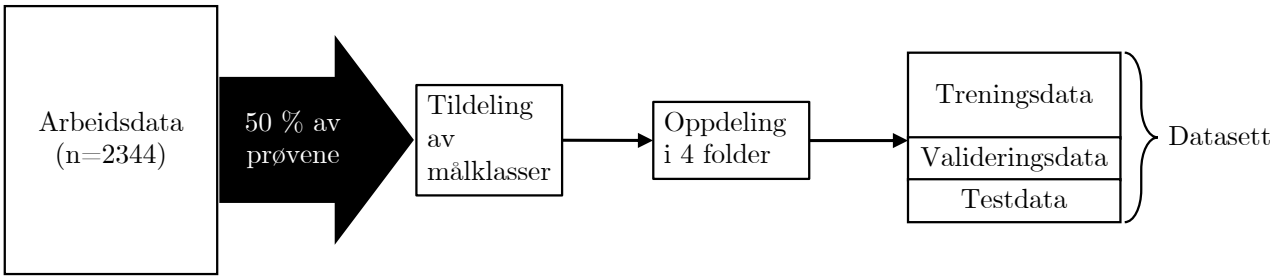
Den utvidede analysen av feilprediksjoner ved hjelp av en ny CNN ble gjort for å se om CNNen kunne finne sammenhenger i bildene som ikke framsto som åpenbare når mennesker så over bildene.

Datasett til øvrige binære modeller

To andre binære modeller ble kort utforsket for å se på EfficientNet-modellers ytelse på andre typer binære problemstillinger. I fordelingen i figur 3.2a kom det fram at det var flere prøver i nivå 1 enn 2 og 3, og det kunne derfor være tidsbesparende dersom en modell oppnår høy ytelse ved klassifisering av nivå 1 mot øvrige.

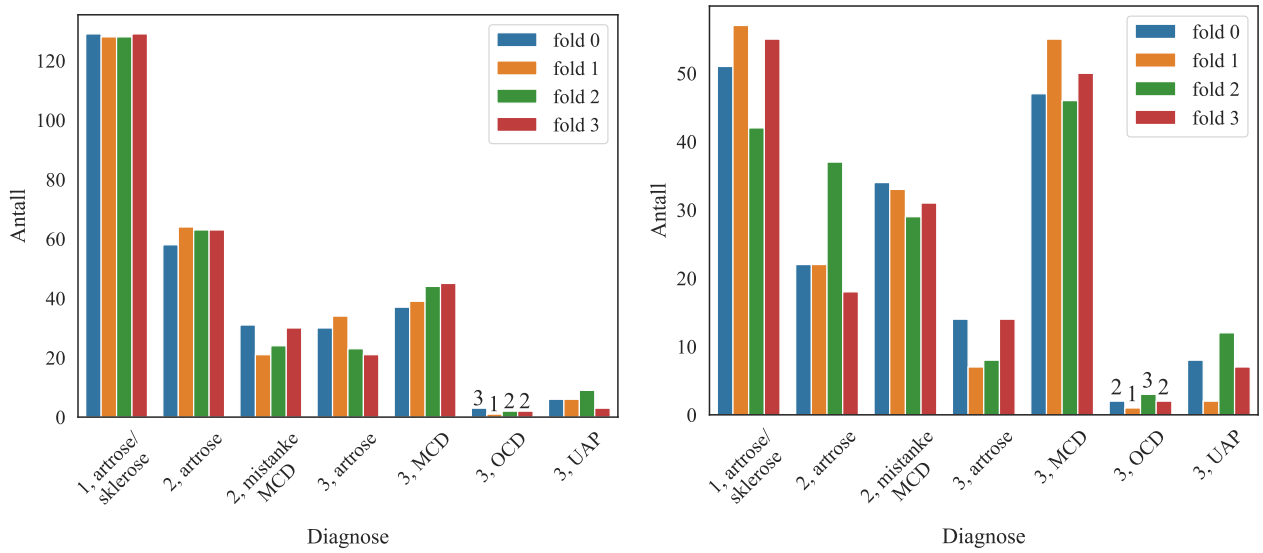
I tabell 3.2 er det også tydelig at det er flest prøver med artrose enn andre sykdommer, og det kunne derfor vært nyttig å kunne skille ut disse fra resten. På denne måten ville prosessen med å analysere bilder manuelt bli mer effektiv enn ved analyse av alle abnormale bilder, i tillegg til at modellen gir en indikator på hvor separerbare artroseprøvene er fra andre diagnosegrupper.

For klassifisering av **nivå 1 vs. øvrige nivå** og **artrose vs. øvrige diagnoser** er fordelingen av datasettene presentert i figur 3.10. For disse problemstillingene ble framgangsmåten ved generering av datasett som forklart over i kapittel 3.5.1 brukt, men med andre prosentandeler av arbeidsdataen, se figur 3.9. Det ble ikke lagd eksterne datasett for modeller ved klassifisering av nivå 1 vs. resten og artrose vs. resten.



Figur 3.9: *Prosessen for generering av datasett med* gen_dataset.py*. Omtrent 50% av røntgenbildene i arbeidsdataen med diagnose-encoding 1-7 (se tabell 3.2) ble brukt i datasettet ved klassifisering av nivå 1 vs resten. For klassifisering av artrose vs resten ble 20 % av røntgenbilder med artrose, og 60 % av prøver med andre diagnoser brukt. Det ble ikke generert eksterne datasett ved disse klassifiseringene.*

I datasettet til nivå 1 vs. resten-modeller ble halvparten av prøvene i hver diagnosegruppe brukt, som resulterte i forholdstall på omtrent 1:1 for negativ og positiv klasse. For artrose vs. resten-modeller ble 20 % av prøvene i diagnosegruppene artrose nivå 1, 2 og 3 brukt, samt 60 % av prøvene i resterende diagnosegrupper, bortsett fra nivå 3 OCD (n=8). En detaljert oversikt over antall prøver fra hver diagnosegruppe i datasettene brukt til trening, validering og testing av modeller for klassifisering av nivå 1 vs. resten og artrose vs. resten er gitt i tabell 3.4.



(a) *Nivå 1 vs. resten* (b) *Artrose vs. resten*

Figur 3.10: *Fordeling av diagnosegrupper i hver fold i datasettene brukt til trening, validering og testing ved klassifisering av (a) nivå 1 vs. resten og (b) artrose vs. resten. Antall prøver med nivå 3 OCD er markert med tall over stolpene i figuren.*

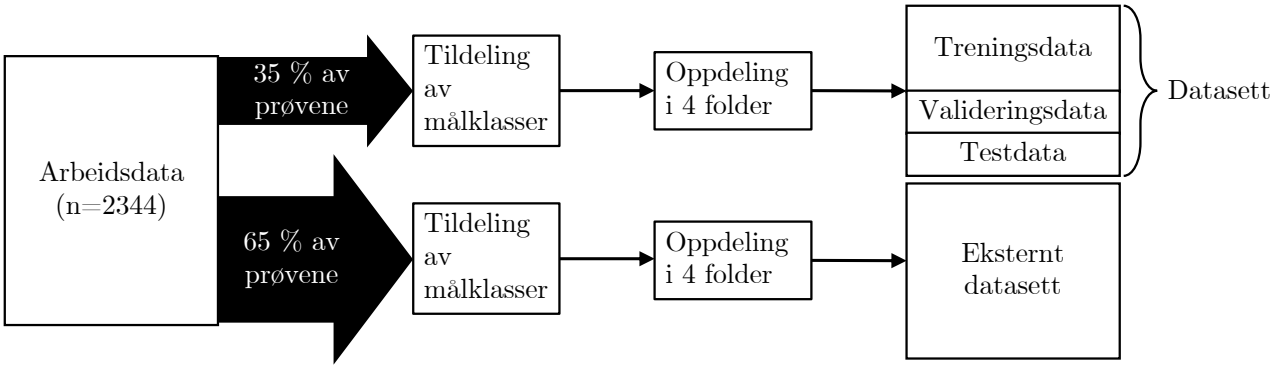
Tabell 3.4: *Tabell med oversikt over antall bilder fra hver diagnosegruppe utenom normale i arbeidsdataen, i tillegg til datasett brukt til klassifisering av nivå 1 vs. resten, og artrose vs. resten.*

Diagnose	**Totalt**	**Nivå 1 vs. resten**	**Artrose vs. resten**
Nivå 1 artrose	1027	514	205
Nivå 2 artrose	496	248	99
Nivå 2 MCD	212	106	127
Nivå 3 artrose	216	108	43
Nivå 3 MCD	330	165	198
Nivå 3 OCD	16	8	8
Nivå 3 UAP	49	24	29
Sum	2344	1173	709

3.5.2 Flerklasse-datasett

Datasett for klassifisering av nivå 1, 2 og 3

For flerklasseproblem ble samme framgangsmåte som i kapittel 3.5.1 brukt, med unntak av tildeling av klasser. Prosessen ved oppdeling av arbeidsdataen i datasett er vist i figur 3.11. Klassifisering av nivå 1, 2 og 3 er klassifiseringen som brukes ved screening av AD hos hunder per dags dato, og derfor var dette en prioritet å utforske med EfficientNet-modeller. Målet var å skille mellom de tre sykdomsnivåene på prøver med sykdom, altså utelatt normale prøver. En detaljert oversikt over antall prøver fra hver diagnosegruppe brukt i datasettene ved klassifisering av nivå 1, 2 og 3 er gitt i tabell 3.5.

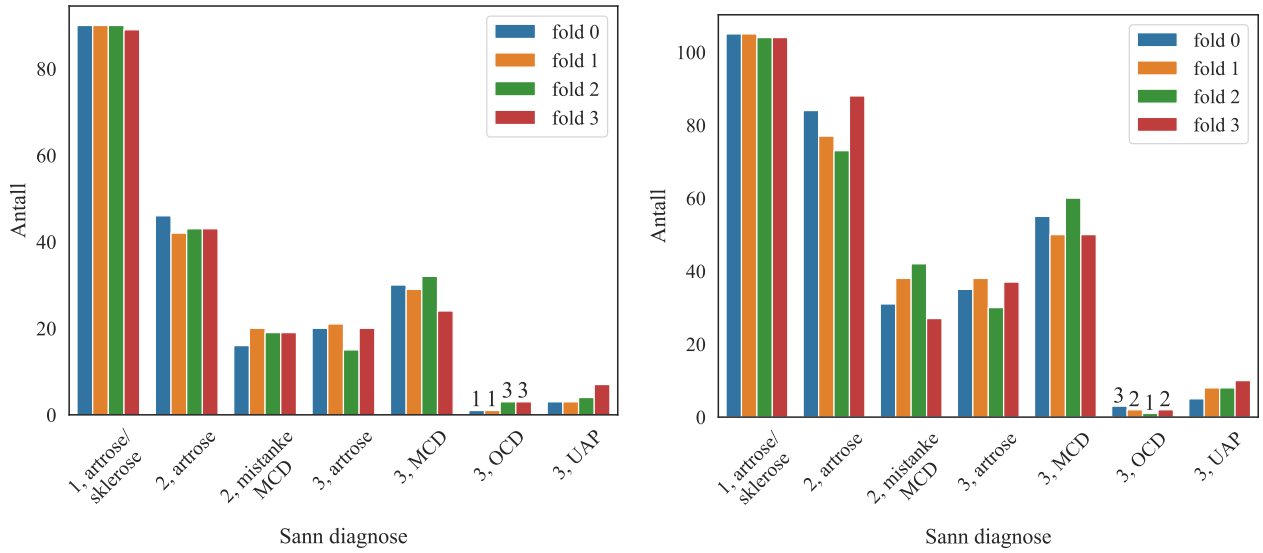


Figur 3.11: *Prosessen ved generering av datasett med* gen_datasett.py *for datasett brukt til klassifisering av nivå 1, 2 og 3. Omtrent 35 % av alle prøvene i arbeidsdataen som ikke var nivå 0, normale ble brukt til trening, validering og testing av EfficientNet-modeller. Resterende 65 % ble brukt som eksternt datasett. Samme oppdeling ble brukt ved generering av datasett til klassifisering av alle diagnosegrupper.*

Tabell 3.5: *Tabell med oversikt over antall bilder fra hver diagnosegruppe, utenom normale, i arbeidsdataen, i tillegg til datasett brukt til trening, validering og testing (tren/val/test) av modeller som klassifiserer nivå 1, 2 og 3. I siste kolonne er antall eksterne prøver fra hver diagnosegruppe gitt. For prøver brukt til trening, validering og testing av modeller som klassifiserer alle diagnosegrupper, gjelder også antall prøver gitt i tren/val/test-kolonnen i denne tabellen.*

Diagnose	**Totalt**	**Tren/val/test**	**Eksterne**
Nivå 1 artrose	1027	359	418
Nivå 2 artrose	496	174	322
Nivå 2 MCD	212	74	138
Nivå 3 artrose	216	76	140
Nivå 3 MCD	330	115	215
Nivå 3 OCD	16	8	8
Nivå 3 UAP	49	17	31
Sum	2344	823	1272

Fordelingen av prøver fra forskjellige diagnosegrupper i hver fold i datasettet brukt til klassifisering av nivå 1, 2 og 3 er presentert i figur 3.12. Figur 3.12a viser fordelingen i datasettet brukt til trening, validering og testing av ulike EfficientNet-modeller, mens figur 3.12b viser fordelingen i det eksterne datasettet. Det eksterne datasettet benyttet ved klassifisering av nivå 1, 2 og 3 ble brukt til omtrening av EfficientNet-modellen med høyest ytelse etter trening, validering og testing.



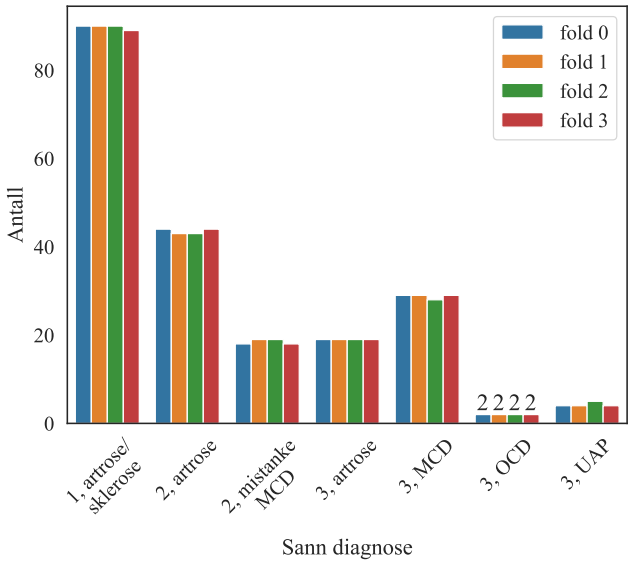
(a) *Datasett brukt til trening, validering og testing* (b) *Eksternt datasett brukt til ekstern evaluering*

Figur 3.12: *Fordeling av antall prøver fra hver diagnosegruppe i hver fold i datasett brukt til (a) trening, validering og testing av modeller ved klassifisering av nivå 1, 2 og 3 og (b) ekstern evaluering av klassifisering av nivå 1, 2 og 3. Antall prøver fra diagnosegruppe 3, OCD er notert over stolpen til tilhørende fold.*

Datasett for klassifisering av alle diagnosegrupper

Det ble også undersøkt om det var mulig å skille mellom alle 7 diagnosegrupper (se tabell 3.2), utelatt nivå 0 normale prøver. Dette ble gjort for å undersøke EfficientNet-modellers evne til å skille mellom de ulike diagnosegruppenes detaljer. Resultatet på disse modellene ble sett på som et mål på hvilke klasser som var vanskelig for modellene å skille mellom, og dette kan til en viss grad forklare hvilke prøver som ville være vanskelige å klassifisere i andre problemstillinger også.

Til denne problemstillingen ble det ikke tatt hensyn til klassebalanse. Dette kom av at det kun fantes 16 prøver totalt i gruppen med færrest prøver, i tillegg til at denne problemstillingen ble gjort som utforskning av predikerbarhet til diagnosegruppene, uten omfattende eksperimentering og evaluering. Antallet prøver fra hver diagnosegruppe brukt i datasettet for trening, validering og testing av modeller som skiller mellom alle diagnosegrupper er det samme som for trening, validering og testing av modeller som klassifiserer nivå 1, 2 og 3, se tabell 3.5. Fordelingen av prøver i hver fold i datasettet er presentert i figur 3.13, og denne er ikke lik fordelingen av prøver til datasett tilhørende klassifisering av nivå 1, 2 og 3 i figur 3.12a.



Figur 3.13: *Fordeling av antall prøver fra hver diagnosegruppe i hver fold i datasettet brukt til trening, validering og testing av modeller ved klassifisering av alle diagnosegrupper. Antall prøver fra diagnosegruppe 3, OCD er notert over stolpen til tilhørende fold.*

3.6 Konfigurering av EfficientNet-modeller

Som nevnt i kapittel 3.5, ble ulike standardstørrelser på røntgenbildene prøvd ut i ulike eksperimenter. I tillegg til å endre på standardstørrelsen, ble andre innstillinger, eller *konfigurasjoner*, variert. Dette skjer i steget kalt "fininnstilling av konfigurasjoner" i arbeidsflyten i figur 3.1, på basis av resultatene til modeller med andre konfigurasjoner. Konfigurasjonene besto av ulike kombinasjoner av læringsrate, kompleksitet og augmentering, se tabell 3.6 for detaljer. Konfigurasjonene ble generert med `gen_config.py` (se tabell 3.1), og bestemte hvordan maskinlæringsmodellen skulle settes opp.

Tabell 3.6: *Oversikt over variabler som justeres i eksperimentene.*

Variabel	**Verdier**				
Bildestørrelse	640	800	1280		
Kompleksitet	B1	B2	B3	B4	
Læringsrate	0.00005	0.0001	0.0005	0.001	0.005
Augmentering	Ja	Nei			

Modellen med lavest kompleksitet i EfficientNet-familien er EfficientNet-B0 (kompleksitet B0), som står forklart i kapittel 2.3.6. I denne masteroppgaven ble to referansemodeller basert på EfficientNet-B0 brukt, en til klassifisering av normale og abnormale prøver, og en til klassifisering av prøver fra nivå 1, 2 og 3. Ved trening av referansemodellene ble det ikke brukt augmentering, og læringsraten ble tilfeldig satt til 0.0005. Disse to modellene ble brukt som utgangspunkt for konfigurasjoner av modeller, og ytelsen ble ansett som en referanse. Målet med å prøve ut andre konfigurasjoner, var å oppnå modeller med høyere ytelse enn denne referanseytelsen.

Ved trening, validering og testing med ulike innstillinger ble det observert hvilke verdier i hver variabel som ga høyest ytelse. Fokuset ved eksperimentering var å oppnå høyest mulig ytelse med minst mulig ressursbruk. Modeller med lavest kompleksitet ble prøvd ut først, og ulike kombinasjoner av kompleksiteter og læringsrater ble prøvd ut. For utvalgte kombinasjoner av læringsrater og kompleksiteter ble også augmentering utprøvd, for å utforske om det førte til økning i ytelse.

Augmentering er en måte å utvide datasettet brukt til trening av en maskinlæringsmodell ved å gjøre små endringer på bildene før hver epoke [2]. Augmenteringsteknikkene brukt ved augmentering i denne masteroppgaven, presentert i tabell 3.7, omfattet zooming, å gjøre bildene uklare (støy), å flytte bildene langs vertikal og/eller horisontal akse (forskyvning), kontrastendringer og endring i pikselintensitet (lysstyrke). Disse teknikkene endrer utseendet til bildene noe, uten å fjerne detaljer som kjennetegner diagnosen tilhørende bildene. Andre teknikker, som rotering og flipping, ble ikke brukt, siden det allerede var stor variasjon i bildene i arbeidsdataen. Som regel ble verdi-intervallene under "standard" kolonne i tabell 3.7 brukt, men det ble utprøvd en utvidet versjon, med større intervall på noen av metodene.

Tabell 3.7: *Oversikt over metoder brukt i bildeaugmentering, med verdi-intervaller for standard og utvidet augmentering. Alle metoder bruker relative verdier til bildene slik de er i datasettet, mens forskyvning er gitt ved antall piksler forskjøvet, og støy er oppgitt som variansen til en gaussisk fordeling.*

Metode	**Standard**	**Utvidet**
Zoom	0.8-1.2	0.8-1.2
Kontrast	0.7-1.3	0.6-1.4
Forskyvning	30-30	30-30
Støy	0.05	0.05
Lysstyrke	0.8-1.2	0.7-1.3

Det ble lagt lite vekt på utforskning av ulike bildestørrelser, da det krevde mer arbeid å lage nye datasett for hver bildestørrelse, enn det det gjorde å justere øvrige innsillingsvariabler. De fleste modeller ble kjørt med høyde og bredde på 800 piksler, siden det lå rundt gjennomsnittet på beskjærte biler, se figur 3.6.

Til sammen ble det trent 29 modeller med ulike konfigurasjoner for problemstillingen normal vs. abnormal, og til sammen 72 eksperimenter ble kjørt for alle problemstillinger, inkludert ekstern evaluering. En oversikt over alle eksperimenter utført i denne masteroppgaven ligger i excel-fila Elbow_Experiments.xlsx på github, se tabell 3.1.

Ytelsesmålene brukt til å analysere ytelsen til alle binære modeller var AUC, nøyaktighet, MCC og F1. For flerklassemodeller ble AUC, nøyaktighet og MCC brukt. Disse ytelsesmålene er forklart i kapittel 2.3.7.

Kapittel 4

Resultat

4.1 Optimering av konfigurasjon

Som forklart i kapittel 3.6 ble modeller med ulike konfigurasjoner testet for å oppnå høyere ytelse enn referansemodellene. For å finne høyest helhetlig ytelse blant modellene ble MCC-score brukt, siden denne gir det mest balanserte resultatet mellom feil- og rettprediksjoner med hensyn på både positive og negative prøver, se kapittel 2.3.7. I kapittel 2.3.7 er også de andre ytelsesmålene brukt til evaluering av modeller forklart.

For hver modell som ble trent, ble resultater for alle epoker etter epoke 20 registrert. Modellene ble trent med mellom 50 og 100 epoker, og et gjennomsnitt på omtrent 70 epoker per trening. Epoken som ga høyest ytelse ble satt til den epoken som ga det høyeste gjennomsnittet på alle ytelsesmålene på valideringssettet. I gjennomsnitt ga epoke 59 høyest ytelse, men tallet varierte fra 23 til 98.

Tidsbruken per eksperiment varierte mellom omtrent 40 minutt og 9 timer, avhengig av kompleksitet på modeller, antall prøver brukt, og dataressurser tilgjengelig. For kun evaluering uten trening varierte tidsbruken mellom 16 og 60 minutter. Dette betyr omtrent 1 sekund per prediksjon per bilde på det meste. Tidsbruken er regnet for hele slurm job-en, altså inkludert blant annet oppsett av eksperiment som gjøres før kjøring av modell i Python-filene som for eksempel experiment_binary.py.

Til sammen i denne masteroppgaven ble det registrert 71 eksperiment, og en oversikt over alle eksperiment med konfigurasjoner og ytelsesmål er å finne i fila Elbow_Experiments.xlsx. Tabell 4.1 gir en oversikt over ytelsene til modellene med høyest ytelse for hver problemstilling listet under binære og flerklasseeksperiment i kapittel 3.5. Modellene er plukket ut på basis av høyest MCC-score oppnådd per problemstilling.

Tabell 4.1: *Oversikt over ytelsen til modellene med høyest MCC-score i hver problemstilling. For normal/abnormal og nivå 1 vs nivå 2 vs nivå 3 oppnådde omtrente modeller høyest MCC-score. F1-kolonnen gir F1-verdier beregnet som i ligning 2.14, mens F1_0 gir F1-verdier med hensyn på negativ klasse, som forklart i kapittel 2.3.7.*

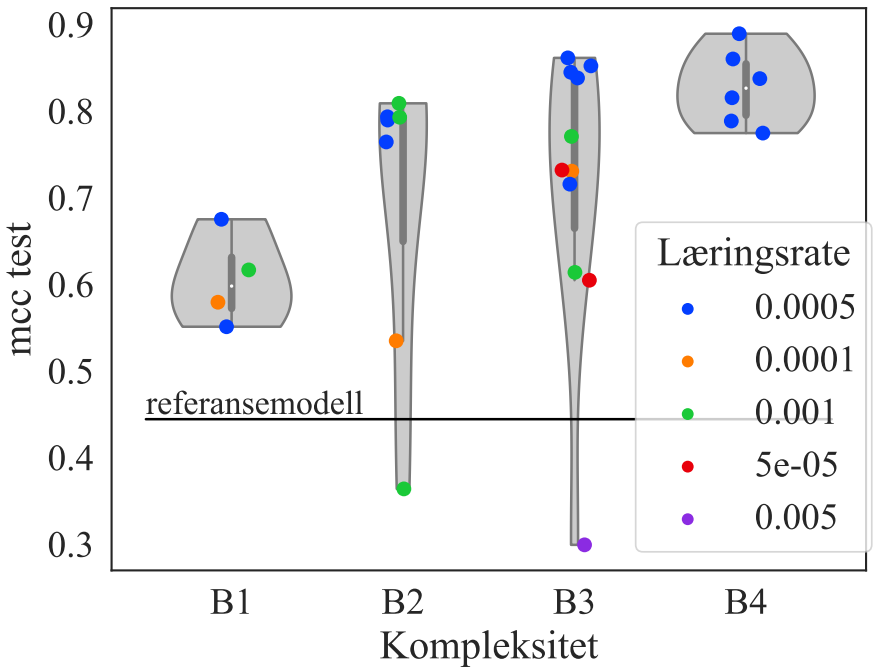
Problemstilling	**Nøyaktighet**	**MCC**	**F1**	**F1_0**	**AUC**
Normal/abnormal	0.956	0.912	0.955	0.956	0.988
Nivå 1 vs andre	0.737	0.477	0.752	0.720	0.781
Artrose vs andre	0.718	0.446	0.691	0.740	0.772
Klassifiserbare vs. ikke klassifiserbare	0.588	0.180	0.625	0.543	0.659
Nivå 1 vs nivå 2 vs nivå 3	0.668	0.502	-	-	0.845
Alle 7 diagnosegrupper	0.597	0.422	-	-	0.864

4.2 Binære modeller

4.2.1 Normal vs. abnormal

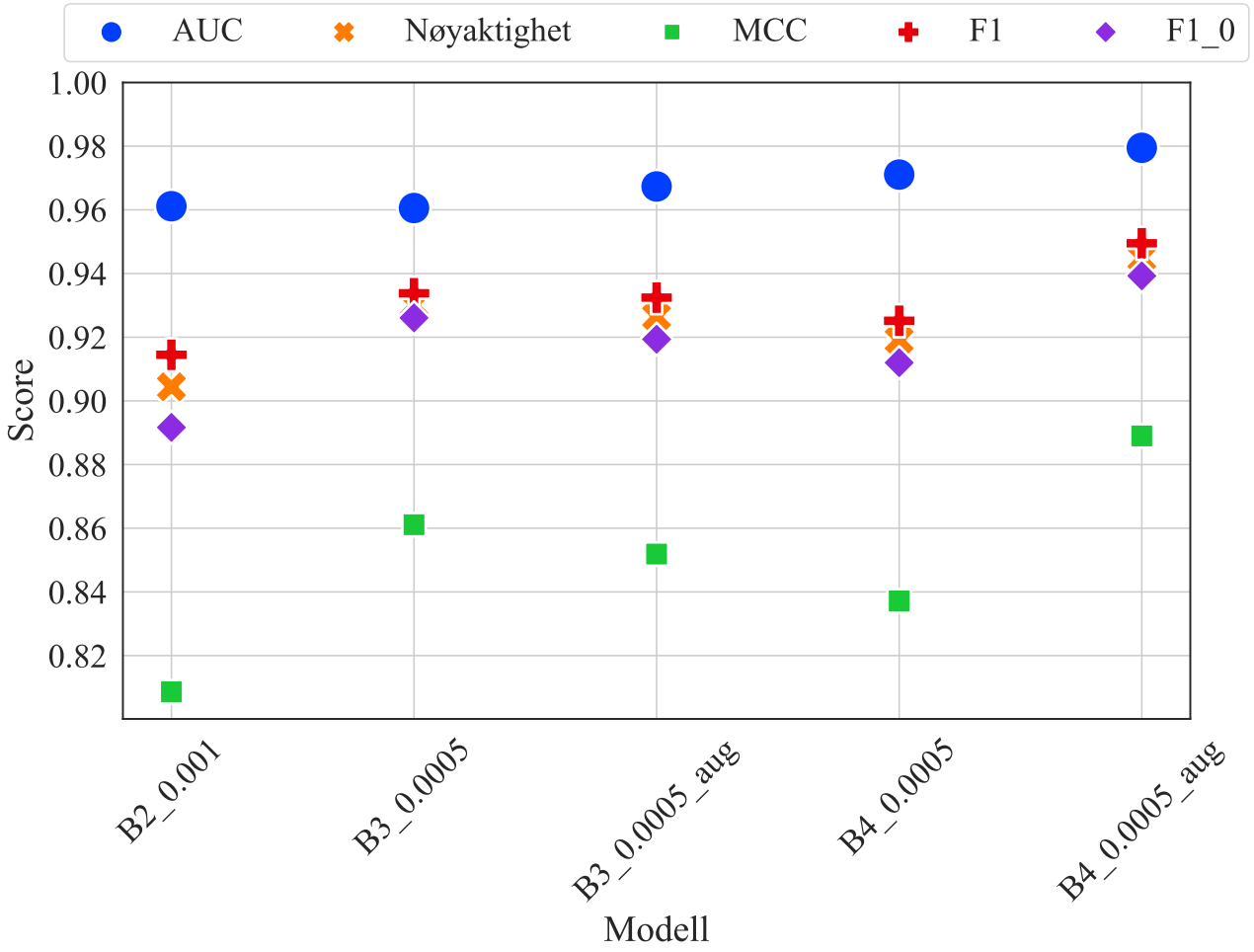
Det ble trent 29 modeller for problemstillingen med klassifisering av normale og abnormale prøver. I tillegg ble to av modellene med høyest ytelse evaluert med det eksterne datasettet, altså på alle prøvene som ikke var brukt i treningen av modellen, se kapittel 3.5.1 for detaljer. Dette ble gjort for å se den generaliserte ytelsen til modellen, i tilfelle noe av informasjonen i testsettet under trening hadde lekket inn i modellen.

En oversikt over MCC-scoren til testsettet til alle 29 trente modeller er vist i figur 4.1, med MCC-scoren til referansemodellen markert som en svart linje. Både B3- og B4-kompleksitetene ga høyere score enn B2 og B1. I tillegg virker det som at en læringsrate på 0.0005 gir høyest ytelse for alle modellkompleksiteter, og derfor ble denne læringsraten brukt i flest modeller.



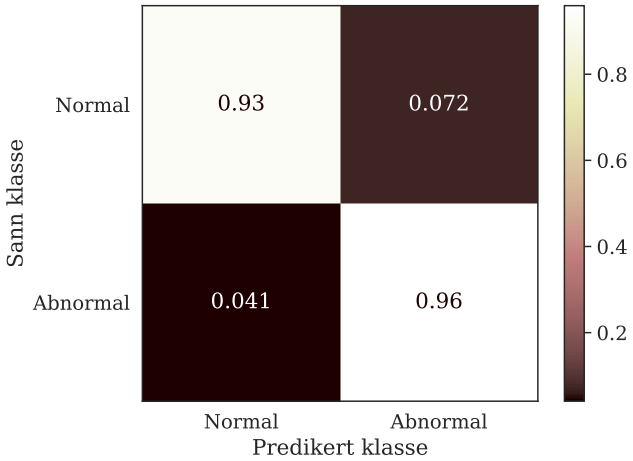
Figur 4.1: *Oversikt over MCC-scoren til testsettet til alle 29 modeller som ble trent til å skille mellom normale og abnormale prøver, inkludert referansemodellen med kompleksitet B0 og læringsrate 0.0005. Hvert punkt i figuren tilsvarer et eksperiment, og den svarte linja tilsvarer referansemodellens MCC-score.*

De fem modellene med høyest ytelse på testsettet ved klassifisering av normale og abnormale prøver er vist i figur 4.2. Alle modellene ga en nøyaktighet på mellom 0.92 og 0.93 på testsettet. MCC-scoren varierte mellom omtrent 0.82 og 0.89, som viser god balanse mellom feil- og sannprediksjoner i begge målklasser. Modellen med aller høyest ytelse på testsettet var B4_0.0005_aug, det vil si modellen med kompleksitet B4, læringsrate 0.0005, og med standard augmentering.



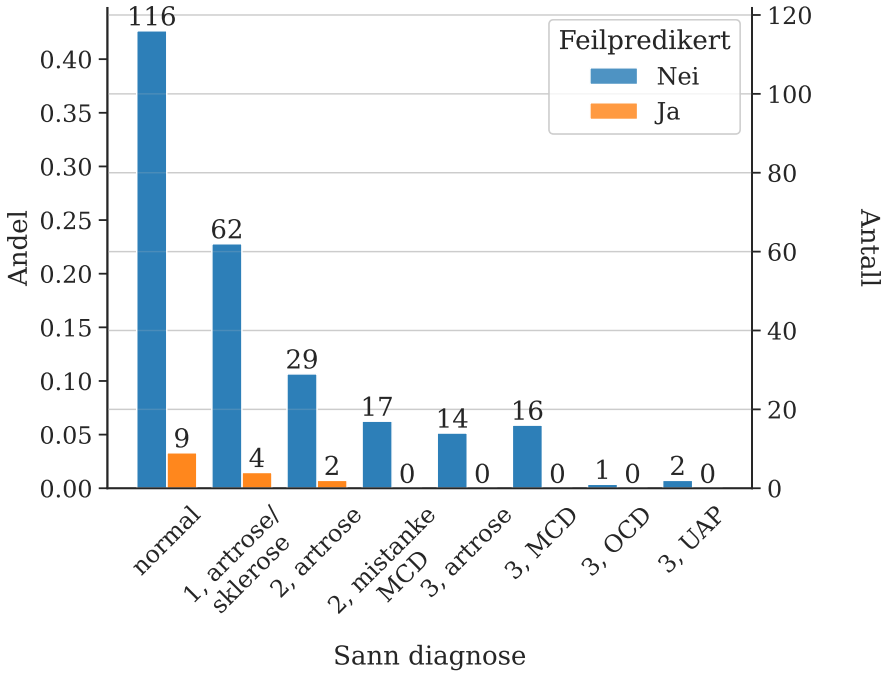
Figur 4.2: *De fem modellene med høyest helhetlig ytelse ved testing av trente modeller som klassifiserte normale og abnormale prøver. Alle, bortsett fra én av de fem modellene, hadde læringsrate 0.0005, og kompleksitet B2, B3 eller B4. Alle modellene hadde MCC-score på mellom 0.8 og 0.9, og nøyaktigheter på mellom 0.90 og 0.94 ved testing. To av disse fem modellene ble trent med augmentering, resten hadde ikke augmentering. Navnene til modellene, slik de står oppført på x-aksen, refererer til kompleksitet og læringsrate. Dersom augmentering ble brukt, ender navnet til modellen på "_aug".*

Modellen kalt B4_0.0005_aug ble analysert videre, og forvirringsmatrisa tilhørende modellen er presentert i figur 4.3. Det kommer fram at det er en høyere andel sanne positive (96 %) enn negative (93 %). Dette vises også av F1- og F1_0-scorene i figur 4.2, men for B4_0.0005_aug ligger disse scorene nærmere hverandre enn for nesten alle de andre modellene i figuren. Det vil si at det er relativt god balanse mellom sanne positive og sanne negative for modellen med høyest ytelse ved klassifisering av normale og abnormale prøver.



Figur 4.3: *Forvirringsmatrisa til den modell B4_0.0005_aug, som hadde høyest ytelse på testsettet blant alle modeller som klassifiserte normale og abnormale prøver (se figur 4.2). Matrisa viser hvor stor andel av normale og abnormale prøver som ble predikert riktig og feil. 7,2% av de normale prøvene ble predikert som abnormale.*

En detaljert oversikt over hvilken diagnose hver av de feilpredikerte prøvene fra testsettet til modell B4_0.0005_aug hadde er presentert i søylediagrammet i figur 4.4. Det kommer fram av figuren at de feilpredikerte prøvene var fra følgende diagnosegrupper: normal, nivå 1 artrose og/eller sklerose og nivå 2 artrose.

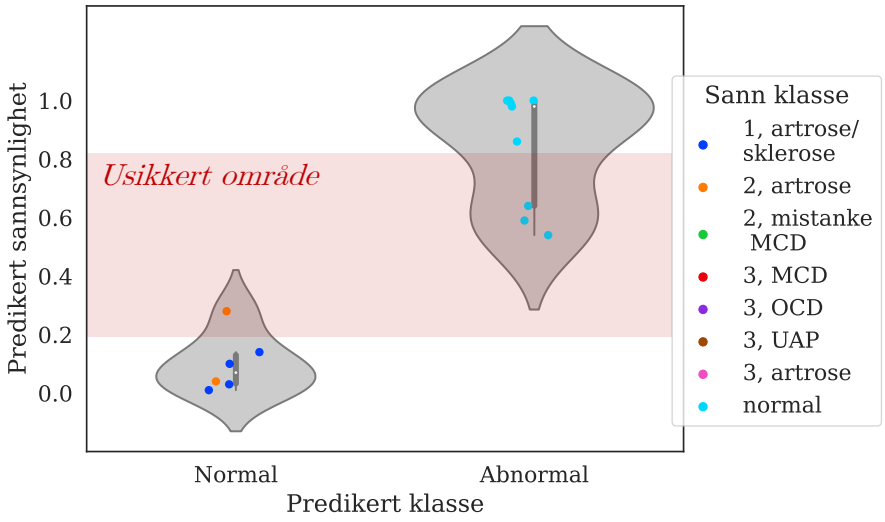


Figur 4.4: *En oversikt over hvor stor andel av prøver i testsettet som ble feil og riktig klassifisert for hver diagnosegruppe. Dette er resultatet etter testing av modell B4_0.0005_aug, som hadde høyest ytelse for klassifisering av normale og abnormale prøver (se figur 4.2). Høyre akse og tallene på toppen av stolpene angir antall prøver i absolutte tall.*

Også modellens "selvsikkerhet" ved prediksjon ble analysert. Selvsikkerhet regnes som hvor

sikker modellen er på at en prøve tilhører positiv eller negativ klasse, altså hvor stor sannsynlighet den avgjør at prøven har for å tilhøre positiv klasse. Dersom prøven er predikert som abnormal med en sannsynlighet på 1.0 har prøven 100% sannsynlighet for å være abnormal i følge modellen. Dersom prøven har mindre enn 50% sannsynlighet for å tilhøre abnormal klasse regnes den som normal klasse. "Sikre" prediksjoner kan regnes som prediksjoner rundt 1.0 og 0.0, og i denne masteroppgaven er intervallet 0.2-0.8 brukt som et utgangspunkt for "usikre" prediksjoner.

Fordelingen av sannsynligheter for hver prøve til modellen er delt opp i to fiolinplott, én for feilpredikerte prøver (figur 4.5) og én for riktig predikerte prøver (figur 4.7).

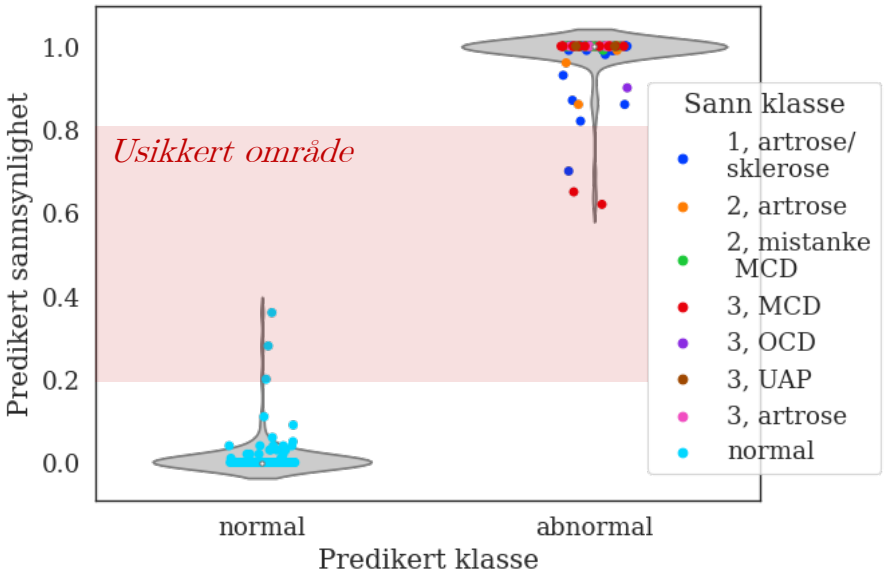


Figur 4.5: *Fordeling av predikerte sannsynligheter for feilpredikerte prøver i testsettet til modellen med høyest helhetlig ytelse (B4_0.0005_aug) i problemstillingen normal vs. abnormal, se figur 4.2. Alle feilpredikerte prøver er markert i figuren som prikker med farge tilhørende sann diagnose. "Usikkert område", farget rødt, er intervallet av sannsynligheter som regnes som usikre prediksjoner (her omtrent 0.2-0.8). For eksempel vil en prøve predikert som abnormal med sannsynlighet 0.56, regnes som "usikker".*

Få av de feilpredikerte prøvene i figur 4.5 befinner seg i nærheten av 1.0 eller 0.0, altså ser det ut til at mange feilpredikerte prøver ble predikert med lav grad av sikkerhet. Dette kommer også fram av fordelingene i fiolinplottet, som gir sannsynlighetstettheten til prøver for ulike prediksjonssannsynligheter. Fiolin-fordelingene i figur 4.5 viser høyest tetthet av prediksjoner med sannsynligheter under 0.2 for prøver predikert som normale, og over 0.8 for prøver predikert som abnormale. Noen av røntgenbildene tilhørende feilpredikerte prøver er vist i figur 4.6, og en oversikt over alle røntgenbildene til feilprediksjonene i figur 4.5 er å finne i vedlegg B.

Figur 4.6: *Eksempler på feilpredikerte prøver ved klassifisering av normale og abnormale prøver med modell B4_0.0005_aug. Hvert bilde er markert med diagnosegruppen den ble markert å tilhøre av veterinærer, "sann diagnose". "Predikert sannsynlighet" er sannsynligheten for at hvert bilde tilhører abnormal klasse, predikert av modell B4_0.0005_aug. Dersom en prøve har en predikert sannsynlighet på mer enn 0.5, klassifiseres den som abnormal, dersom sannsynligheten er under 0.5, klassifiseres den som normal.*

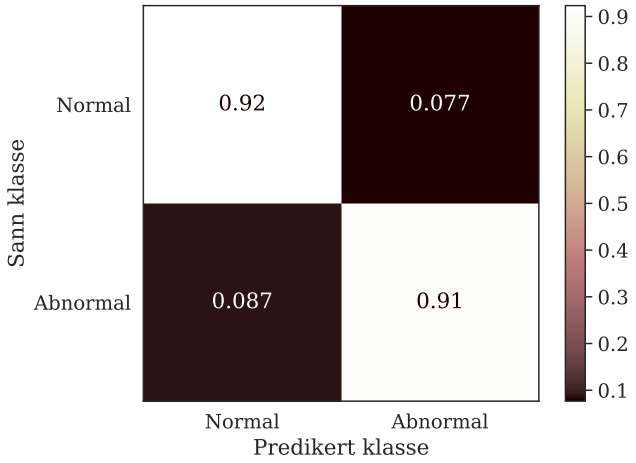
For riktig predikerte prøver er det tydelig fra fiolinplottet i figur 4.7 at tettheten på fordelingen av sannsynligheter er størst i ytterpunktene, dvs. ved sannsynligheter rundt 0.0 og 1.0. Dette viser at modellen har predikert riktig med stor sikkerhet, og kun et fåtall prøver er predikert med sannsynligheter på "usikkerhetsintervallet" (0.2-0.8). Til sammenligning er det høyere tetthet for prediksjoner på det usikre intervallet i figur 4.5, med feilprediksjoner.



Figur 4.7: *Fordeling av predikerte sannsynligheter for riktig predikerte prøver i testsettet til modellen med høyest helhetlig ytelse (B4_0.0005_aug) i problemstillingen normal vs. abnormal, se figur 4.2. Hver prikk på figuren tilsvarer en prøve, og fargen indikerer hvilken diagnosegruppe prøven faktisk tilhører. "Usikkert område", farget rødt, er intervallet av sannsynligheter som regnes som usikre prediksjoner (her omtrent 0.2-0.8).*

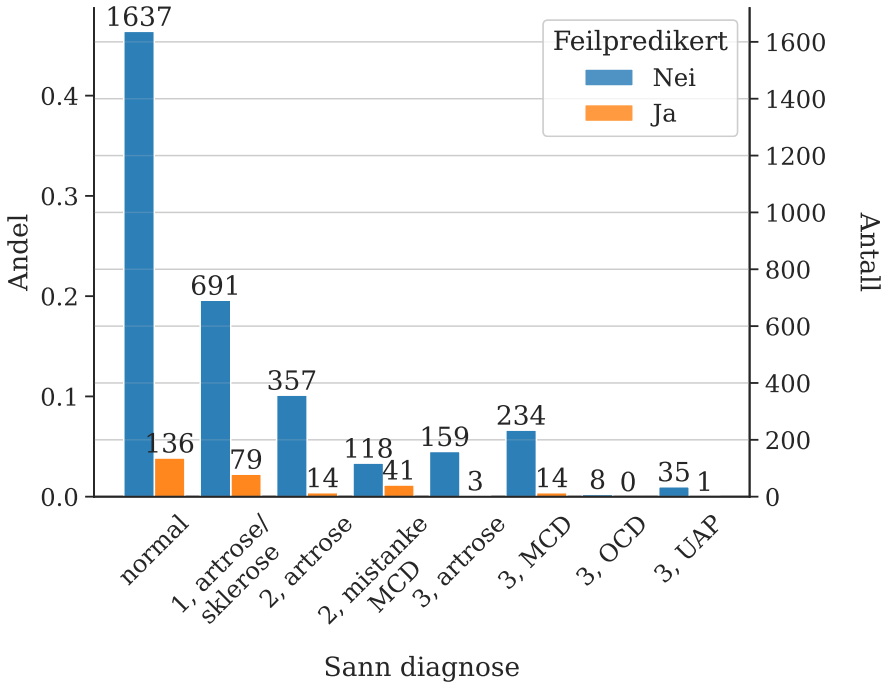
4.2.2 Ekstern evaluering, normal vs. abnormal

For å få et bedre bilde på hvor godt modell B4_0.0005_aug yter på et tilfeldig utvalg bilder, ble modellen evaluert ved å predikere klassetilhørighet til prøver i det eksterne datasettet (se kapittel 3.5.1). Disse bildene var helt usett for modellen, siden de ikke var brukt til å trene, validere eller teste noen av modellene med problemstillingen normal vs. abnormal tidligere. Forvirringsmatrisa til den eksterne evalueringen er gitt i figur 4.8, og viser at andel sanne positive og negative prøver var noe lavere enn resultatet på testsettet til modellen, gitt i figur 4.3.



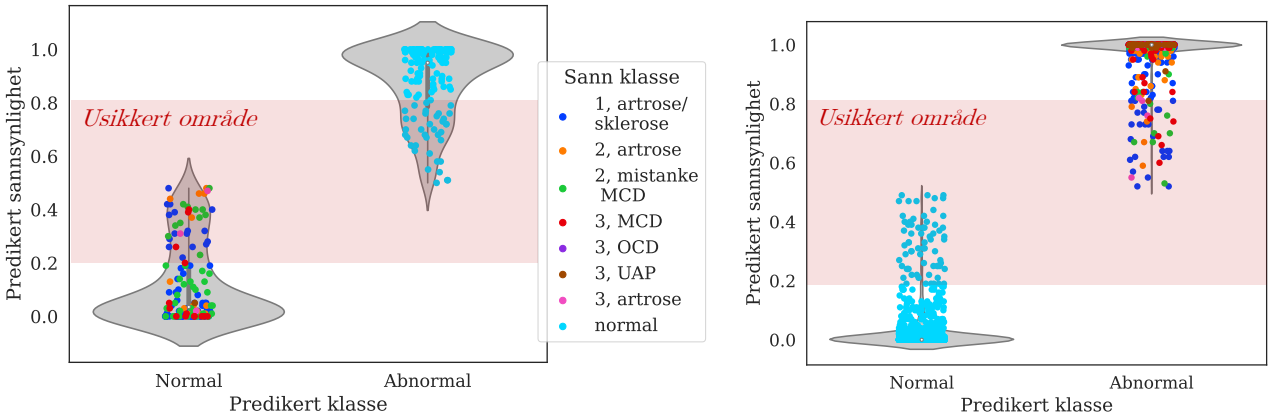
Figur 4.8: *Forvirringsmatrisa til ekstern evaluering av modellen med høyest ytelse ved klassifisering av normale og abnormale prøver, B4_0.0005_aug. Matrisa viser hvor stor andel av sann klasse som ble predikert riktig og feil. 7,7% av de normale prøvene ble predikert som abnormale.*

Blant de feilpredikerte prøvene i den eksterne evalueringen var alle diagnosegrupper representert, bortsett fra nivå 3 OCD, se figur 4.9. Det var flest nivå 0 normale, nivå 1 artrose og/eller sklerose og nivå 2 mistanke MCD i den feilpredikerte gruppen. Totalt ble 288 prøver fra det eksterne datasettet feilpredikert.



Figur 4.9: *En oversikt over hvor stor andel av prøver i ekstern evaluering som ble feil og riktig klassifisert for hver diagnosegruppe. Dette er resultatet etter klassifisering av prøver fra det eksterne datasettet med modell B4_0.0005_aug, som hadde høyest ytelse blant alle normal vs. abnormal-modeller. Bare prøver fra nivå 3 OCD hadde ingen forekomst av feilklassifisering. Høyre akse og tallene på toppen av stolpene angir antall prøver i absolutte tall.*

I den eksterne evalueringen av modell B4_0.0005_aug (se kapittel 4.2.1) var flere prøver predikert med lav selvsikkerhet (sannsynlighet 0.2-0.8) av modellen, se figur 4.10. Fiolinplottene i figur 4.10a viser at tettheten av usikre prediksjoner var lavere for riktig predikerte prøver enn feilpredikerte prøver i figur 4.10b.



(a) Feilklassifiserte prøver.

(b) Riktig klassifiserte prøver.

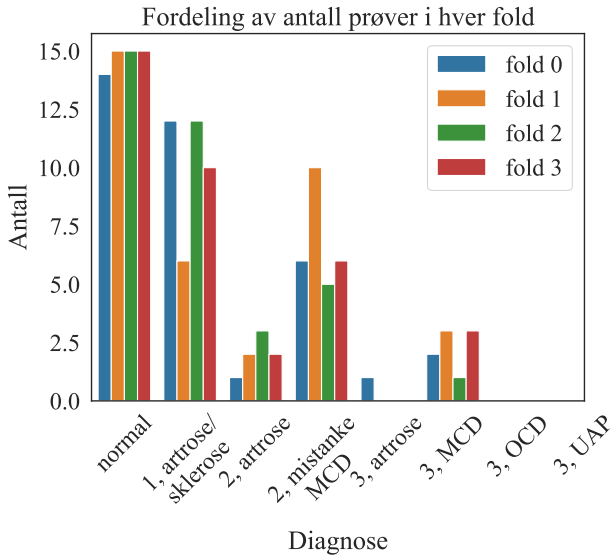
Figur 4.10: *Fordeling av predikerte sannsynligheter for rett- og feilklassifiserte prøver i den eksterne testen på normal/abnormal modell med høyest ytelse (B4_0.0005_aug, figur 4.2). Hver prikk på figuren tilsvarer en prøve, og fargen indikerer hvilken diagnosegruppe prøven faktisk tilhører. Venstre figur (a) viser alle feilklassifiserte prøver, høyre (b) viser alle riktig klassifiserte prøver. "Usikkert område", farget rødt, er intervallet av sannsynligheter som regnes som usikre prediksjoner (her omtrent 0.2-0.8).*

Omtrening av normal/abnormal-modeller

Det ble også gjort omtrening av feilklassifiserte prøver på modellen med høyest ytelse (B4_0.0005_aug, se figur 4.2) etter analyse av den eksterne evalueringen. Dette ble gjort på to måter:

1. trene på halvparten av alle feilklassifiserte prøver i ekstern evaluering,
2. trene på halvparten av alle prøvene i eksternt datasett, med andre halvdel som testsett.

Første omtreningsmetode tok i bruk 144 av 288 feilpredikerte prøver fra den eksterne evalueringen, se figur 4.10. Modell B4_0.0005_aug ble initialisert med de allerede trente vektene, og trent videre med 144 av de feilpredikerte prøvene fra ekstern evaluering i figur 4.9. Dette ble gjort for å se om modellen kunne justere vektene for å tilpasse seg prøver som syntes å være vanskelige å klassifisere ved ekstern evaluering. Datasettet brukt til omtrening med kun feilpredikerte prøver er vist i figur 4.11, med fordeling av antall prøver fra hver diagnosegruppe i hver fold. Fold 0, 1 og 2 ble brukt til trening av modellen, mens fold 3 ble brukt til validering, og alle fold ble brukt som testsett. For å evaluere modellen ble alle prøver i det eksterne datasettet vist i tabell 3.3 brukt, untatt de 144 prøvene brukt til omtrening.



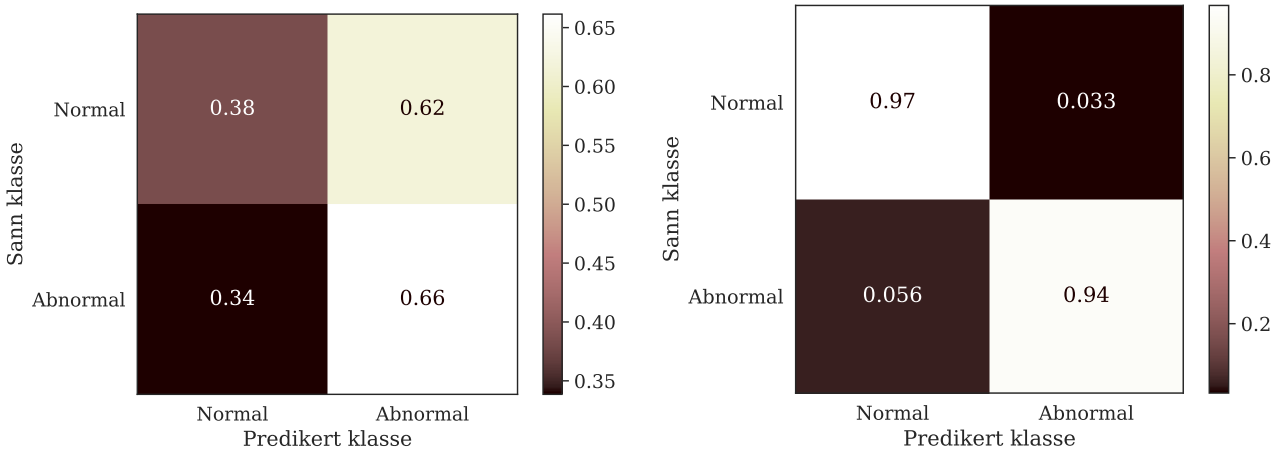
(a)

Figur 4.11: *Fordelingen av prøver fra hver diagnosegruppe i hver fold i datasettet brukt til omtrening av normal/abnormal-modell med kun feilpredikerte prøver fra ekstern evaluering, se figur 4.10.*

Omtreningsmetode nummer to som nevnt over var å trene modell B4_0.0005_aug (se figur 4.2) videre på halvparten av det eksterne datasettet til normal/abnormal-modellen, se kapittel 4.2.1. Det ble antatt at feilpredikerte prøver var representert i hele det eksterne datasettet. Fold 0 og 1 av det eksterne datasettet i figur 3.8 ble brukt til omtrening, mens fold 2 og 3 ble brukt til evaluering av den omtrente modellen.

Forvirringsmatrisene i figur 4.12 viser resultatene etter omtrening med (a) kun feilpredikerte prøver, og (b) med halvparten av det eksterne datasettet til normal/abnormal-modellene (se kapittel 3.5.1).

Forvirringsmatrisa til omtrening med kun feilpredikerte prøver viser nedgang i ytelse fra forvirringsmatrisa i figur 4.3, med kun 38 % sanne negative prøver.

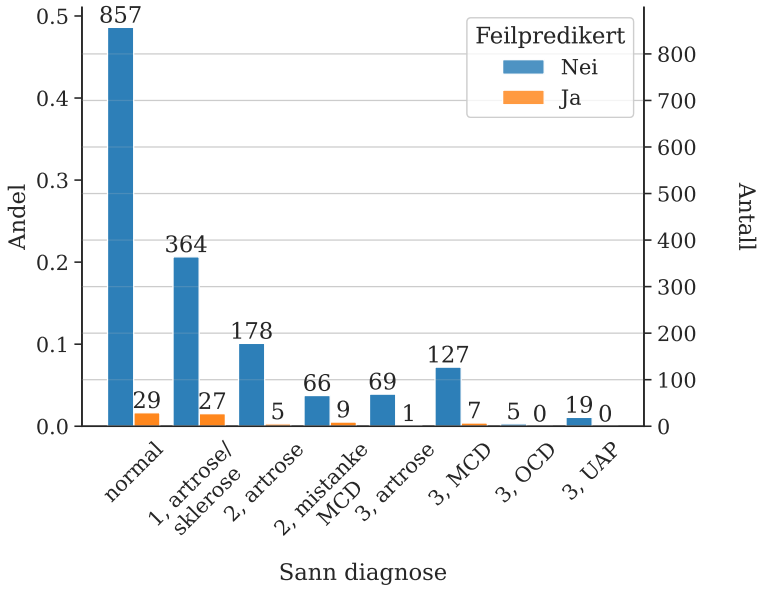


(a) *Omtrening med kun feilklassifiserte prøver.*

(b) *Omtrening med halvparten av det eksterne datasettet.*

Figur 4.12: *Resultat etter omtrening av modell B4_0.0005_aug, som hadde høyest ytelse i normal vs. abnormal problemstillingen. (a) Forvirringsmatrisa ved evaluering av modellen (B4_0.0005_aug) etter å ha trent den videre på halvparten av alle feilklassifiserte prøver i den eksterne evalueringen. (b) Forvirringsmatrisa etter omtrening av modell B4_0.0005_aug på halve det eksterne datasettet.*

Etter omtrening av modell B4_0.0005_aug med halve det eksterne datasettet (metode to, figur 4.12b) var ingen feilpredikerte prøver fra diagnosegruppe nivå 3 OCD og nivå 3 UAP, og bare én prøve fra nivå 3 artrose ble feilpredikert, se figur 4.13. Andelen feilpredikerte i hver diagnosegruppe gikk ned for alle grupper i forhold til ekstern evaluering av modellen før omtrening, se figur 4.9.

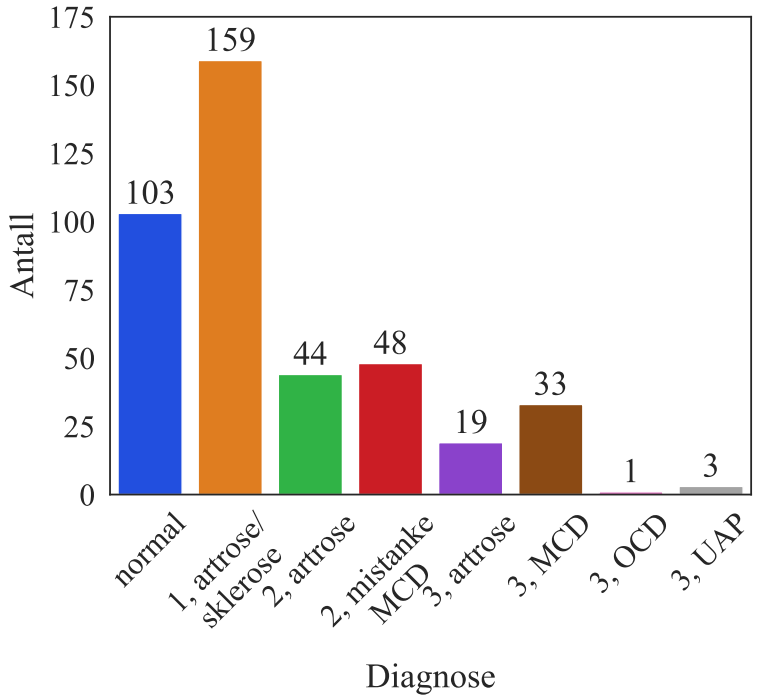


Figur 4.13: *Fordeling av korrekt og feilpredikerte prøver etter original diagnose. Det totale antallet prøver i denne evalueringen er 1767 prøver, altså halvparten av antallet prøver brukt i ekstern evaluering av modell B4_0.0005_aug, se figur 4.9. Høyre akse og tallene på toppen av stolpene angir antall prøver i absolutte tall.*

4.2.3 Andre binære problemstillinger

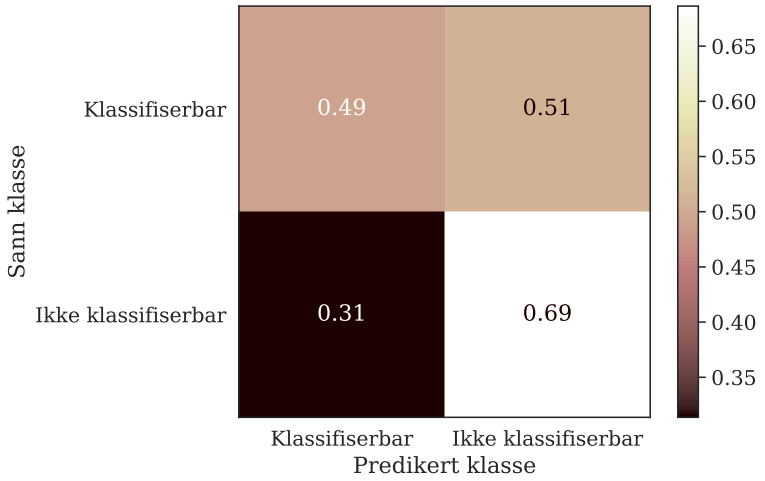
Utvidet analyse: klassifiserbare vs. ikke klassifiserbare

Som en videre analyse av feilprediksjoner i den eksterne evalueringen til normal/abnormal-modellen (figur 4.10) ble det undersøkt om en ny modell kunne trenes opp til å skille mellom feilpredikerte og riktig predikerte prøver. Det vil si at alle feilpredikerte prøver i figur 4.10 ble merket som "ikke klassifiserbare". Disse prøvene ble plukket ut til et datasett sammen med riktig predikerte prøver fra samme eksterne evaluering, som forklart i kapittel 3.5.1. Datasettet besto av prøver fra alle diagnosegrupper, som vist i figur 4.14.



Figur 4.14: *Fordelingen av prøver i datasettet brukt til å skille mellom klassifiserbare og ikke klassifiserbare prøver. Datasettet besto av feil- og riktig predikerte røntgenbilder fra ekstern evaluering av normal/abnormal-modellen, se figur 4.10.*

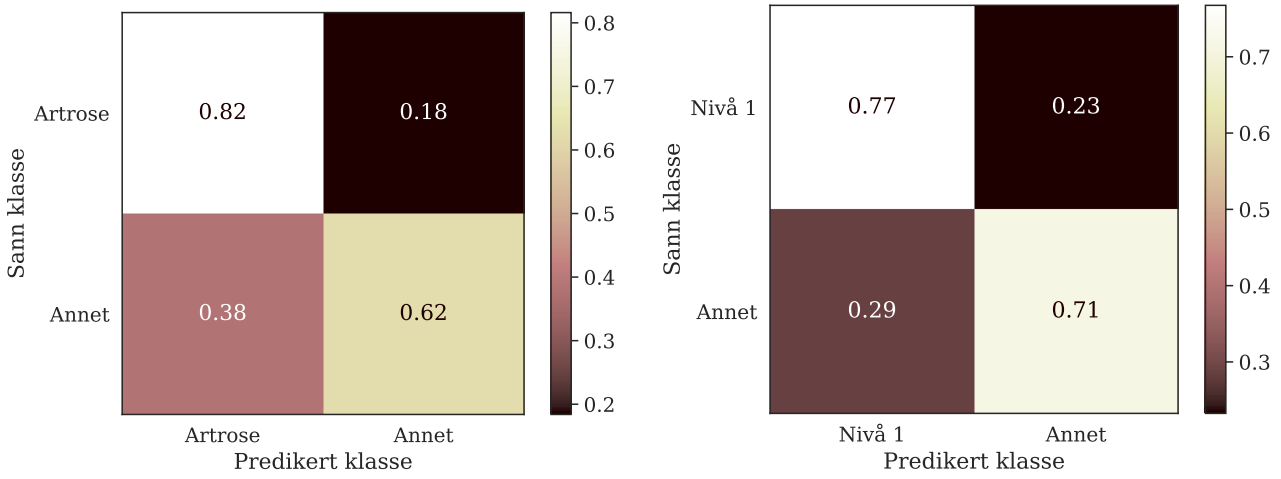
Klassifiseringen av "klassifiserbare" og "ikke klassifiserbare" prøver oppnådde en nøyaktighet på 0.59 og MCC-score på 0.18, se tabell 4.1. Forvirringsmatrisa i figur 4.15 viser at under halvparten av alle prøver som ble riktig klassifisert i den eksterne evalueringen (kapittel 4.2.2), ble kjent igjen som "klassifiserbare". På den andre siden ble nesten 70 % av alle tidligere feilpredikerte prøver gjenkjent.



Figur 4.15: *Forvirringsmatrisa til resultatet av opptrening av modell som predikerte om prøver var klassifiserbare eller ikke.*

Øvrige modeller

I tillegg til å trene modeller som klassifiserte normale og abnormale prøver ble det trent modeller som skulle skille prøver med nivå 1 fra resten, og prøver med artrose fra resten. Disse to modellen oppnådde MCC-scores så vidt over referansemodellen til normal/abnormal problemstilling, se tabell 4.1 for høyest oppnådde ytelse for ulike problemstillinger. Det ble ikke gjort omfattende analyser av resultatene til disse to problemstillingene, men forvirringsmatrisene til begge modellene er vist i figur 4.16.



(a) *Artrose vs. resten* (b) *Nivå 1 vs. resten*

Figur 4.16: *Forvirringsmatriser til binære modeller som skulle skille (a) artrose fra resten og (b) nivå 1 fra resten.*

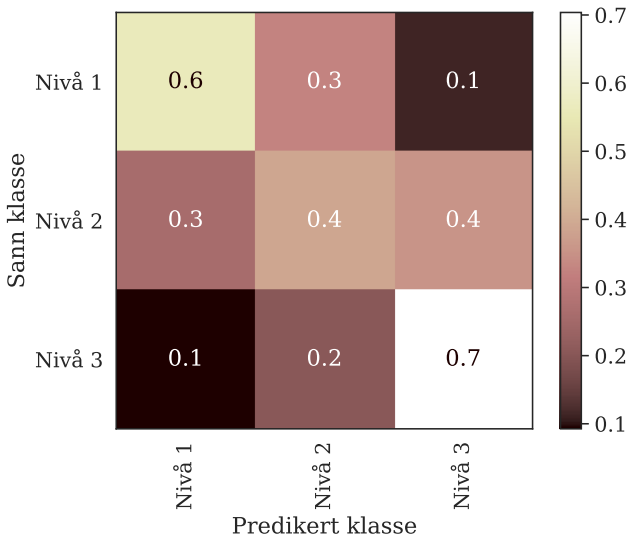
Forvirringsmatrisene til både klassifisering av nivå 1 vs. resten og artrose vs. resten viser høyere andel sanne negative enn positive prøver. Artrose vs. resten-modellen predikerte nesten 40 % falske negative, mens nivå 1 vs. resten-modellen predikerte 29 % falske negative og 23 % falske positive prøver.

4.3 Flerklassemodeller

Blant utprøvde flerklassemodeller var det klassifisering av nivå 1, 2 og 3 som sto i fokus. Det ble også gjennomført eksperiment med skille mellom alle syv diagnosegrupper med gradering, se kapittel 3.5.2.

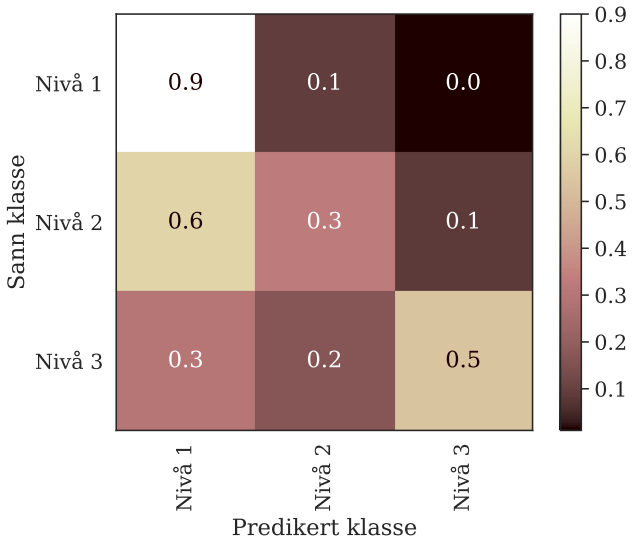
4.3.1 Klassifisering av nivå 1, 2 og 3

Det ble trent en referansemodell for klassifisering av nivå 1, 2 og 3, med kompleksitet B0 og læringsrate 0.0005. Modellen oppnådde en nøyaktighet på 0.546 og MCC-score på 0.321 på testsettet, se figur 4.17. Målet med å prøve ulike konfigurasjoner på modeller ved klassifisering av nivå 1, 2 og 3 var å oppnå høyere ytelse enn dette. Fem modeller oppnådde lavere MCC-score, mens til sammen ni modeller oppnådde høyere score, se fila Elbow_Experiments.xlsx i tabell 3.1.



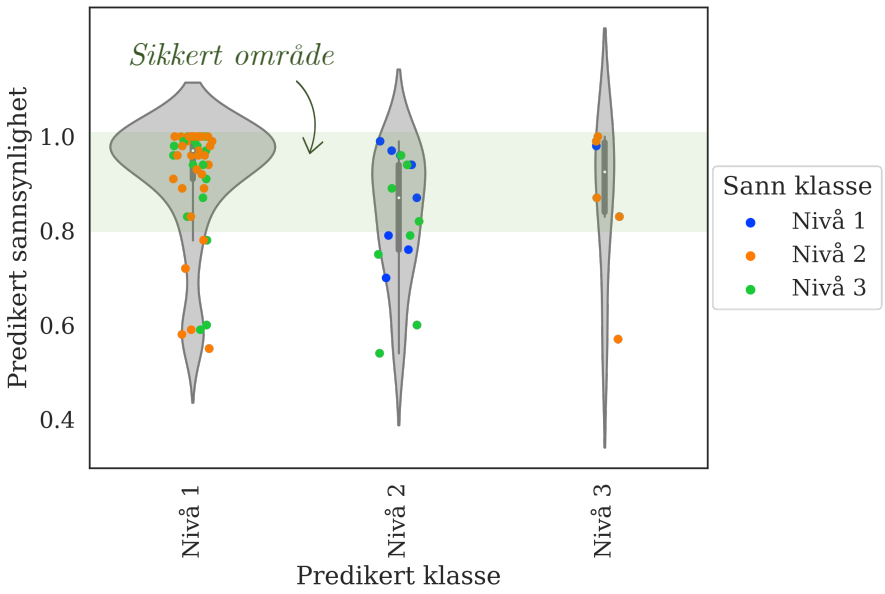
Figur 4.17: *Forvirringsmatrisa til referansemodellen ved klassifisering av nivå 1, 2 og 3. Referansemodellen hadde kompleksitet B0 og læringsrate 0.0005*

Den høyeste MCC-scoren oppnådd ved trening, validering og testing av modeller som skilte prøver med hensyn på nivå var 0.429, med tilhørende nøyaktighet på 0.631. Modellen som oppnådde denne ytelsen hadde kompleksitet B3, læringsrate 0.0005, og det ble brukt standard augmentering, og ble derfor kalt B3_0.0005_aug. Forvirringsmatrisa tilhørende denne modellen er presentert i figur 4.18. Matrisa viser at 90% av prøver fra nivå 1 ble korrekt klassifisert, mens 70% av prøver fra nivå 2 ble feilpredikert. Over halvparten av alle prøver ble predikert som nivå 1. Til tross for at andelene i figur 4.18 av feilpredikerte prøver fra nivå 2 og 3 har gått ned i forhold til figur 4.17, oppnådde modellen en høyere nøyaktighet og MCC-score enn referansemodellen.



Figur 4.18: *Forvirringsmatrisa til modellen med høyest ytelse ved klassifisering av nivå 1, 2 og 3, kalt B3_0.0005_aug.*

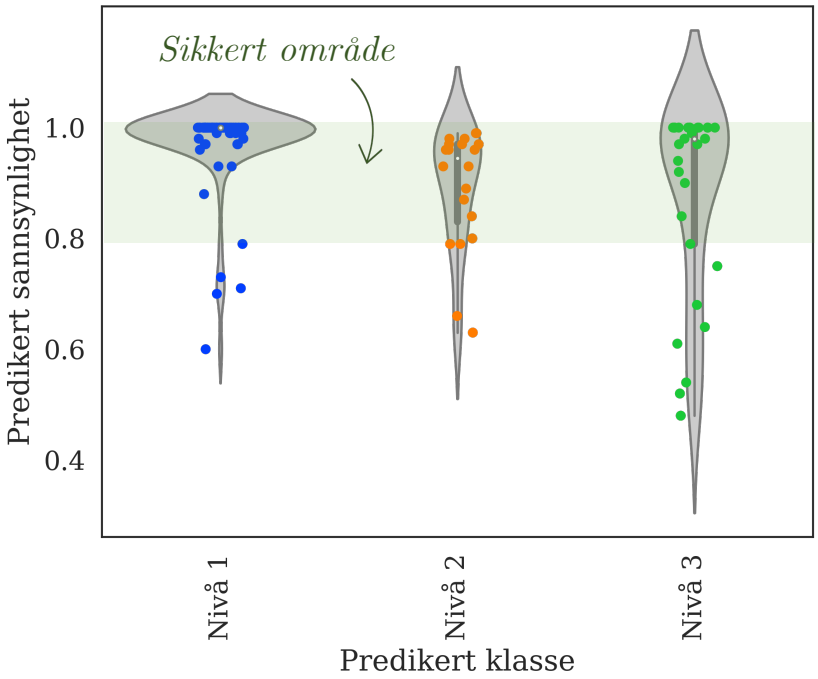
For flerklasseprediksjoner er laveste mulige sannsynlighet for klassetilhørighet 0.33, se kapittel 2.3.7. Figur 4.19 viser at spredningen av sannsynligheter spenner fra omtrent 0.6 til 1.0 for modell B3_0.0005_aug. For flerklasseproblem vil intervallet av usikre prediksjoner være gitt som 0.33-0.8 i denne masteroppgaven, siden ingen prøver predikeres med en sannsynlighet lavere enn 0.33, se kapittel 2.3.7.



Figur 4.19: *Fordeling av feilpredikerte prøver fra hvert nivå av modellen med høyest ytelse ved klassifisering av nivå 1, 2 og 3. Hver prikk på figuren tilsvarer en prøve, og fargen indikerer hvilken diagnosegruppe prøven faktisk tilhører. "Sikkert område", farget grønt, er intervallet av sannsynligheter som regnes som sikre prediksjoner (omtrent 0.8-1.0).*

Fiolinplottene i figur 4.20 av riktig predikerte prøver viser lav tetthet av usikre prediksjoner relativt til sikre prediksjoner, enn i figur 4.19. Det er allikevel høyere spredning i sannsynlig-

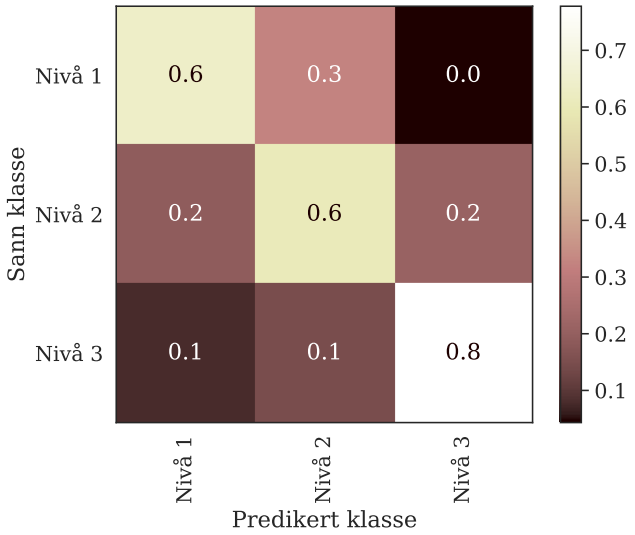
hetene enn for feilprediksjoner, med laveste predikerte sannsynlighet på omtrent 0.5 for riktig predikerte prøver fra nivå 3.



Figur 4.20: *Fordeling av riktig predikerte prøver fra hvert nivå av modellen med høyest ytelse ved klassifisering av nivå 1, 2 og 3. Hver prikk på figuren tilsvarer en prøve, og fargen indikerer hvilken diagnosegruppe prøven faktisk tilhører. "Sikkert område", farget grønt, er intervallet av sannsynligheter som regnes som sikre prediksjoner (omtrent 0.8-1.0).*

Omtrening ved klassifisering av nivå 1, 2 og 3

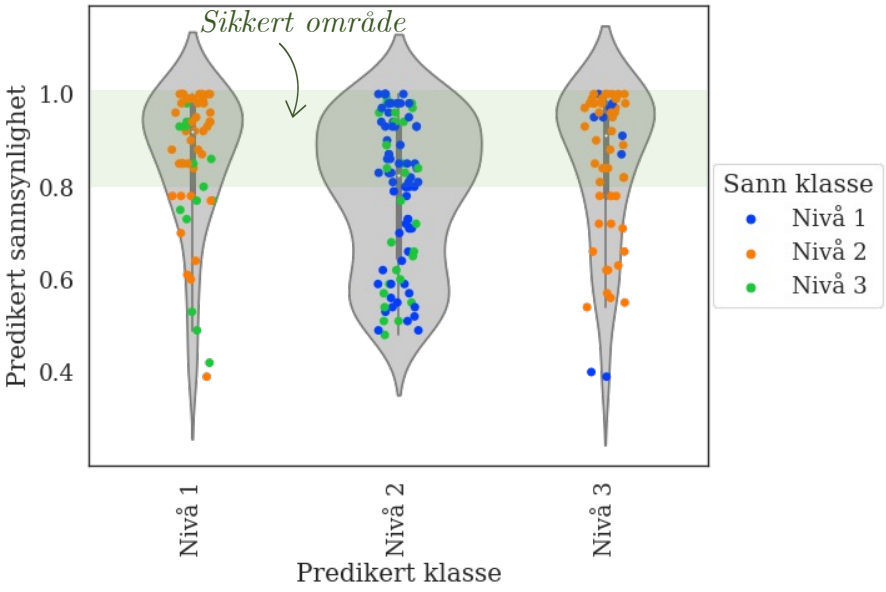
Modell B3_0.0005_aug, som hadde høyest ytelse ved klassifisering av nivå 1, 2 og 3, ble omtrent med 636 prøver fra det eksterne datasettet (se kapittel 3.5.2), og evaluert med resterende 636 prøver. Dette ble gjort for å se om det ville øke ytelsen fra resultatet gitt i figur 4.18. Resultatet er presentert i forvirringsmatrisa i figur 4.21. Ved denne omtreningen ble utvidet augmentering brukt, se tabell 3.7, for å se om dette bidro til høyere ytelse.



Figur 4.21: *Forvirringsmatrisa etter omtrening av modellen med høyest ytelse ved klassifisering etter nivå. Her har modellen trent på 636 nye prøver, som er en økning på omtrent 50% av antall prøver modell B3_0.0005_aug ble trent med i utgangspunktet.*

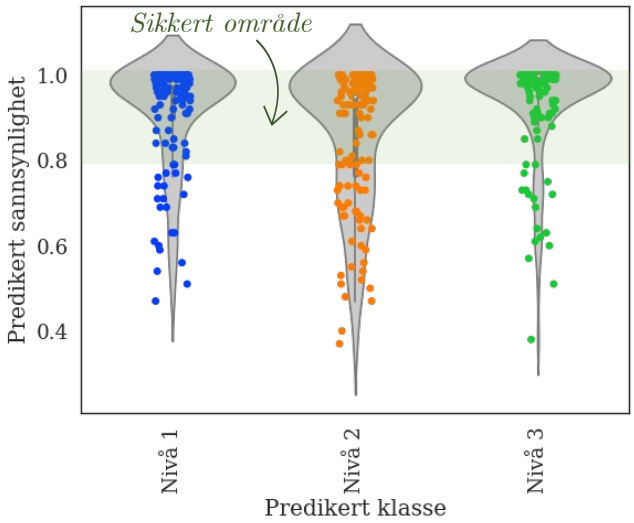
Forvirringsmatrisa i figur 4.21 viser at omtrening med flere prøver ga høyere ytelse enn etter første trening, se figur 4.18. Det er en økning av andel sanne positive for alle nivåer bortsett fra nivå 1, og alle nivåer har 60 % eller flere sanne positive i hver nivå-gruppe.

Figur 4.22 viser fordelingen av feilpredikerte prøver etter omtrening av modellen med høyest ytelse ved klassifisering av nivå 1, 2 og 3, B3_0.0005_aug. Fordelingen viser at spredningen i sannsynligheter er større enn før omtrening (figur 4.19), og de laveste sannsynlighetene ligger rundt 0.4. Dette tyder på at modellen feilpredikerte flere prøver med lav selvsikkerhet, men for alle nivåer er tettheten størst ved sikre sannsynligheter (over 0.8).



Figur 4.22: *Fordeling av feilpredikerte prøver fra hvert nivå etter omtrening av modellen med høyest ytelse ved klassifisering av nivå 1, 2 og 3. Hver prikk på figuren tilsvarer en prøve, og fargen indikerer hvilken diagnosegruppe prøven faktisk tilhører. Laveste predikerte sannsynlighet ligger rundt 0.4, som regnes som en usikker prediksjon. "Sikkert område", farget grønt, er intervallet av sannsynligheter som regnes som sikre prediksjoner (omtrent 0.8-1.0).*

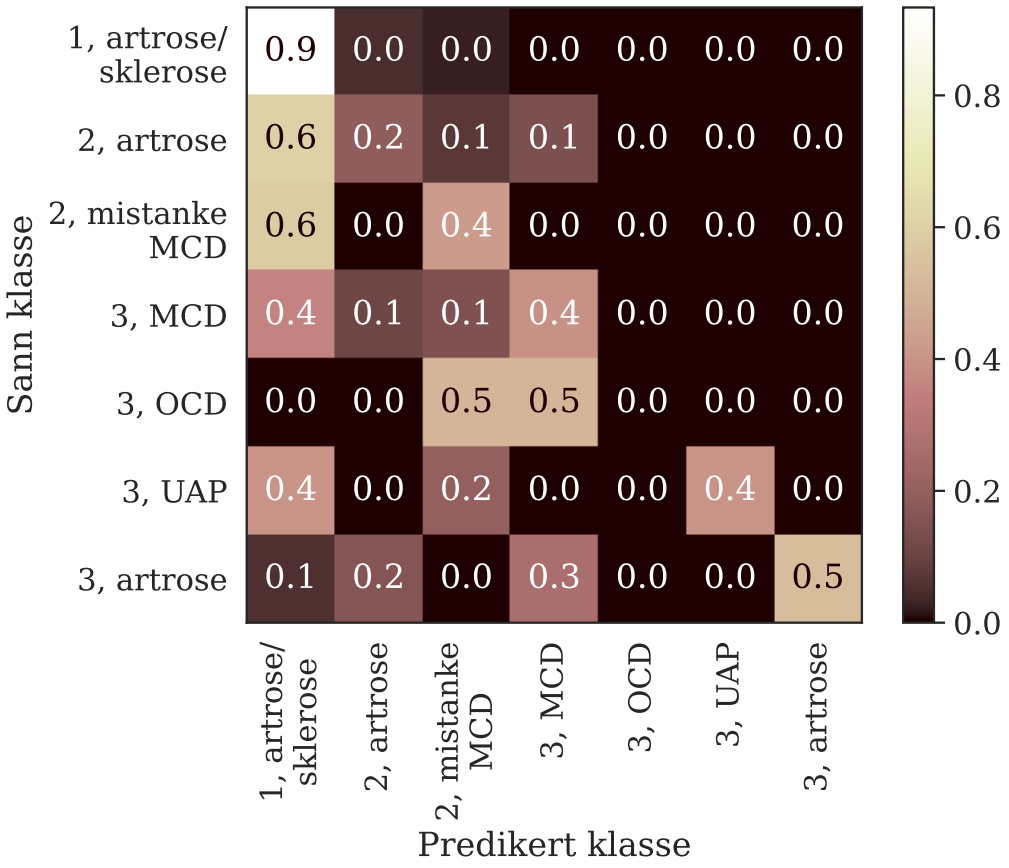
Fordelingen for riktig predikerte prøver er gitt i figur 4.23, og her er alle tettheter samlet mer entydig på intervallet for sikre prediksjoner (over 0.8) relativt til figur 4.22. Det ble allikevel også her stor spredning i sannsynligheter, med sannsynligheter fra omtrent 0.4 til 1.0.



Figur 4.23: *Fordeling av riktig predikerte prøver fra hvert nivå etter omtrening av modellen med høyest ytelse ved klassifisering av nivå 1, 2 og 3. Hver prikk på figuren tilsvarer en prøve, og fargen indikerer hvilken diagnosegruppe prøven faktisk tilhører. "Sikkert område", farget grønt, er intervallet av sannsynligheter som regnes som sikre prediksjoner (omtrent 0.8-1.0).*

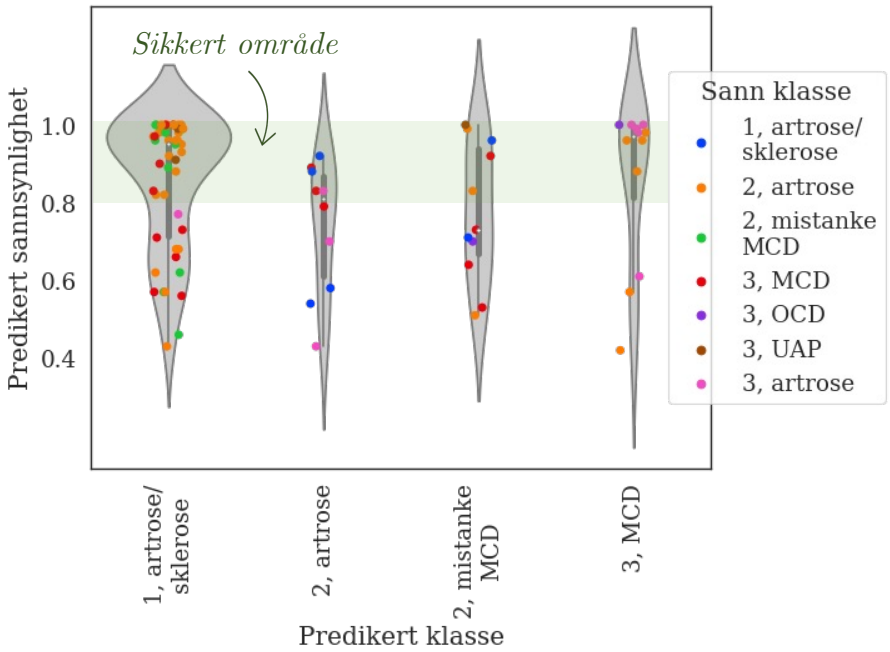
4.3.2 Klassifisering av alle diagnoser

Også flerklasseproblem med klassifisering av alle diagnosegrupper ble utprøvd, altså å skille alle diagnoser fra nivå 1 artrose og/eller sklerose til nivå 3 UAP, se kapittel 3.5.2. Resultatet til modellen med høyest ytelse av fire eksperimenter på denne problemstillingen er presentert som forvirringmatrise i figur 4.24. Denne modellen ble kalt B4_0.0005, og hadde kompleksitet B4 og læringsrate 0.0005. MCC-scoren oppnådd med modell B4_0.0005 var 0.422 med en nøyaktighet på 0.597 (se tabell 4.1), og prøver med nivå 1 artrose og/eller sklerose hadde høyest andel sanne prediksjoner. Ingen av prøvene med nivå 3 OCD ble riktig predikert, og alle ble predikert som MCD i nivå 2 og 3.



Figur 4.24: *Forvirringsmatrisa til modell B4_0.0005, som hadde høyest helhetlig ytelse ved klassifisering av alle diagnosegrupper gitt i tabell 3.2, uten normale prøver.*

Figur 4.25 viser fordelingen av feilpredikerte prøver, predikert å tilhøre fire ulike diagnosegrupper. Prøver fra alle diagnosegrupper ble feilpredikert som noe annet, men som vist i både figur 4.24 og 4.25, var det ingen prøver som ble klassifisert som diagnosegruppe 3, OCD.



Figur 4.25: *Fordeling av feilpredikerte prøver fra hver diagnose ved klassifisering av alle diagnosegrupper med modell B4_0.0005. Hver prikk på figuren tilsvarer en prøve, og fargen indikerer hvilken diagnosegruppe prøven faktisk tilhører. "Sikkert område", farget grønt, er intervallet av sannsynligheter som regnes som sikre prediksjoner (omtrent 0.8-1.0).*

I figur 4.25 kommer det fram at fra diagnosegruppe 2, mistanke MCD, ble alle feilprediksjoner klassifisert som 1, artrose/sklerose. Prøver fra diagnosegruppe 3, UAP ble feilpredikert som 1, artrose/sklerose og 2, mistanke MCD, som antyder at modellen ikke klarte å finne tydelige kjennetegn for diagnosegruppe 3, UAP. Prøver fra diagnosegruppe 1, artrose/sklerose ble feilpredikert som 2, artrose og 2, mistanke MCD.

Kapittel 5

Diskusjon

5.1 Modellytelse

I denne masteroppgaven ble et konvolusjonelt nevralt nettverk (CNN) trent og evaluert på røntgenbilder av hundealbuer med og uten albueleddsdysplasi (AD). Det ble oppnådd en CNN-modell med nøyaktighet på over 0.95, og MCC-score på 0.91, som viser høy grad av klassifiserbarhet mellom normale og abnormale prøver. Datasettet brukt til trening, validering og testing av modellen som skilte abnormale fra normale prøver besto av omtrent halvparten normale og halvparten abnormale prøver, men inneholdt variasjoner innad i abnormal gruppe når det gjaldt antall prøver fra hver diagnosegruppe som vist i tabell 3.3.

Andre lignende studier, som for eksempel studier som bruker dyp læring til binær klassifisering av abnormaliteter, i kneledd eller hofteledd hos hunder, rapporterer gjenkallingsverdi på hhv. 0.53 (McEvoy et al.) og opp til 0.90 (Shim et al.) [15, 50]. Presisjonsverdiene til samme modeller var hhv. 0.81 og 0.96, som gir F1-verdier på hhv. 0.64 og 0.93. I denne masteroppgaven oppnådde modellen med høyest helhetlig ytelse en F1-score på 0.96, som er høyere enn begge nevnte lignende studier.

Også blant studier på humanmedisin er det forsket på bruk av dyp læring på røntgenbilder, som for eksempel ved abnormaliteter som følge av traumer på albuer. To slike studier gjort på albuer ble gjennomført av Rayan et al. og Huhtanen et al. [19, 18], og disse studiene rapporterte gjenkallingsverdier på hhv. 0.91 og 0.89. Rayan et al. brukte tre bilder av barnealbuer om gangen for hver prediksjon, mens Huhtanen et al. fikk høyest ytelse på ett og ett bilde av albuer (for blanding av voksne og barn). F1-scorene tilhørende modellene med nevnte gjenkallingsverdier var hhv. 0.89 og 0.88, som begge var lavere enn F1-scoren til normal/abnormal-modellen med høyest ytelse i denne masteroppgaven.

Et studie av Regnard et al. [17] ble gjort på ytelsen til dyp læring brukt på traumerøntgenbilder av menneskelige lemmer og hofter for binær klassifisering av fire ulike abnormaliteter i knokler (fraksjon, dislokasjon, albueeffusjon og beinlesjon). For hver av disse fire abnormalitetene ble gjenkallingsverdier på hhv. 0.98, 0.90, 0.92 og 0.98 oppnådd, med tilhørende F1-verdier på hhv. 0.76, 0.73, 0.88 og 0.74 [17]. Selv om gjenkallingsverdiene tyder på høy grad av deteksjon av abnormale prøver, vitner F1-verdiene om lavere presisjon, altså mange falske positive. Ved diagnostisering kan det være viktigere å ha høy gjenkalling enn presisjon, slik at abnormaliteter fanges opp og kan behandles, selv om dette fører til flere falske positive prøver.

I denne masteroppgaven oppnådde modellen med høyest helhetlig ytelse ved normal/abnormal

klassifisering like F1-verdier for både F1 med hensyn på positiv og negativ klasse. Dette viser høy grad av separabilitet, som betyr at en stor andel abnormaliteter ble detektert, uten at det gikk på bekostning av antallet falske positive prøver. Dette er bemerkelsesverdig ved sammenligning med lignende studier nevnt over.

Når det gjelder ytelsen til veterinærer ved diagnostisering av røntgenbilder, ble det ikke gjort en analyse på dette i denne masteroppgaven. Studien av Shim et al. [50] gjort på binær klassifisering av abnormaliteter i kneledd på hunder registrerte derimot ytelsen til radiologer. De rapporterte nøyaktigheter på mellom 0.87 og 0.91 for ulike abnormaliteter i kneleddet. For eksempel gjaldt den ene abnormaliteten osteofyttformasjon, som i likhet med artrose er utvekster av benvev. Til tross for denne likheten kan det ikke gås ut ifra at samme nøyaktigheter vil gjelde for diagnostisering av AD, men det gås ut ifra at radiologer heller ikke ved diagnostisering av AD oppnår en nøyaktighet på 1.0.

Tidsbruken ved diagnostisering av AD fra røntgenbilder var på omtrent 1 sekund per bilde for CNNer i denne masteroppgaven. Dette er en markant forskjell fra de estimerte 5 minuttene radiologer bruker på samme arbeid. Spesielt når det gjelder å skille ut alle abnormale prøver fra normale prøver kan det spares mye tid for radiologene å ta i bruk maskinlæring. Som vist i figur 3.2a er omtrent 90 % av alle hunder som gjennomgår screening sykdomsfrie, og det vil derfor være et effektivt tiltak at radiologene kun utfører gradering av albuer der det er kjent at abnormalitet er til stede.

5.2 Begrensninger

Denne masteroppgaven har i utgangspunktet prøvd ut vilkårlige konfigurasjoner, med mål om å oppnå høyest mulig ytelse på EfficientNet-modellene. Det ble ikke tatt hensyn til parvise kombinasjoner for statistisk testing av signifikansen til ulike konfigurasjoner med hensyn på ytelse. En av grunnene til dette var begrensningen i ressurser ved kjøring av eksperimenter, da det tidvis var lange køer og andre problemer ved kjøring av slurm jobs på Orion. Selv om forskjellen i konfigurasjoner ikke kan dokumenteres statistisk, viser resultatene at visse konfigurasjoner ga høyere ytelse enn andre.

Et viktig punkt å ta hensyn til når det gjelder å skille mellom abnormale klasser er at flere diagnoser opptrer samtidig. Radiologer vil alltid kategorisere prøver som den "verste" (høyeste diagnosegrad) sykdommen til stede, og dette ligger til grunn i kategorisering av grunnsannhetene i denne masteroppgaven. For eksempel for en prøve med artrose nivå 2 kombinert med MCD nivå 3 vil prøven alltid klassifiseres som MCD nivå 3. Dette kan føre til at en maskinlæringsmodell klassifiserer prøven som artrose nivå 2 heller enn MCD. Et tiltak for å øke korrekte prediksjoner ved kombinasjoner av diagnoser kan være å prøve flerklasseproblem med flere målklasser (multi-label[1]). Det kan også være en idé å analysere for eksempel de tre høyeste sannsynlighetene (top-3 nøyaktighet) for hver prøve, siden den verste sykdommen kan ha bare en liten forskjell i sannsynlighet fra den høyeste sannsynligheten.

For å analysere feilprediksjoner, blant annet i lys av flere diagnoser per prøve, kan explainability[2] benyttes [53]. Dette vil gi en indikator på om det hovedsakelig er modellen som ikke finner sykdomstegnene, eller om det er prøvene som har dårlig predikerbarhet. Explainability vil også "åpne den svarte boksen", som maskinlæringsmodeller ofte kan minne om, siden informasjons-

[1] https://scikit-learn.org/stable/modules/multiclass.html

[2] Markere piksler modellen har fokusert på på bildet.

uthentingen til modellen ikke er kjent. Dersom en modell viser høy ytelse kan explainability bidra til å hjelpe radiologer til å se tegn de ikke ville sett uten, som vil være en god bruk av maskinlæring innen radiologi, heller enn å erstatte radiologer.

5.3 Selvsikkerhet ved prediksjon

Fra fordelinger av feilpredikerte prøver i alle problemtyper er det tydelig at modellene feilpredikerer med sterk "selvsikkerhet" (overconfidence). Fiolinplottene viser tettheten på sannsynlighetene for at prøvene tilhører abnormal klasse, og tettheten er klart høyest i ytterpunktene (sannsynligheter ved 0.0 og 1.0). Dette gjelder for både rett og feil predikerte prøver, men spredningen er større for feilpredikerte prøver.

For prøver som ikke ligger i ytterpunktene, altså for prøver med sannsynligheter for eksempel over 0.2 og under 0.8, finnes flere alternativer for å redusere antall feilprediksjoner. Det ene alternativet er å flagge disse usikre prøvene, slik at alle flaggede, i tillegg til abnormale, prøver blir vurdert manuelt av radiologer. Dette vil medføre ekstra arbeid for radiologene, som må gå gjennom flere prøver dersom "flaggingsområdet" settes til et intervall på mellom 0.2 og 0.8. Selv om tettheten til feilpredikerte er større enn rett predikerte ved "usikre" sannsynligheter, er det et høyere antall rett predikerte i absolutte tall, og dermed er det sannsynlig at radiologen må gå gjennom flere prøver som faktisk var rett predikert enn feilpredikert. Dette blir et kost/nytte-regnskap, men spesielt ved klinisk diagnostisering er dette et nyttig grep for å unngå å gi pasienter feil diagnose.

Alternativ to er å heve eller senke terskelverdien på sannsynligheten for at en prøve havner i abnormal klasse, slik at for eksempel alle prøver med sannsynlighet over 0.7 ble klassifisert som abnormal. En heving vil medføre at man ikke fanger opp abnormale prøver (flere falske negative), mens senking medfører flere falske positive prøver. I en klinisk sammenheng vil senking av terskelverdien fungert bra for å fange opp hunder med albueleddsdysplasi, selv om dette ville medført flere falske positive. Hovedproblemet med falske positive prøver er, som nevnt i kapittel 1.1, at hunder som kunne vært avlet på ikke får brukes mer i avl, noe som kan føre til økonomiske tap for hundeeiere.

5.4 Omtrening av EfficientNet-modell

Å kjøre omtrening av normal/abnormal-modellen med bare feilklassifiserte prøver (figur 4.12a) ga dårlige resultater på resten av det eksterne datasettet. Dette kan skje dersom feilklassifiserte prøver fra ekstern evaluering var veldig forskjellige fra prøvene modellen hadde trent på under første trening (outliers). Alternativt kan modellen ha blitt overtilpasset prøvene gitt i omtreningen, og dermed ha mistet viktige koblinger som gjorde at modellen generaliserte godt opprinnelig. Fordelingen av diagnosegrupper til stede blant prøvene brukt i denne omtreningen ble vist i figur 4.11 i kapittel 3.5.1. Denne fordelingen var ulik fordelingen til de andre datasettene brukt ved trening av normal/abnormal modell, og dette kan ha medført at modellen "glemte" hvordan den skulle klassifisere for eksempel 3, OCD- og 3, UAP-prøver, som ikke var representert i det hele tatt i datasettet.

Etter omtrening med halve det eksterne datasettet (figur 4.12b) økte ytelsen betraktelig fra 0.837 (ekstern evaluering) til 0.912 (evaluering etter omtrening) i MCC-score. Dette kan komme av at det kun var halve det eksterne datasettet som ble brukt til evaluering etter omtrening, altså at færre antatt "ikke klassifiserbare" prøver måtte predikeres av modellen. Det er ikke

kjent hvor mange av de feilpredikerte prøvene fra den eksterne evalueringen som ble brukt ved omtrening og evaluering av omtrent modell, og det kan tenkes at de fleste av prøvene havnet i treningsdataen, og at det på den måten var færre "vanskelige" prøver i evalueringen av den omtrente modellen. Allikevel er det verdt å nevne at det var prøver fra alle diagnosegrupper til stede i både trenings- og evalueringsettet ved omtrening, siden noen diagnosegrupper så ut til å bli feilpredikert oftere enn andre generelt i denne masteroppgaven, se for eksempel figur 4.13.

Forvirringsmatrisa til denne omtrente modellen (figur 4.12b) viser nedgang i begge feilklassifiserte grupper, og spesielt falske positive har en veldig lav andel. Dette er viktig i forbindelse med screening, siden det kan medføre økonomiske tap dersom hunden blir merket som syk når den ikke er det, altså falsk positivt utfall. I en screeningsammenheng er det derfor viktigere å oppnå høy ytelse på F1-verdier av negativ klasse enn av positiv klasse.

Omtrening av modellen med høyest ytelse på klassifisering av nivå 1, 2 og 3 ga også høyere ytelse. Også denne modellen ble trent videre på flere prøver enn ved første trening, og resultatene av omtrening med flere prøver indikerer generelt at modeller yter bedre dersom de er trent på et større datasett. Ved trening av modeller i denne masteroppgaven kunne det vært hensiktsmessig å trene alle modeller med mange flere prøver, men det ville ført til et mindre eksternt datasett for hver problemstilling. Et stort eksternt datasett er nyttig for å sikre variasjon i bildene modellen aldri har sett før. Når det er stor variasjon i bildene ved ekstern evaluering kan evalueringen regnes som en god representasjon på hvordan modellen ville ytt i praksis.

5.5 Evaluering av andre modeller

Modellene med høyest ytelse ved flerklasseproblem i denne masteroppgaven oppnådde ikke like høy ytelse som den binære normal/abnormal-modellene. Klassifisering av nivå 1, 2 og 3 fikk høyest ytelser, med nøyaktighet på 0.67 og MCC på 0.50, mot klassifisering av alle diagnosegrupper på hhv. 0.597 og 0.422. Som nevnt tidligere, kan for eksempel top-3 nøyaktighet benyttes ved evaluering, slik at modellen ikke får "feil", dersom modellen velger sykdommen med lavest gradering. Ved flerklasseproblem med tre målklasser vil ikke dette være relevant, da modellen alltid må treffe én av tre målklasser, men ved klassifisering av alle diagnoser kan dette være nyttig. Da kan sannsynlighetsdistribusjonen til hver prøve brukes som et verktøy av veterinærer for å finne riktig diagnose under analyse av røntgenbilder.

Forvirringsmatrisa til modellen med høyest ytelse for klassifisering av alle diagnosegrupper (figur 4.24) viser at spesielt diagnosegruppa 3, OCD var vanskeligere enn andre å klassifisere. Ingen prøver ble predikert som 3, OCD, som tyder på at modellen kanskje ikke fant noen kjennetegn i det hele tatt for denne gruppa. Også diagnosegruppa 2, artrose viste dårlig separabilitet fra andre diagnosegrupper, og en stor andel (60 %) prøver med grunnsannhet 2, artrose ble diagnostisert som 1, artrose/sklerose. Dette kan komme av de små ulikhetene i disse to gruppene, der det kun handler om millimeterforskjeller på benavleiringsklumper ved artrose.

Det ble også prøvd å bruke maskinlæring til å skille mellom prøver som forventes å bli feilklassifisert i modellen som skiller abnormale fra normale prøver (*utvidet analyse*, figur 4.15). Denne modellen fungerte dårlig på negativ klasse (klassifiserbar), da modellen predikerte under 50 % sanne negative. Modellen plukket derimot ut prøver som ikke var klassifiserbare i den binære normal/abnormal-modellen nesten 70 % av gangene (sanne positive), og det ville vært interessant å utforske denne modellen mer. Altså vil nesten 70 % av alle prøver som ville blitt feilpredikert i normal/abnormal-modellen fanges opp og analyseres manuelt i stedet. Dersom denne modellen ble brukt uten videre justering, ville potensielt 60 % av alle prø-

ver sendes direkte til manuell diagnostisering, hvorav nesten 60 % ville blitt feilpredikert av normal/abnormal-modellen.

Selv om en større andel prøver ikke ville blitt feilpredikert med denne metoden ville det gått på bekostning av tiden spart ved å ikke gå gjennom prøver manuelt i det hele tatt. Dersom 60 % av alle prøver måtte analyseres manuelt hvert år, ville i snitt 3000 prøver måtte gjennomgå manuell analyse direkte, mens de resterende 2000 ville blitt automatisk klassifisert med normal/abnormal-modellen.

Den største begrensningen ved opptrening av en modell som skal skille klassifiserbare og ikke klassifiserbare prøver er tilgangen på feilpredikerte prøver fra tidligere modeller. I dette forsøket ble 205 prøver brukt i treningssettet, som er omtrent halvparten av antallet brukt til trening av normal/abnormal-modellen. Siden det er ukjent hvilke egenskaper de feilpredikerte bildene i normal/abnormal-modellen hadde, er det grunn til å tro at eventuelle fellesnevnere i egenskapene er komplekse, og krever mer eksperimentering for å kunne oppdages.

5.6 Bruk av flere bilder per prøve

Radiologer bruker aktivt røntgenbilder av begge albueleddene for å stille diagnose [11]. Det er en fordel å ha begge sider tilgjengelig for vurdering samtidig, slik at subtile forskjeller mellom albuene blir enklere å legge merke til. Alle modellene i denne oppgaven har bare tatt inn ett og ett bilde, og har ikke hatt tilgang til annen informasjon om hunden eller kunnet sammenligne høyre og venstre albueledd. Spesielt for flerklassemodeller kan det være nyttig å se på potensialet til en slik modell, siden modellene oppnådd i denne masteren ikke hadde utpreget høy ytelse.

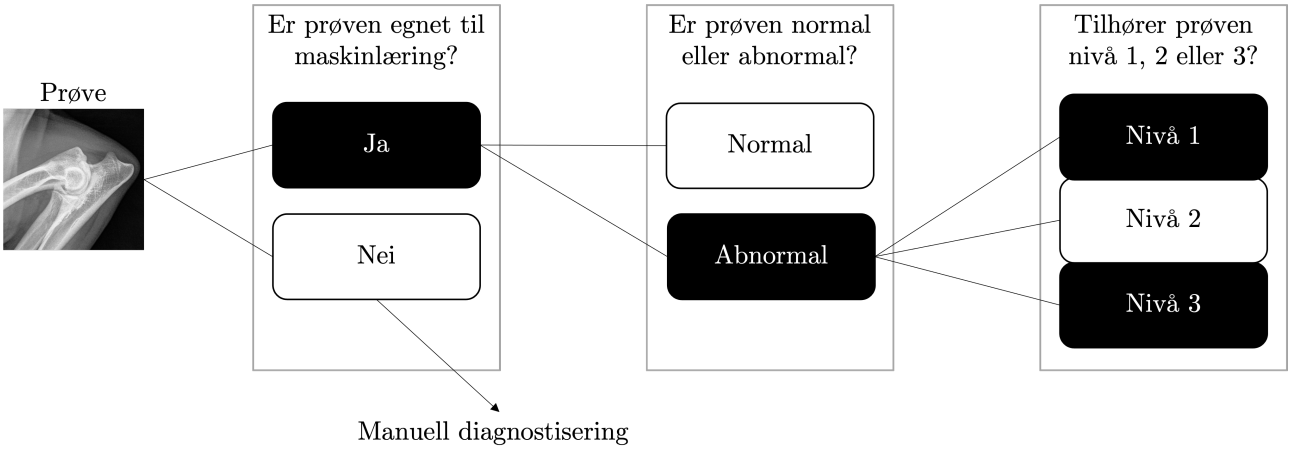
5.7 Preprosessering av røntgenbilder

I tillegg til preprosessering av originalbildene, har veterinærer mulighet til å manipulere DICOM-filene rett etter røntgenbildene er tatt. Dersom det ble undersøkt nærmere hva individuelle kontrastforskjeller og bildekvaliteter har å si for klassifisering, kunne det vært aktuelt å utarbeide en slags standardisering av DICOM-filer når bildene tas. På denne måten kan bildene bli likere, og modellen vil ikke bli påvirket av slike forskjeller. Ved analyse av alle feilpredikerte bilder fra testsettet til modell B4_0.0005_aug (se figur 4.5), kom det fram at omtrent 1/3 av de feilpredikerte bildene var ansett kontrastfylte, altså at benvev ikke har jevn hvitfarge, men mørke flekker der det burde være hvitt. Et eksempel på et kontrastfylt bilde feilpredikert som abnormal med en sannsynlighet på 0.59 er vist i figur 5.1. Dette kan antyde at kontraster i bildene virker inn på prediksjonene, og at mer utforsking av kontraster i forbindelser med feilprediksjoner kan gi bedre resultat. Både nærmere analyse av feil-/ og riktig predikerte prøver fra ekstern evaluering, og større variasjon i augmenteringsteknikkene (se tabell 3.7) brukt ved trening av modellen, er eksempler på tiltak som kan tenkes å bidra til en reduksjon av feilprediksjoner som følge av kontrastforskjeller.

Figur 5.1: *Eksempel på kontrastfylt bilde som ble feilpredikert av modellen med høyest helhetlig ytelse for klassifisering av normale og abnormale prøver. Denne prøven hadde sann diagnose normal, men ble predikert som abnormal med en sannsynlighet på 0.59.*

5.8 Diagnostisk pipeline

Som en sammenfatning av resultatene i denne masteroppgaven kan en diagnostisk pipeline foreslås for å understreke potensialet til maskinlæring som hjelpemiddel ved diagnostisering. Pipelinen er skissert i figur 5.2, og viser tre steg en pipeline kan inneholde: klassifiserbarhet av prøven, abnormalitet og gradering av abnormalitet. Steg én er dermed en utluking av prøver som ikke vil la seg klassifisere i neste steg av pipelinen, og disse prøvene blir sendt rett til manuell diagnostisering av en veterinær. Neste steg plukker ut alle abnormale prøver, som i siste steg automatisk graderes fra nivå 1 til 3.



Figur 5.2: *Forslag på komplett pipeline til screening av AD.*

Dette oppsettet kan brukes som en støtte under manuell diagnostisering, altså at veterinæren får

et forslag på diagnose før han eller hun gjør sin egen vurdering. I framtiden kan modellen forskes videre på og utvikles til å brukes i stedet for manuell gjennomgang av bilder, i første omgang ved screening. Modellene er ikke klare for å erstatte veterinærer, spesielt i klinisk sammenheng, men i screeningsammenheng gjøres forebyggende (i motsetning til akutt) behandling, og det kan derfor være mer aktuelt å vurdere større menneskelig uavhengighet i slike tilfeller.

Kapittel 6

Konklusjon og videre arbeid

6.1 Konklusjon

Denne masteroppgaven hadde som mål å undersøke potensialet for bruk av konvolusjonelle nevrale nettverk for binær klassifisering av albueleddsdysplasi fra røntgenbilder av hundealbuer. På grunn av høye ytelsesverdier på store bildedatasett ble modeller i EfficientNet-familien brukt til klassifisering. Eksperimenter gjort i denne masteroppgaven viser at konvolusjonelle nevrale nettverk i høy grad kan skille mellom normale og abnormale hundealbuer.

Røntgenbilder som ble feilpredikert viser ikke åpenbare tegn til fellestrekk, hverken ved menneskelig evaluering eller ved forsøk på å skille prøvene ved hjelp av en EfficientNet-modell, se figur 4.6 og vedlegg B. Det ble allikevel lagt merke til likheter i kontraster for noen av bildene.

Andre problemstillinger som klassifisering av grad av albueleddsdysplasi ble også utprøvd med EfficientNet-modeller, men disse viste ikke like høy grad av predikerbarhet. Heller ikke klassifisering av albuer med artrose vs. øvrige diagnoser, eller andre utprøvde problemstillinger oppnådde like høy ytelse. Klassifisering av spesifikk diagnose og/eller diagnosegrad viser seg å være mindre predikerbart enn prediksjon av tilstedeværelse av albueleddsdysplasi, som kan følge av at flere diagnoser og/eller diagnosegrader kan være til stede i samme albue.

6.2 Videre arbeid

Selv om arbeidet i denne masteroppgaven indikerer stort potensiale for klassifisering av albueleddsdysplasi i hundealbuer, bør metoden evalueres og testes videre før den kan implementeres i en screeningsammenheng.

En analyse av radiologers treffsikkerhet ved diagnostisering av AD kan være et nyttig verktøy i diskusjonen om bruk av CNNer i dette arbeidet. Dette bør prioriteres ved videre arbeid for å kunne sammenligne modeller oppnådd, som et mål på relevansen til CNNene.

En mer omfattende analyse av feilprediksjoner vil være et viktig steg for å forstå i hvilke tilfeller prediksjonene svikter. Blant annet er kontrastfylte bilder nevnt i denne masteroppgaven som en mulig grunn til feilprediksjoner, og dette kan betyr at kontrastfylte bilder er outlier-bilder. Andre ukjente outliere kan også synliggjøres ved mer utforskning.

Et annet punkt ved analyse av feilprediksjoner er å undersøke om predikert diagnosegruppe ved feilprediksjoner egentlig var til stede i prøven eller ikke. Altså å gjennomgå alle feilpredikerte

bilder og markere alle diagnosegrupper som er til stede på hvert bilde. Et nyttig steg i videre arbeid med feilprediksjoner er også analyse av explainability, som vil gi indikasjoner på hvilke piksler modellen fokuserer på ved prediksjon [53]. Dette kan også sammenlignes med hvilke områder av røntgenbildene veterinærene fokuserer på ved screening, for å se om veterinærenes og CNNenes fokusområder sammenfaller.

Omtrening av modeller med et større treningssett har vist gode resultater på økning i ytelse, og dette kan brukes for å utvikle modeller med enda høyere treffsikkerhet. For flerklassemodeller kan også implementering av flere bilder per prøve være en metode for å øke ytelsen til modellen. I tillegg kan det implementeres kode som gjør at modellen alltid velger høyeste diagnosegrad dersom det er liten forskjell i sannsynligheter mellom to diagnosegrader, for å overkomme utfordringen med at flere diagnoser opptrer samtidig. Et forslag til videre arbeid er derfor å kombinere røntgenbilder fra høyre og venstre albue for å se om dette vil ha en effekt på resultatene ved ulike problemstillinger, men spesielt flerklasse-problemstillinger.

Dersom flere modeller blir utprøvd, med flere konfigurasjoner, kan med fordel en statistisk analyse av betydningen av konfigurasjoner gjøres. Dette vil gi konklusjonen mer tyngde, og modellen kan implementeres i screeningarbeid med større sikkerhet. Dersom det er tilgang på flere dataressurser kan også mer komplekse modeller utprøves, da eksperimenter i denne oppgaven generelt viser høyere ytelse for modeller med høyere kompleksitet.

Referanser

[1] Spencer A. Johnston og Karen M. Tobias. *Veterinary surgery : small animal*. 2. utg. Bd. 1. St. Louis, Missouri: Elsevier, 2018. ISBN: 0323320651.

[2] François Chollet. *Deep Learning with Python*. 2017.

[3] Lovdata. *Dyrevelferdsloven (LOV-2009-06-19-97)*. Jul. 2021. URL: https://lovdata.no/lov/2009-06-19-97/%C2%A725.

[4] Norsk Kennel Klub. *Registreringstjenester hos Norsk Kennel Klub*. URL: https://www.nkk.no/for-hundeeiere/registrering/.

[5] Norsk Kennel Klub. *2022 - Registreringstall pr. fylke og kommune*. Jan. 2023. URL: https://www.nkk.no/statistikk/category1098.html.

[6] Guro Hatlo og Nils Fridtjof Skumsvoll. *Angrer på at de skaffet seg hund under pandemien – NRK Vestfold og Telemark – Lokale nyheter, TV og radio*. Sep. 2021. URL: https://www.nrk.no/vestfoldogtelemark/angrer-pa-at-de-skaffet-seg-hund-under-pandemien-1.15670383.

[7] Aftenposten. *Over 2 millioner søk etter hund i koronaåret*. Des. 2020. URL: https://www.aftenposten.no/okonomi/i/0K94r2/over-2-millioner-soek-etter-hund-i-koronaaaret.

[8] Aud Darrud mfl. *Hundeoppdrettar klaga inn etter at sju kvalpar frå ulike kull fekk same liding – NRK Norge – Oversikt over nyheter fra ulike deler av landet*. Mar. 2023. URL: https://www.nrk.no/norge/hundeoppdrettar-klaga-inn-etter-at-sju-kvalpar-fra-ulike-kull-fekk-same-liding-1.16206778.

[9] Mari Nyborg Hauback. *Muntlig kommunikasjon med Mari Nyborg Hauback ved NMBU*. Ås, feb. 2023.

[10] Norsk Kennel Klub. *AD-albueleddsdysplasi - Norsk Kennel Klub*. URL: https://www.nkk.no/ad-albueleddsdysplasi/category1014.html.

[11] Hege Kippenes Skogmo. *Muntlig kommunikasjon*. Ås, mar. 2023.

[12] Stuart W.S. MacDonald, Lars Nyberg og Lars Bäckman. «Intra-individual variability in behavior: links to brain structure, neurotransmission and neuronal activity». I: *Trends in Neurosciences* 29.8 (aug. 2006), s. 474–480.

[13] Geoff Currie og Eric Rohren. «Intelligent imaging: Applications of machine learning and deep learning in radiology». I: *Veterinary Radiology & Ultrasound* 63.S1 (des. 2022), s. 880–888.

[14] Erin Hennessey mfl. «Artificial intelligence in veterinary diagnostic imaging: A literature review». I: *Veterinary Radiology and Ultrasound* 63.S1 (des. 2022), s. 851–870.

[15] Fintan J. McEvoy mfl. «Deep transfer learning can be used for the detection of hip joints in pelvis radiographs and the classification of their hip dysplasia status». I: *Veterinary Radiology & Ultrasound* 62.4 (jul. 2021), s. 387–393.

[16] Norsk Kennel Klub. *Hofteleddsdysplasi (HD)*. URL: https://www.nkk.no/hd-hofteleddsdysplasi/category1281.html.

[17] Nor Eddine Regnard mfl. «Assessment of performances of a deep learning algorithm for the detection of limbs and pelvic fractures, dislocations, focal bone lesions, and elbow effusions on trauma X-rays». I: *European Journal of Radiology* 154 (sep. 2022).
[18] Jarno T. Huhtanen mfl. «Deep learning accurately classifies elbow joint effusion in adult and pediatric radiographs». I: *Scientific Reports* 12.11803 (des. 2022).
[19] Jesse C. Rayan mfl. «Binomial Classification of Pediatric Elbow Fractures Using a Deep Learning Multiview Approach Emulating Radiologist Decision Making». I: *Radiology. Artificial intelligence* 1.e180015 (jan. 2019).
[20] Mingxing Tan og Quoc V. Le. «EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks». I: *36th International Conference on Machine Learning, ICML 2019* 2019-June (mai 2019), s. 10691–10700.
[21] M Flower. *Webb's Physics of Medical Imaging*. Red. av M Flower. 2. utg. Bova Raton: CRC Press, des. 2012. ISBN: 9780429099571.
[22] Hiroaki Hayashi mfl. *Photon Counting Detectors for X-ray Imaging*. 1. utg. Springer International Publishing, 2021.
[23] Wade Allison. *Fundamental Physics for Probing and Imaging*. 1. utg. Oxford University Press, Incorporated, 2006. ISBN: 0191525332.
[24] Shultis J. Kenneth og Faw Richard E. *Fundamentals of Nuclear Science and Engineering*. 3. utg. Productivity Press, 2016.
[25] Edward L Alpen. *Radiation biophysics*. eng. 2nd ed. San Diego, Calif.: Academic Press, 1998, s. 74–76. ISBN: 0-12-053085-6.
[26] J S Lilley. *Nuclear physics : principles and applications*. eng. Chichester, 2001.
[27] Norsk Kennel Klub. *Avl*. URL: https://www.nkk.no/nkk-mener/category1051.html%20chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.nkk.no/getfile.php/131969702-1516957651/Dokumenter/Om%20NKK/Organisasjonen/NKK%20mener/Fakta%20om%20Avl.pdf.
[28] Norsk Kennel Klub. *Raser med registreringsrestriksjoner*. URL: https://www.nkk.no/raser-med-krav-for-registrering/category964.html.
[29] Gabriela Baers mfl. «Heritability of Unilateral Elbow Dysplasia in the Dog: A Retrospective Study of Sire and Dam Influence». I: *Frontiers in Veterinary Science* 6 (nov. 2019), s. 422.
[30] Donald E Thrall. *Textbook of veterinary diagnostic radiology*. Red. av Donald E. Thrall. 7. utg. Elsevier, 2018. ISBN: 9780323482479 (hbk.)
[31] Cristi R Cook mfl. «Diagnostic Imaging of Canine Elbow Dysplasia: A Review». I: *Veterinary Surgery* 38.2 (feb. 2009), s. 144–153.
[32] Norsk Kennel Klub. *NKKs avlsstrategi*. URL: https://www.nkk.no/nkks-avlsstrategi/category1026.html.
[33] Gry Løberg, Wenche Eikeseth og Olaf A. Roig. *hund*.
[34] Ana Válega mfl. «Digital Analysis of Subtrochlear Sclerosis in Elbows Submitted for Dysplasia Screening». I: *Frontiers in Veterinary Science* 8 (mai 2021), s. 365.
[35] Jacob Michelsen. «Canine elbow dysplasia: Aetiopathogenesis and current treatment recommendations». I: *The Veterinary Journal* 196.1 (apr. 2013), s. 12–19.
[36] S. Ekman og C. S. Carlson. «The pathophysiology of osteochondrosis». I: *The Veterinary clinics of North America. Small animal practice* 28.1 (1998), s. 17–32.
[37] L. Sjöström. «Ununited anconeal process in the dog.» I: *The Veterinary clinics of North America. Small animal practice* 28.1 (1998), s. 75–86.
[38] S. A. Johnston. «Osteoarthritis: Joint Anatomy, Physiology, and Pathobiology». I: *Veterinary Clinics of North America: Small Animal Practice* 27.4 (jul. 1997), s. 699–723.

[39] Borghild Roald. *sklerose*. URL: https://sml.snl.no/sklerose.
[40] Sebastian Raschka. *Python machine learning : machine learning and deep learning with Python, scikit-learn and TensorFlow 2*. 3. utg. Packt Publishing, 2019. ISBN: 978-1-78995-575-0.
[41] Sinno Jialin Pan og Qiang Yang. «A Survey on Transfer Learning». I: *IEEE Transactions on Knowledge and Data Engineering* 22 (2010), s. 1345–1359.
[42] Keras. *EfficientNet B0 to B7*.
[43] Anh T. Dang. *Accuracy and Loss: Things to Know about The Top 1 and Top 5 Accuracy | by Anh T. Dang | Towards Data Science*. Jan. 2021. URL: https://towardsdatascience.com/accuracy-and-loss-things-to-know-about-the-top-1-and-top-5-accuracy-1d6beb8f6df3.
[44] Aashish Nair. *Baseline Models: Your Guide For Model Building*. Apr. 2022. URL: https://towardsdatascience.com/baseline-models-your-guide-for-model-building-1ec3aa244b8d.
[45] scikit-learn developers. *sklearn.metrics.accuracy_score*. Nov. 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html.
[46] Davide Chicco og Giuseppe Jurman. «The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation». I: *BMC Genomics* 21.1 (des. 2020), s. 6–29.
[47] Bao Ngoc Huynh. «Visualization of deep learning in auto-delineation of cancer tumors». I: (2020).
[48] Orion HPC. *Home*. URL: https://orion.nmbu.no/en/home.
[49] Tsung-Yi Lin mfl. *Focal Loss for Dense Object Detection*. Tekn. rapp. Aug. 2017.
[50] Hyesoo Shim mfl. «Deep learning-based diagnosis of stifle joint diseases in dogs». I: *Veterinary Radiology and Ultrasound* 64.1 (jan. 2022), s. 113–122.
[51] TensorFlow. *keras/metrics.py at v2.11.0 · keras-team/keras*. 2022. URL: https://github.com/keras-team/keras/blob/v2.11.0/keras/metrics/metrics.py#L144-L184.
[52] Esam M.A. Hussein. «Mechanics». I: *Radiation Mechanics*. Elsevier Science Ltd, 2007. Kap. 1, s. 1–65. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780080450537500021.
[53] Mauricio Reyes mfl. «On the interpretability of artificial intelligence in radiology: Challenges and opportunities». I: *Radiology: Artificial Intelligence* 2.3 (mai 2020). DOI: 10.1148/RYAI.2020190043/ASSET/IMAGES/LARGE/RYAI.2020190043.FIG5.JPEG.
[54] Mari Nyborg Hauback. *Personlig kommunikasjon*. 2023.
[55] David Sterratt. *Principles of computational modelling in neuroscience*. eng. Cambridge, United Kingdom: Cambridge University Press, 2011. ISBN: 978-0-521-87795-4.
[56] Valentino Zocca. *Python deep learning : next generation techniques to revolutionize computer vision, AI, speech and data analysis*. 1. utg. 2017. ISBN: 1-78646-066-1.
[57] Magne Brekke og Arne Borthne. *røntgenundersøkelse*. Apr. 2022. URL: https://sml.snl.no/r%C3%B8ntgenunders%C3%B8kelse.
[58] Witold Skrzynski. «X-Ray Detectors in Medical Imaging». I: *Advanced X-ray Detector Technologies* (2022), s. 135–149.

Vedlegg A

Tabell 6.1: *Tabell med alle datasett brukt en eller flere ganger i denne masteroppgaven. For hvert datasett er problemtypen datasettet er brukt til oppgitt, i tillegg til antall prøver fra hver diagnosegruppe i datasettet. En oversikt over datasettene er også gitt i* Elbow_Experiments.xlsx *på* https://github.com/huynhngoc/cubiai.

Navn på datasett	**Problemtype**	**Normale**	**1 artrose/ sklerose**	**2 artrose**	**2 mistanke MCD**	**3 artrose**	**3 MCD**	**3 OCD**	**3 UAP**	**Total**
800_normal_abnormal_1	Skille mellom normal og abnormal	266	256	131	55	48	92	5	9	862
800_normal_abnormal_2	Skille mellom normal og abnormal	500	257	124	53	54	82	8	12	1090
1280_normal_abnormal_2	Skille mellom normal og abnormal	500	257	124	53	54	82	8	12	1090
640_normal_abnormal_2	Skille mellom normal og abnormal	500	257	124	53	54	82	8	12	1090
800_level_3	Skille mellom nivå 1, nivå 2 og nivå 3	0	359	174	74	76	115	8	17	823
800_complete_3	Skille mellom alle diagnosegrupper, uten normale.	0	359	174	74	76	115	8	17	823
800_ext_binary_2	Skille mellom normal og abnormal	601	770	371	159	162	248	8	36	2355
800_arthrosis_vs_rest	Skille mellom alle prøver med artrose og øvrige diagnoser, uten normale.	0	205	99	127	43	198	8	29	709
800_lvl1_vs_rest	Skille mellom alle prøver fra nivå 1 og øvrige diagnoser, uten normale.	0	514	248	106	108	165	8	24	1173
800_binary_error_predictor_2	Skille ikke klassifiserbare fra klassifiserbare prøver	103	159	44	48	19	33	1	3	410
800_complete_ext_binary_2	Skille mellom normal og abnormal	1779	770	371	159	162	248	8	36	3533
800_complete_ext_binary_feedback_2	Skille mellom normal og abnormal	1714	730	363	132	161	239	8	36	3383
800_only_wrong_2	Kun feilpredikerte prøver fra ekstern test på klassifisering av normale og abnormale prøver. Dette datasettet brukes kun til omtrening, ikke til testing	59	40	8	27	1	9	0	0	144
800_complete_ext_level_3	Skille mellom nivå 1, nivå 2 og nivå 3	0	418	322	138	140	215	8	31	1272

Vedlegg B

Alle feilpredikerte røntgenbilder fra modellen med høyest ytelse uten omtrening ved binær klassifisering av normale og abnormale prøver vises i dette vedlegget. Hver prøve er merket med sann diagnosegruppe, og predikert sannsynlighet for å tilhøre abnormal gruppe. Sann diagnose indikerer grunnsannheten til hvert bilde med spesifikk diagnosegruppe, selv om modellen bare skilte mellom normale og abnormale røntgenbilder.

Dersom predikert sannsynlighet var mer enn 0.5 ble prøven klassifisert som abnormal, og for sannsynligheter lavere enn 0.5 ble prøvene klassifisert som normale. For eksempel indikerer markeringen "sann diagnose normal og predikert sannsynlighet 0.59" at en prøve uten albueleddsdysplasi ble predikert som abnormal med en sannsynlighet på 0.59.

Fordelingen av sannsynlighetene til disse feilprediksjonene er presentert i figur 4.5.

Sann diagnose: normal,
Predikert sannsynlighet: 1.00

Sann diagnose: normal,
Predikert sannsynlighet: 0.59

Sann diagnose: normal,
Predikert sannsynlighet: 1.00

Sann diagnose: normal,
Predikert sannsynlighet: 0.86

Sann diagnose: normal,
Predikert sannsynlighet: 0.64

Sann diagnose: normal,
Predikert sannsynlighet: 1.00

Sann diagnose: normal,
Predikert sannsynlighet: 0.98

Sann diagnose: normal,
Predikert sannsynlighet: 0.99

Sann diagnose: normal,
Predikert sannsynlighet: 0.54

Sann diagnose: 1, artrose/
sklerose,
Predikert sannsynlighet: 0.10

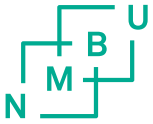
Sann diagnose: 1, artrose/
sklerose,
Predikert sannsynlighet: 0.03

Sann diagnose: 1, artrose/
sklerose,
Predikert sannsynlighet: 0.14

Sann diagnose: 1, artrose/
sklerose,
Predikert sannsynlighet: 0.01

Sann diagnose: 2, artrose,
Predikert sannsynlighet: 0.04

Sann diagnose: 2, artrose,
Predikert sannsynlighet: 0.28



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

Postboks 5003
NO-1432 Ås
Norway