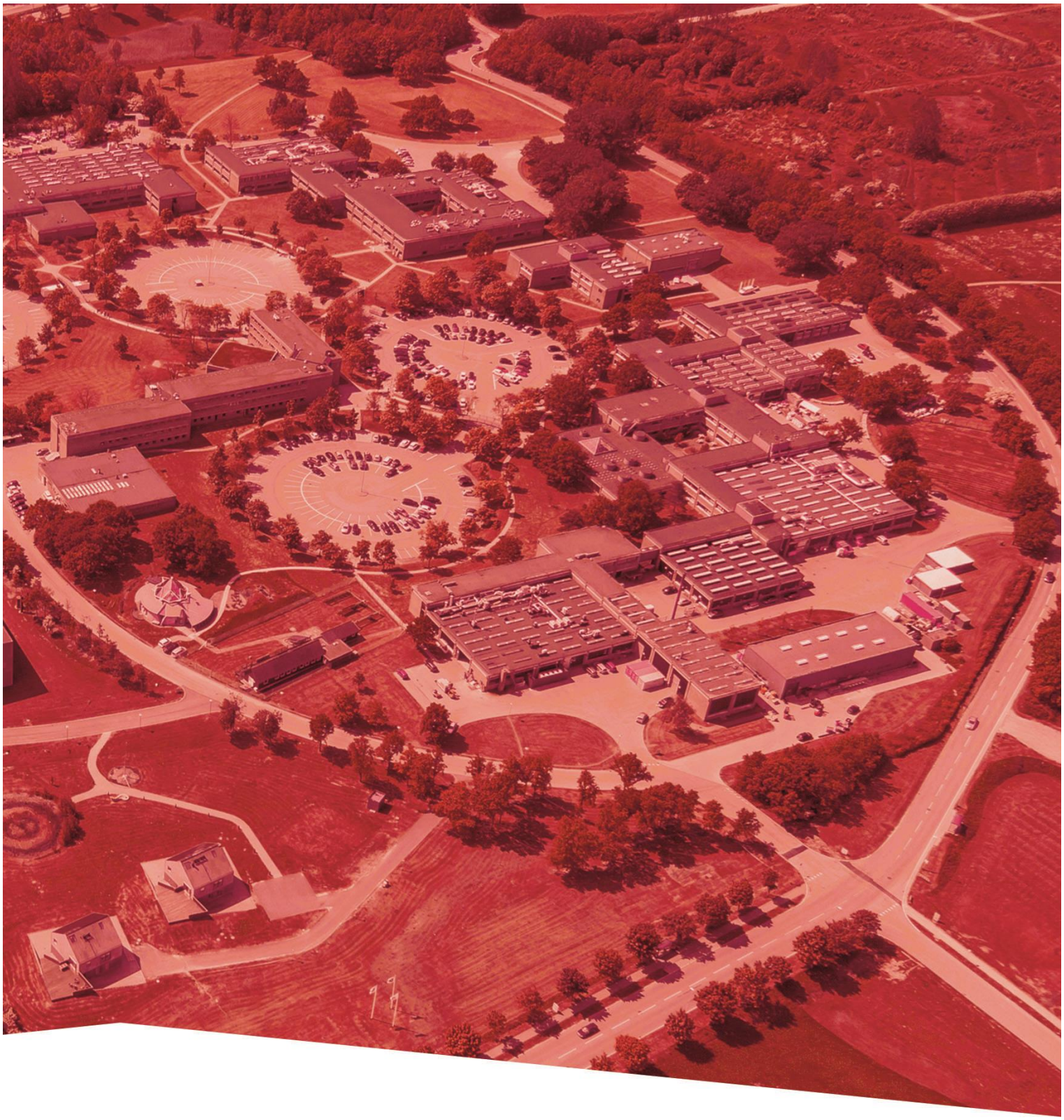


Computer vision fæces-score klassifikation for smågrise

TEKNOLOGISK
INSTITUT

TEKNOLOGISK
INSTITUT

Computer vision fæcesscore klassifikation for smågrise

Udarbejdet for:
Svinerådgivningen I/S
Birk Centerpark 24, 7400 Herning

Udarbejdet af
Teknologisk Institut
Agro Food Park 15
8200 Aarhus N
Landbrug og Digitalisering
Svend Vendelbo Nielsen (sven@teknologisk.dk)

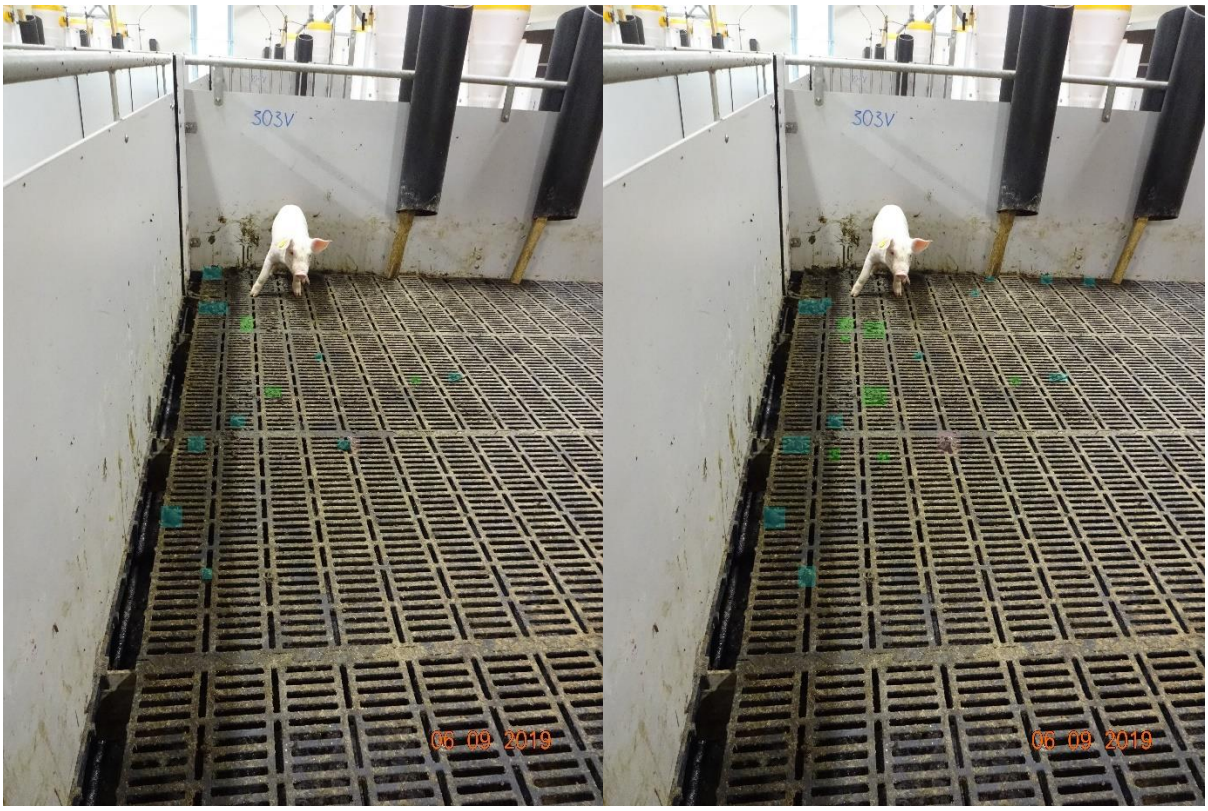
Maj 2023

TEKNOLOGISK
INSTITUT

Introduktion

I dette projekt udført som et A.I Danmark projekt har Teknologisk Institut i samarbejde med Svinerådgivningen lavet en computer vision model, som kan estimere lokation og score af fæces på et billede af et spaltegulv i en svinestald (Se Figur 1).

Da fæces på et staldgulv forekommer i forskellige størrelser og friskheder, og samtidig kan være trådt på, er det svært for menneskelige eksperter at være 100% konsistente i annoteringen af fæces. Vi har optimeret algoritmen efter at ramme den gennemsnitlige fæcesscore over hele billedet ud fra én eksperts annoteringer.



Figur 1: Et eksempel på et billede fra en svinestald samt modellens forudsigelser. På venstre side ses de manuelt annoterede fæces af modellen markeret med blå (karakter 1), grøn (karakter 2) og lilla (karakter 4). På højre side ses modellens prædikterede fæces. Prædiktionsmodellen er ikke trænet med det pågældende billede.

Arbejdsfordeling

Svinerådgivning har taget alle billederne og annoteret alle fæcesscorer. Teknologisk Institut har valgt, trænet og valideret modellen og skrevet denne rapport.

Metode

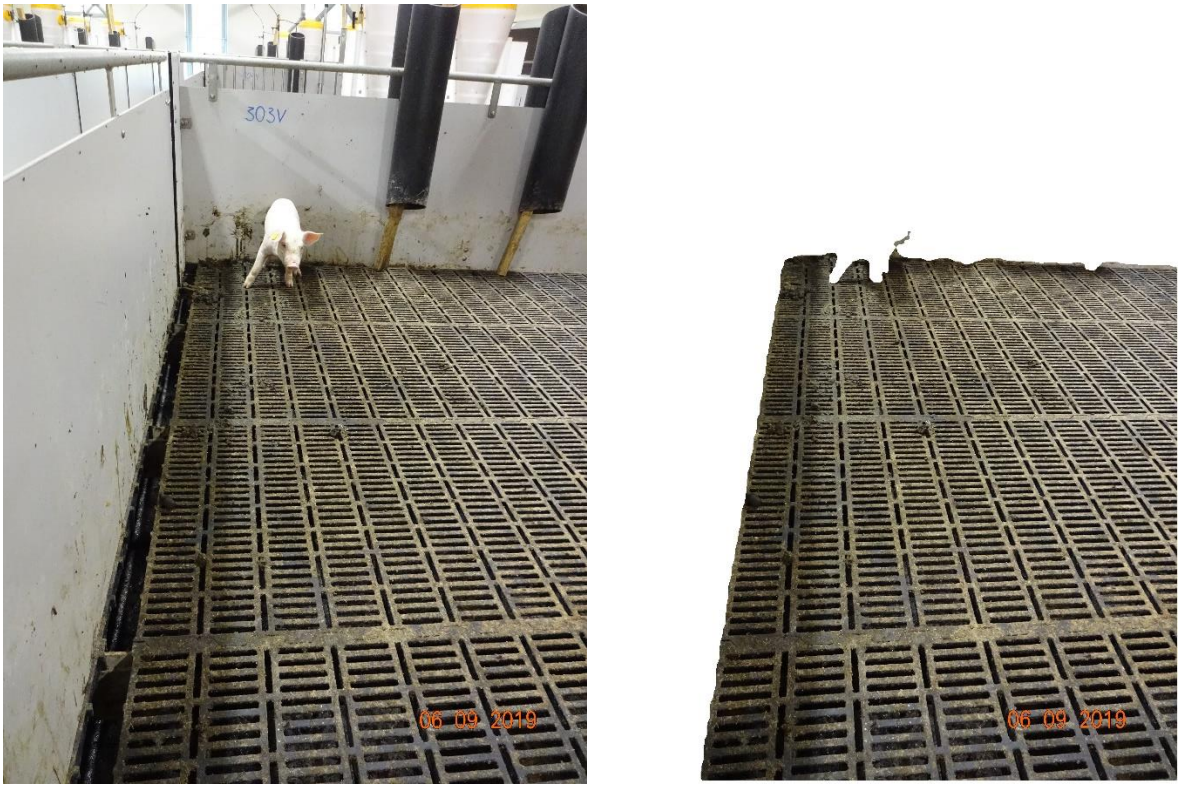
Billederne er indsamlet i et afsluttet projekt, hvor de indgik som dokumentation for manuelt opgjorte fæcesscorer. De er alle højtopløselige billeder taget indendørs i dagslys med det samme kamera.

Placering og scorer for fæces er annoteret på 219 billeder. Et eksempel kan ses på Figur 1. Heraf blev 24 billeder brugt som test-sæt mens 39 blev brugt som et valideringssæt. På alle billederne i trænings-, test- og valideringssættet er de vigtigste fæces annoteret, mens der på testsættet også er annoteret flere små fæces.

Til at lave detektionerne har vi gentrænet en pre-trænet Yolo V8 model, der bygger på et convolutional neural network til at finde og klassificere fæces ud fra 3 scorer: 1,2 og 4, der indikerer, hvor flydende fæces er, hvilket er en indikator for diarré. Oprindeligt arbejdede vi med scorerne 1,2,3 og 4, men scoren 3 var næsten uadskillelig fra scoren 2 – både i annotering og de første modellers præcision – så vi slog den sammen med score 2. Vi overvejede samtidigt at ændre skalaen fra 1,2 og 4 til 1,2 og 3, men vi beholdt 1-2-4-skalen for at de mest kritiske fæces af score 4 vægtede højere, når man udregner en gennemsnitlig fæcesscore.

Fordi størrelsen på en enkelt fæces er meget lille i forhold til det samlede antal pixels i billedet, kan det kræve rigtigt meget tid at lære det neurale network ikke at fjerne den information, der ligger i de få pixels. Dette kaldes med et fagord Small Object Detection. For at undgå problemer med dette, splittede vi alle billeder op i subbilleder af størrelse 512x512 og trænede algoritmen på disse subbilleder.

For at forbedre modellens performance har vi trænet en Yolo V8 segmenteringsmodel til at genkende gulvet og begrænse prædiktionerne til at skulle ligge på dette genkendte gulv (Se Figur 2). Denne model blev trænet og valideret på blot 60 billeder.



Figur 2: Detektion af gulvet. Inden modellens prædiktion fjernes alt det, som ikke er gulv, ved hjælp af en computer vision model.

Den mest optimale følsomhed for modellen blev estimeret ved at vælge en detektionsgrænse, der både gav en høj R^2-værdi for den gennemsnitlige fæcesscore, men også proportionen af score 4 fæces.

Evaluering af model

For at afgøre om en annoteret boks er det samme som en prædikteret boks, bruger vi det lempelige kriterium at Intersection over Minimum skal være mindst 0.5. Det betyder at en annoteringsboks (Figur 1, venstre side) siges at være i overensstemmelse med en prædiktionsboks (Figur 1, højre side), hvis deres overlappende areal er større end halvdelen af den mindste af de to bokse. Tilsvarende for hvornår en prædiktionsboks er i overensstemmelse med en annoteringsboks. Det vil sige at hvis en fæces bliver prædikteret som flere fæces bliver alle prædiktionerne talt som rigtige.

Vi bruger også målene Recall og Precision, som skal fortolkes som følgende:

- Recall angiver, hvor stor en procentdel af de annoterede bokse, der er i overensstemmelse med mindst en prædikteret boks

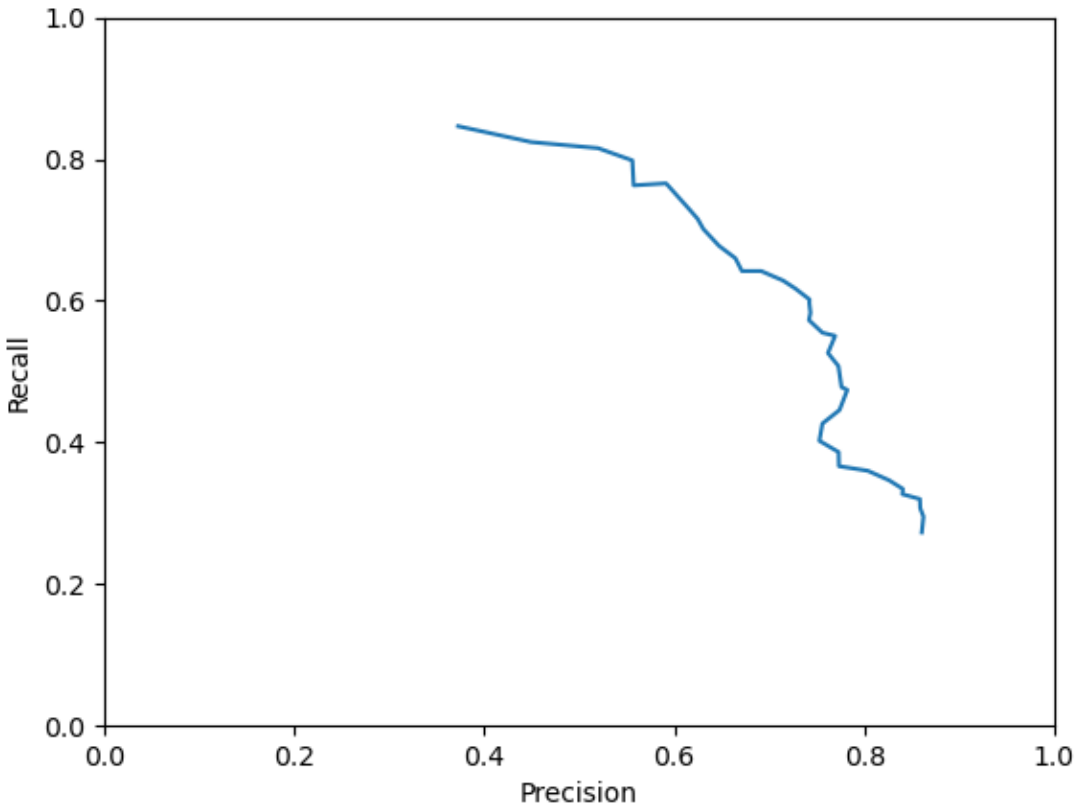
- Precision angiver, hvor stor en procentdel af de prædikterede bokse, der er i overenstemmelse med mindst en annoteret boks.

En følsom model, som kommer med rigtigt mange prædiktioner, som den ikke er sikker på, vil oftest have et højt recall, fordi der er så mange muligheder for at en annoteret boks bliver dækket af en prædikteret boks. Dog vil den ofte have en lav precision, fordi mange af de prædikterede bokse ikke er annoterede. Omvendt vil en konservativ model, som kun prædikterer de bokse den er meget sikker på, have et lavt recall men en høj precision.

Resultater

Vi evaluerer modellen ved at sammenligne de manuelt markerede annoteringer på testsættet med modellens predikterede på de samme billeder.

Ud fra det annoterede datasæt har vi udregnet en såkaldt ROC-kurve der sammenligner de mulige precision og recall, som modellen kan have (Figur 3). Den viser, at vi kan justere modellens følsomhed således at modellen kan have recall på 80% og precision på 50% - eller på den anden side have en precision på 80% og en recall på 35%. Dette spænd svarer til at variere detektionsgrænsen – hvilket vil sige hvor sikker modellen er på hver enkelt prædikteret boks – mellem 1 og 40%.

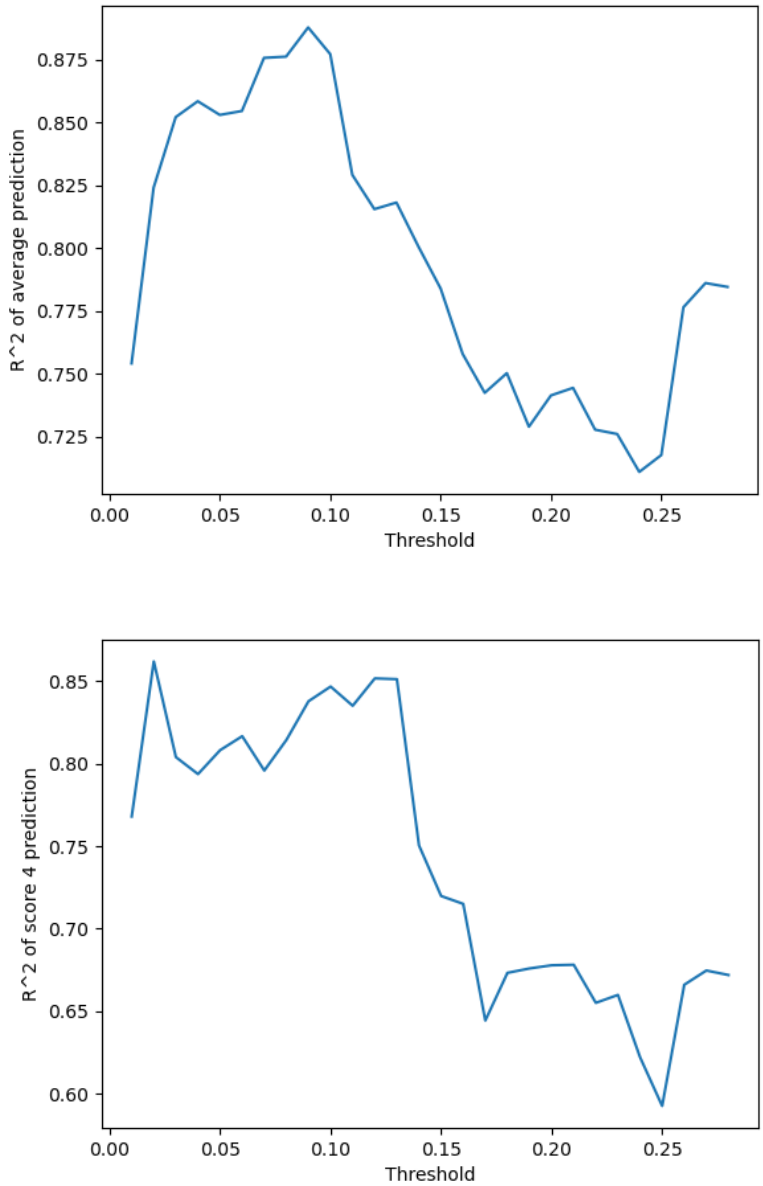


Figur 3: Denne figur viser Recall og Precision for forskellige detektionsgrænser. Den ujævne facon skyldes det lille testsæt.

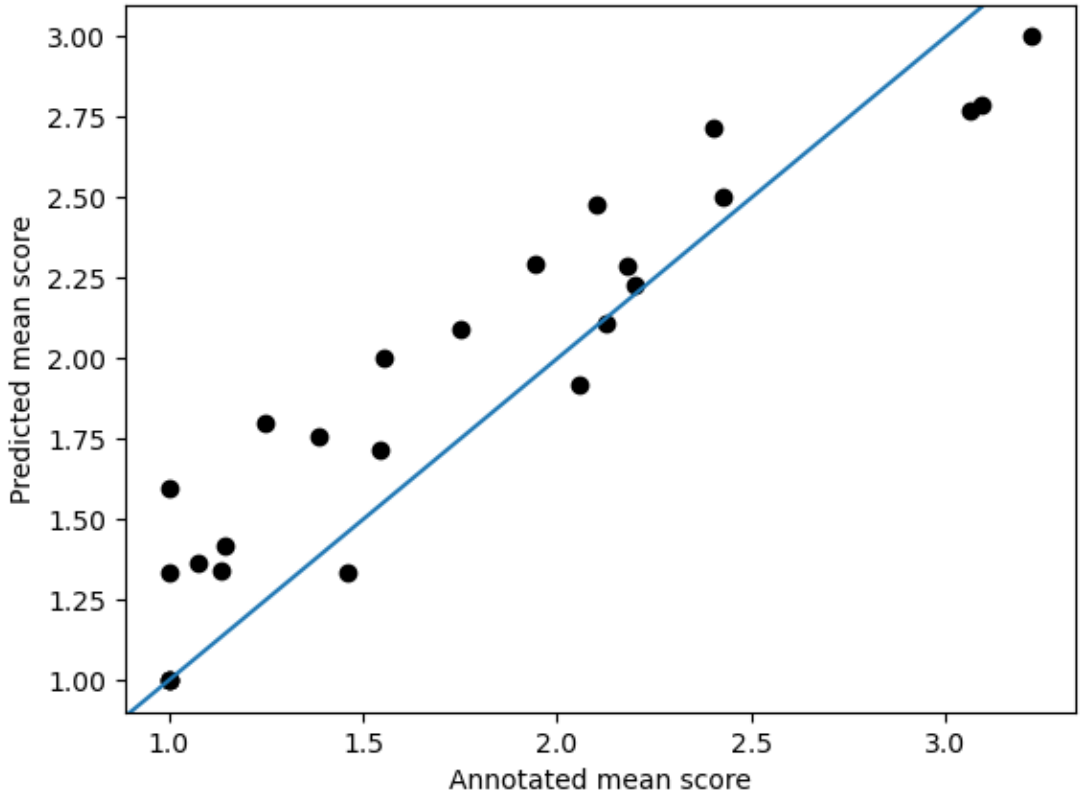
Da et af hovedmålene er at estimere en gennemsnitlig fæcesscore, valgte vi den afvejning mellem præcision og recall, der ville give gennemsnitsscoren den bedste prædiktionsevne. På Figur 4 ses hvordan præcisionen af henholdsvis gennemsnitsfæcesscoren og proportion af score 4 fæces afhænger af detektionsgrænsen. På baggrund af disse grafer valgte vi detektionsgrænsen på 10%, fordi det giver en R^2 på 0.88 for den gennemsnitlige fæcesscore og en 0.83 for proportionen af score 4 fæces.

Valget af detektionsgrænse på 10% certainty giver en precision på 65% og recall på 68%. For den gennemsnitlige fæcesscore giver det en gennemsnitlig afvigelse på 0.24. For proportionen af fæces med score 4 giver det en gennemsnitlig, procentpoint afvigelse på 6.4%.

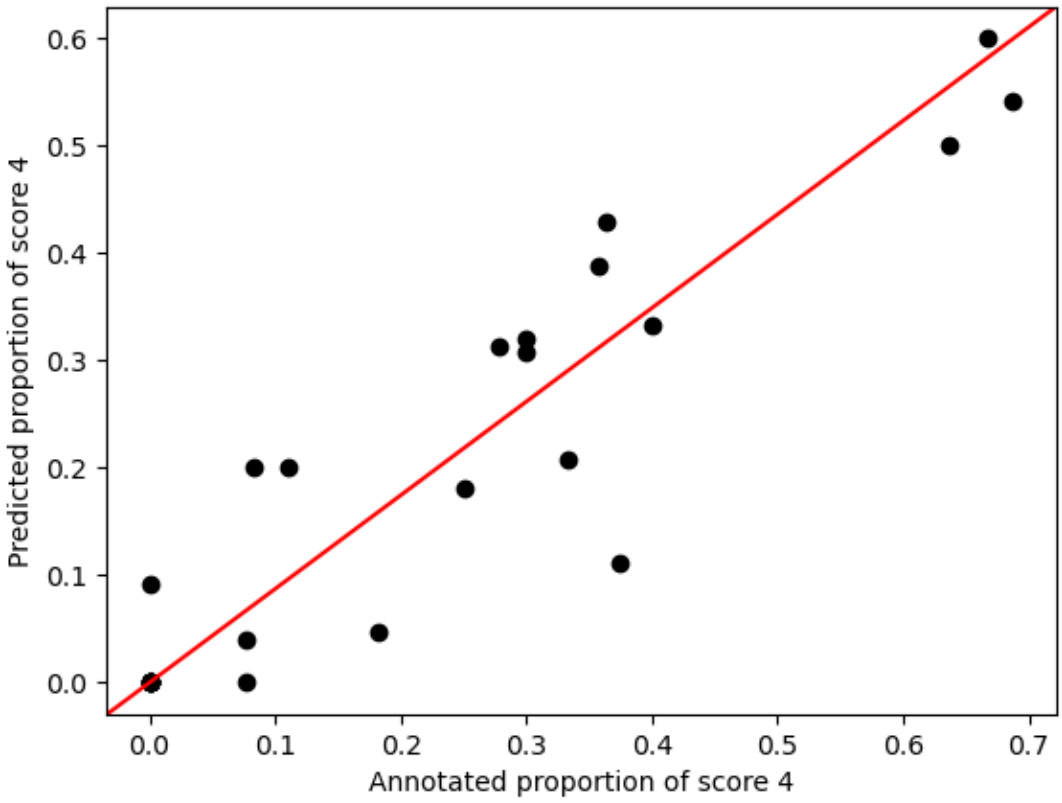
På Figur 5 og 6 ses de prædikterede og annoterede gennemsnitsfæcesscorer samt proportion af fæcescore 4 på testdata'et, når detektionsgrænsen er sat til det optimale 0.1.



Figur 4: Øverst vises R^2 mellem af prædiktionens gennemsnitlige fæcesscore for forskellige detektionsgrænser/følsomheder og nederst vises tilsvarende R^2 for proportion af fæcesscore 4.



Figur 5: Den gennemsnitlige fæcesscore i annoteringen sammenlignet med den gennemsnitlige fæcesscore i prædiktionen på de 24 testbilleder. Det ses at modellen er biased i den forstand at den prædikterer for høje værdier, når den sande score er lav og den prædikterer for lave værdier når den sande gennemsnitsscore er høj



Figur 6: Proportionen af klasse 4 fæces i annoteringen sammenlignet med proportionen af fæcesscore 4 i prædiktionen på de 24 testbilleder.

Tabel 1 og 2 angiver de observerede prædiktionsfejl på datasættet. Det ses der er få fejl, hvor en stadie 4 fæces bliver prædikteret som 1 eller 2 - og omvendt.

Tabel 1: Denne tabel viser hvor godt de prædikterede bokse var i overenstemmelse med de annoterede bokse på testsættet. Det vil sige at blandt de bokse, der var prædikteret med score 1, var 102 også annoteret som 1, mens 3 var annoteret som 2 og 55 var slet ikke annoteret.

Prædikteret	Ikke annoteret	Annoteret som 1	Annoteret som 2	Annoteret som 4
1	55	102	2	0
2	75	33	24	3
4	17	1	5	53

Tabel 2: Tabellen viser hvor godt de annoterede bokse var i overenstemmelse med prædikterede bokse på test-sættet.

Annoteret	Ikke prædikteret	Prædikteret som 1	Prædikteret som 2	Prædikteret som 4
1	74	102	33	1
2	5	2	24	5
4	19	0	3	53

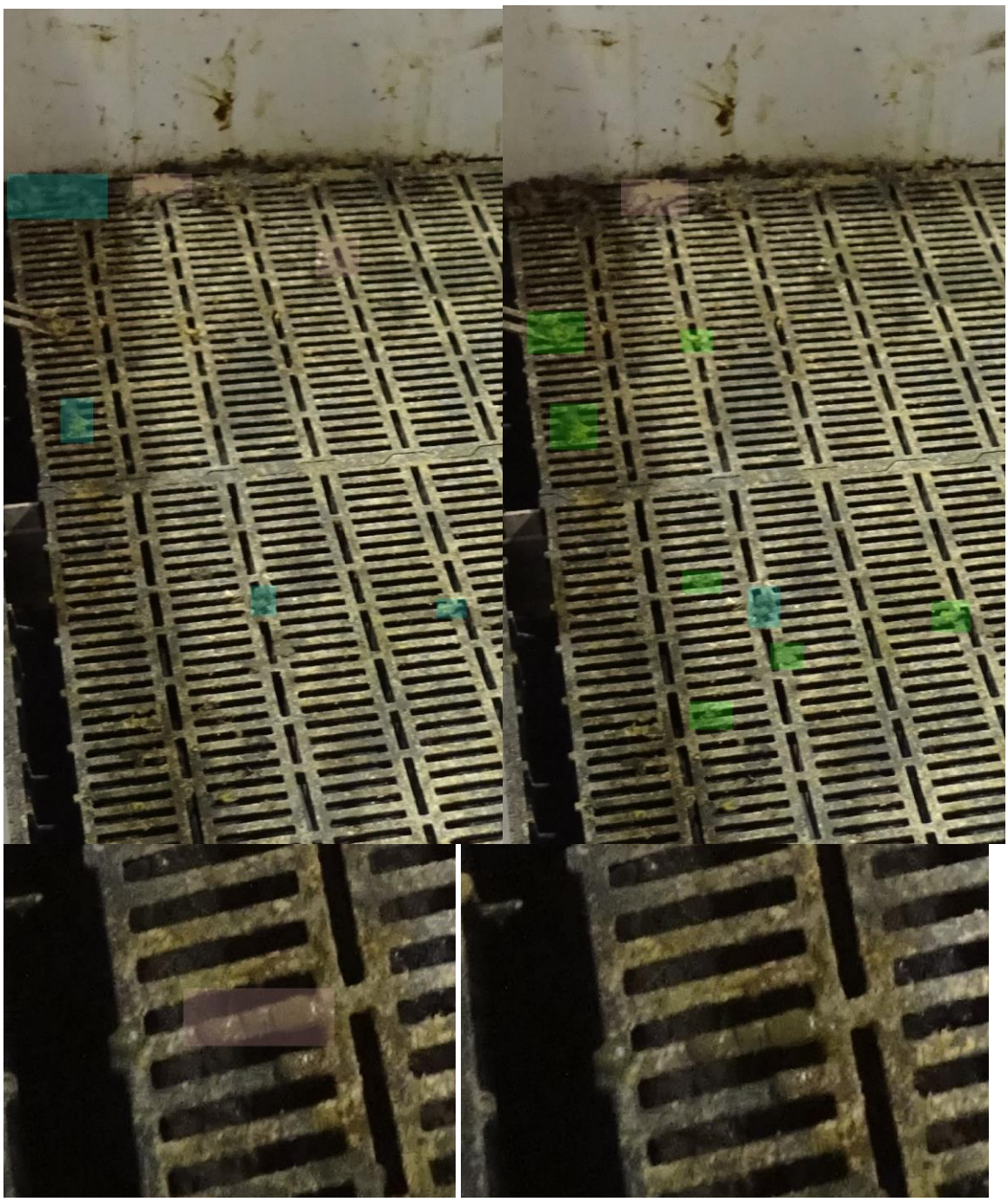
I det ovenstående har vi udregnet den gennemsnitlige fæcesscore ved at tælle det rå antal fæces af hver score i billederne. Vi har også eksperimentet med at vægte antallet af fæces med arealet og modellens confidence af dem. Det er opsummeret i Tabel 3

Tabel 3: Performance for den gennemsnitlige fæcesscore og proportionen, der er 4 for forskellige måder at vægte de enkelte fæces (Weighing) og for forskellige detektionsgrænser (threshold). Tabellen viser, at der ikke er nogen måder, der giver en stor forbedring over for de rå counts med detektionsgrænsen 0.10. Dog er vægtningen med både confidence(certainty) og areal ikke god for proportionen af score 4 fæces.

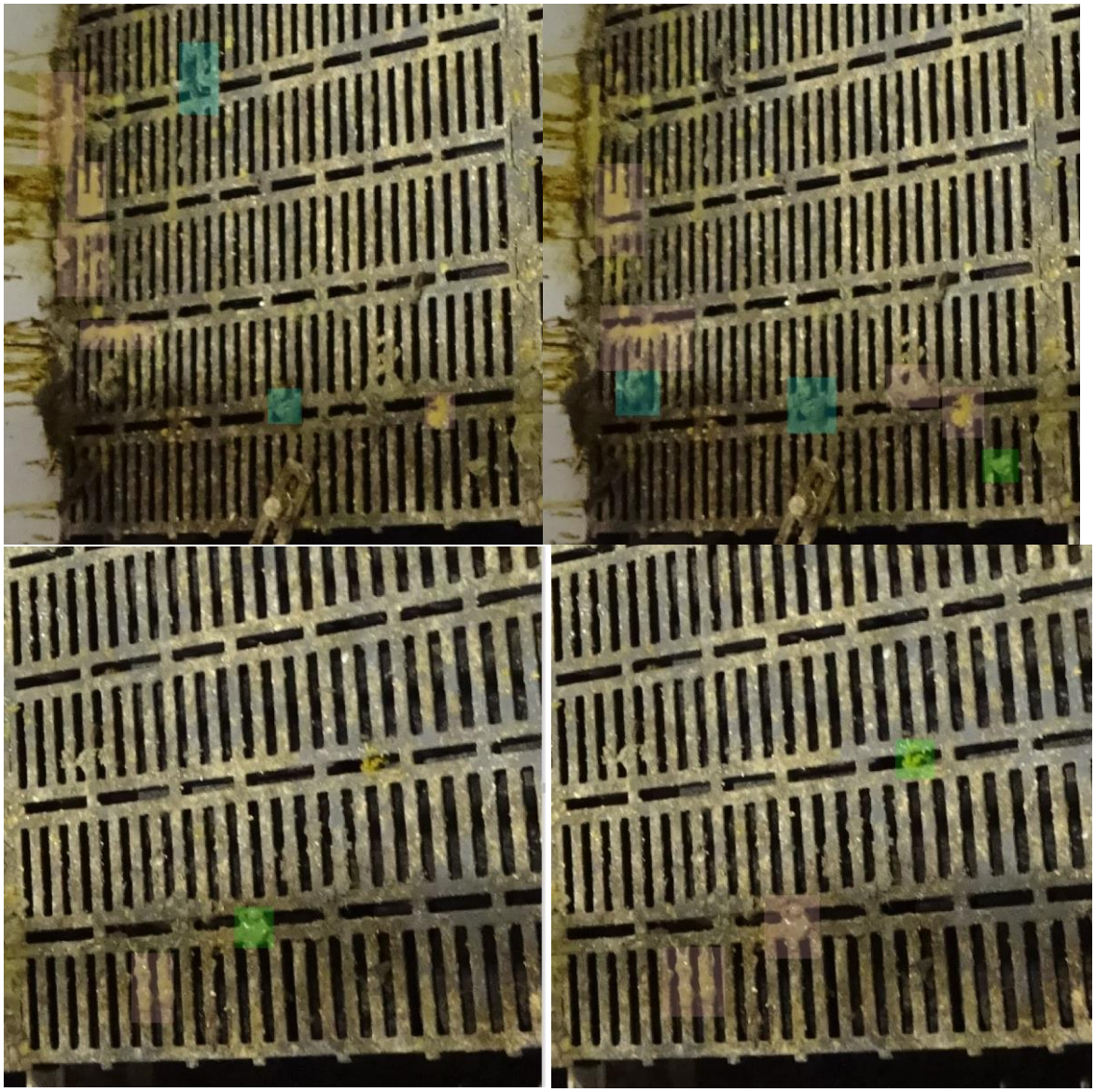
	Measure	Weighing	R2	Spearman rho	deviance	threshold
0	average	count	0.853120	0.934317	0.267573	0.05
4	average	weighted_area	0.893542	0.941272	0.224598	0.05
2	average	weighted_certainty	0.870034	0.949000	0.216048	0.05
6	average	weighted_certainty_area	0.841607	0.936635	0.240801	0.05
8	average	count	0.877320	0.941453	0.237077	0.10
12	average	weighted_area	0.861139	0.933544	0.233009	0.10
10	average	weighted_certainty	0.868190	0.942817	0.209263	0.10
14	average	weighted_certainty_area	0.820253	0.928907	0.250789	0.10
1	proportion	count	0.836731	0.891057	0.058696	0.05
5	proportion	weighted_area	0.859851	0.930054	0.068325	0.05
3	proportion	weighted_certainty	0.804154	0.918166	0.067177	0.05
7	proportion	weighted_certainty_area	0.747052	0.903176	0.094050	0.05
9	proportion	count	0.829918	0.915274	0.064144	0.10
13	proportion	weighted_area	0.793780	0.909579	0.084859	0.10
11	proportion	weighted_certainty	0.784742	0.921670	0.072122	0.10
15	proportion	weighted_certainty_area	0.713318	0.906399	0.099821	0.10

Eksplorativ undersøgelse af fejlprædiktioner

Testbillede 17 er det billede med størst (procentpoints) afvigelse på antallet af fæces med scoren 4. Dens prædiktioner er undersøgt i Figur 7



Figur 7: Annoteret mod prædikteret på billede 17. De venstre billeder er de annoterede og de højre, de prædikterede. Det ses at prædiktionen finder score 4 fæces oppe i toppen af billedet men ikke de to andre. Dog genkender den møtrikken som en score 2 fæces, mens den genkender flere af de udtværede fæces som score 2.



Figur 8: Annoterede (venstre) mod prædikterede (højre). Der ses en forskel mellem klasser og hvilke fæces der bliver detekteret.

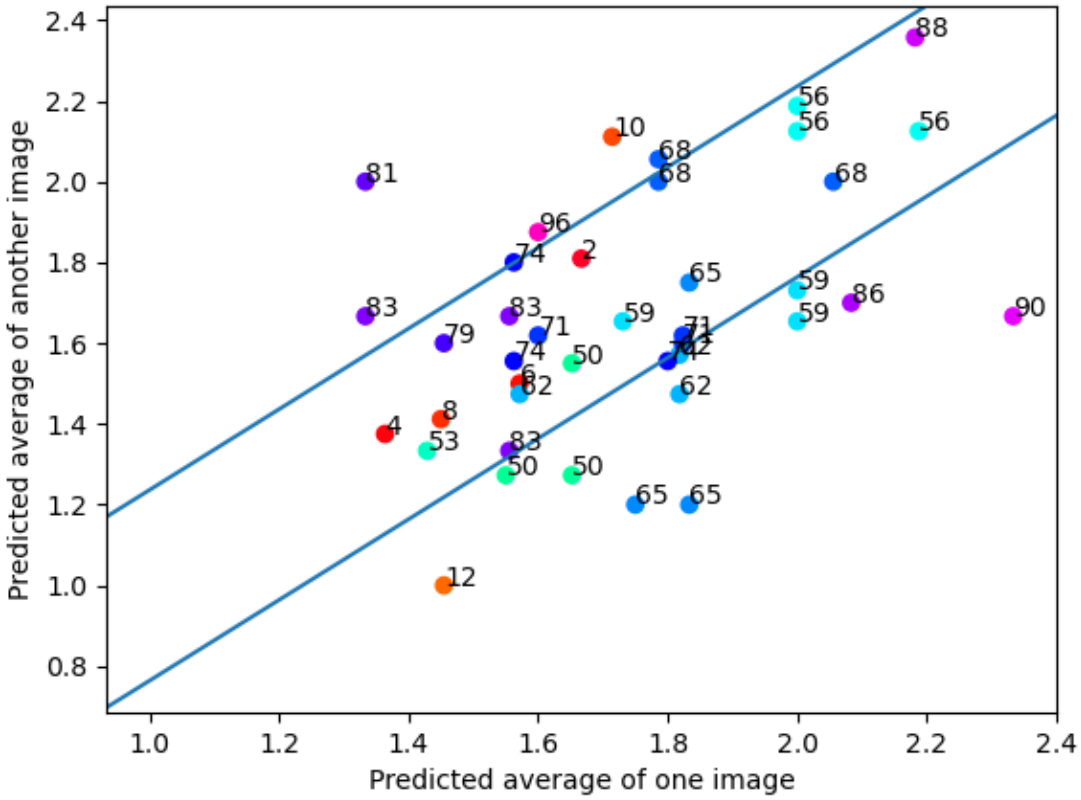
Test med nyt kamera

En anvendelse af denne model kunne involvere, at brugere tager billeder med deres egne kameraer. For at teste brugbarheden af dette, er der taget nogle nye billeder fra forskellige vinkler i en anden svinestald. Det nye Kamera havde opløsning 1536x2048, hvorimod træningsdataet havde opløsning

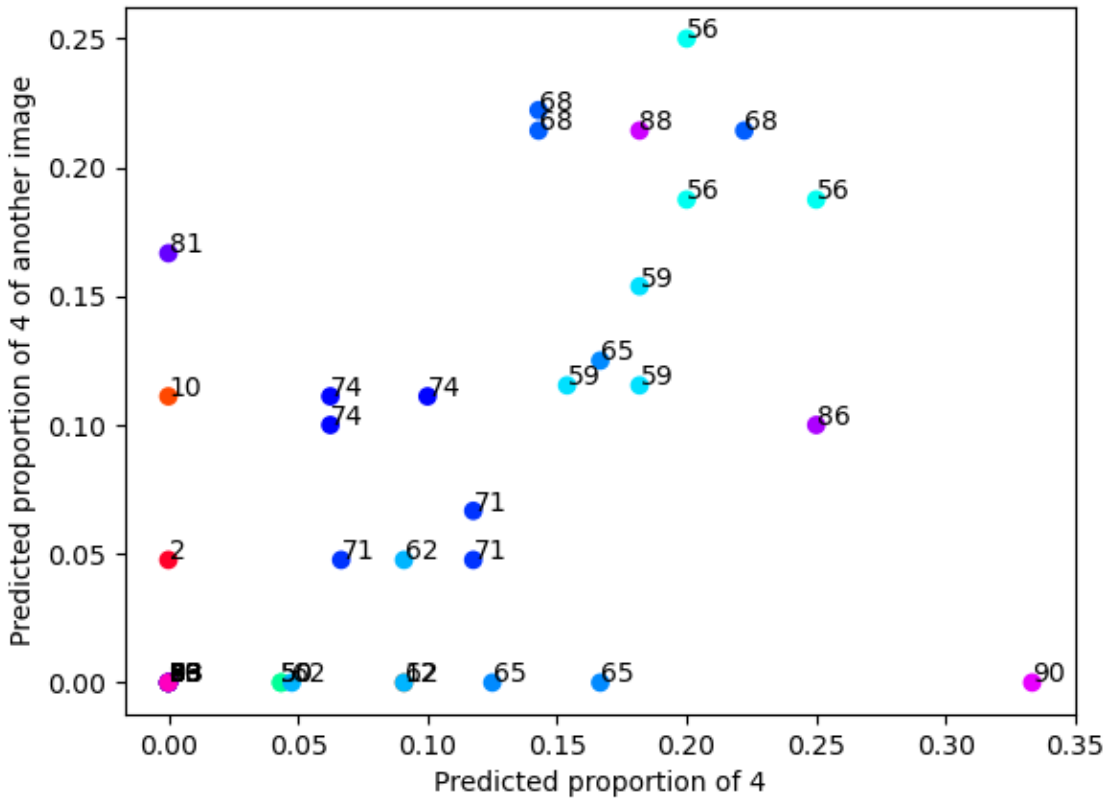
3672x4896. På det nye testsæt blev taget 66 billeder af 28 unikke gulve, hvoraf der blev taget 2-3 billeder af hvert gulv. På Figur 9 ses en side-by-side sammenligning

Figur 9: Modellen er trænet på billeder som vist til venstre. Billedet til højre viser de nye billeder, som modellen også bliver testet på i denne rapport. Bemærk forskellen i farvetone.

Algoritmen for at finde gulvet på billedet fejlede på 8 ud af 66 billeder – oftest på de billeder hvor gulvet udgjorde mere end 80% af billedet eller på uskarpe billeder. På alle de resterende billeder blev der estimeret fæces. Blandt disse 58 billeder var den gennemsnitlige forskel mellem to fæcesscorer fra forskellige billeder, men på det samme gulv, 0.24 (Se Figur 10). Den gennemsnitlige forskel på proportionen af stadie 4 fæces var 5.5 procentpoint.

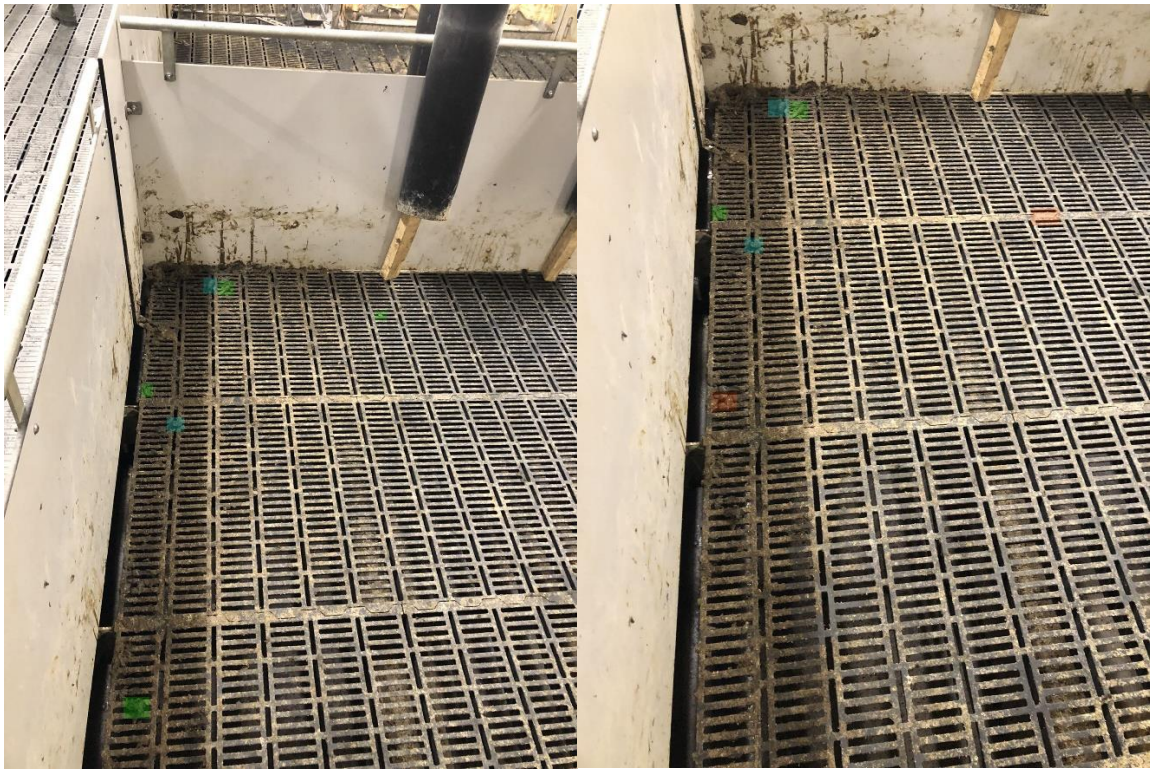


Figur 10: Et plot af den gennemsnitlige fæcesscorer blandt alle par af test-billeder taget af samme staldgulv. Tallet og farven angiver hvilket gulv, de hører til. De to linjer angiver den gennemsnitlige forskel på 0.24.

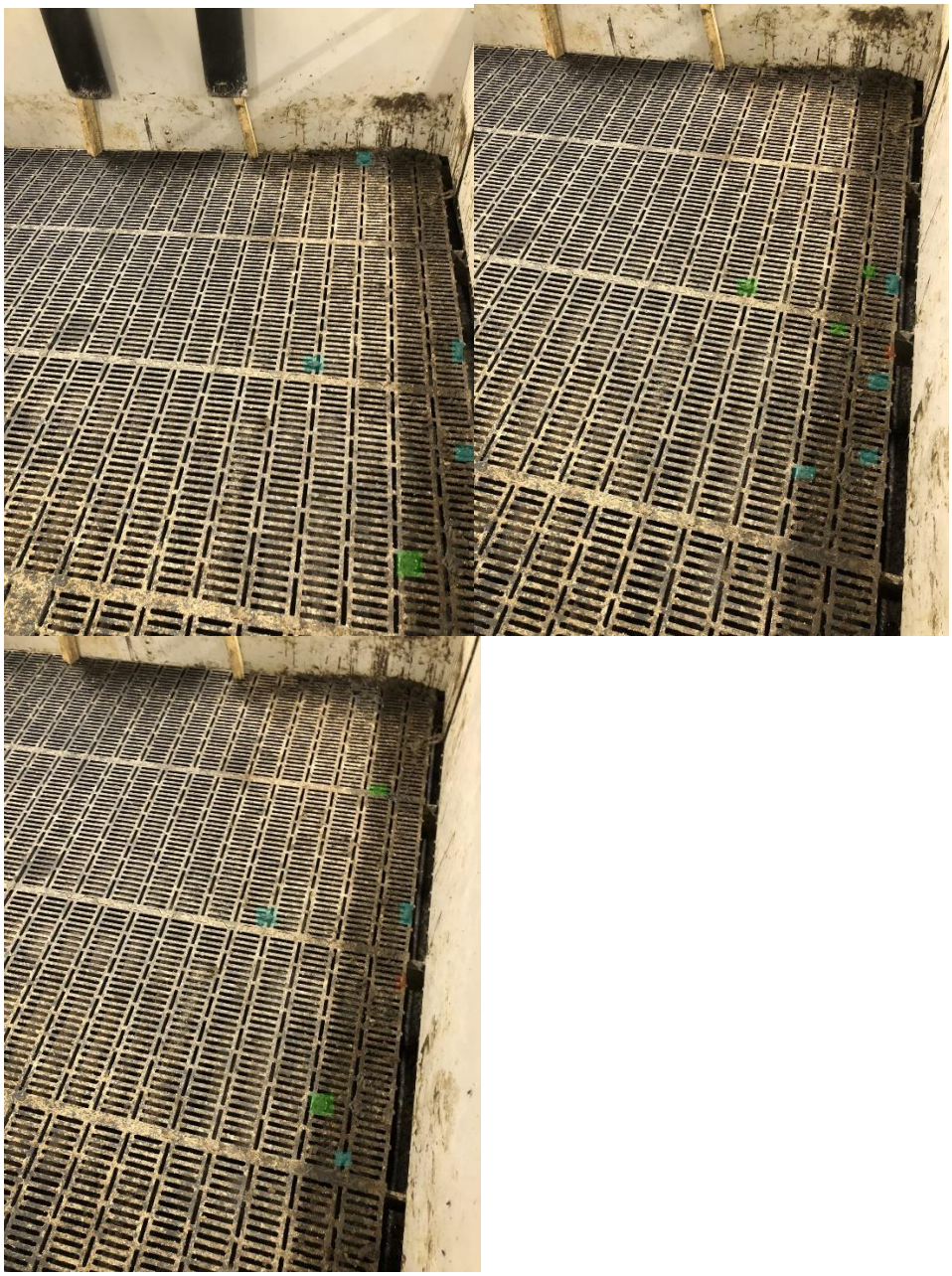


Figur 11: Et plot af proportionen af fæces score 4 blandt alle par af test-billeder taget af samme staldgulv. Tallet og farven angiver hvilket gulv, de hører til.

Gulv nummer 90 havde den største usikkerhed i prædiktionen (Figur 12), og der var også stor forskel på gulv nummer 65 (Figur 13).



Figur 12: Billederne på gulv 90 med den største forskel i gennemsnitlig fæcesscore. Forskellen skyldes at billedet til venstre ikke klassificerede den legitime score 4 fæces i venstre side af billedet, samt at billedet til højre ved en fejl klassificerede et stykke gulv som score 4.



Figur 13: Billeder med prædiktioner for gulv 65, der gav den næsthøjeste forskel i gennemsnitlig fæcesscore. Forskellen skyldes hovedsageligt den manglende detektion af fæces af score 4 på det første billede. På dette første billede lå fæces også yderst til højre i billedet, hvor det kan forventes at detektionsgraden er lavest.

Diskussion

Den endelige model har en precision på 65% og et recall på 68% målt for Intersect over minimum på mindst 0.5. Det er forholdsvist lavt sammenlignet med computer vision metoder for andre problemstillinger. Grunden til den numerisk lave performance skyldes blandt andet at det er svært at annotere billederne konsistent. Fæces varierer i størrelse, friskhed og i hvor høj grad de er blevet trådt på, og dermed varierer det også hvordan forskellige eksperter ville annotere et billede, og der kan også være variation inden for den samme ekspert på forskellige dage. Vi kan derfor forestille os dette hierarki af estimatorer, hvor vi starter med den bedste

1. Et antal eksperter annoterer alle fæces på billedet, og vi bruger en form for gennemsnit af deres annoteringer.
2. En enkelt ekspert annoterer alle fæces på billedet på forskellige dage, og vi bruger en form for gennemsnit af disse annoteringer.
3. En enkelt ekspert annoterer alle fæces på billedet en enkelt gang, og vi bruger disse annoteringer.

I dette projekt har vi brugt estimator 3 som "sandheden" og vist at det har givet os en computer vision estimator for den gennemsnitlige fæcesscore med en gennemsnitlig afvigelse på 0.24. Ved at bruge estimator 1 eller 2 som "sandheden" i vores computer vision model kunne vi formentligt få en mere præcis estimator for den gennemsnitlige fæcesscore.

I testen med et andet kamera, så vi at den gennemsnitlige afstand mellem den gennemsnitlige fæcesscore af to billeder på det samme spaltegulv var 0.24. Dette kunne antyde at prædiktionens variation på grund af kameravinkel kan forklare hele testerror'eren som også var på præcist 0.24. Det kan vi dog ikke entydigt afgøre da testen med dette alternative kamera har en forventeligt højere usikkerhed. Ikke desto mindre tyder det på at modellen kan forbedres ved at trænes på mere alsidigt data. I Figur 12 så vi for eksempel en tydelig misklassifikation af en fæces score 4 på revnen mellem to spaltegulvsplader. Dette kunne undgås hvis modellen havde set flere billeder i det lys. Derudover fejlede algoritmen for at finde spaltegulvet på 8 ud af 66 billeder, hvilket også kan repareres ved mere træning.

Endelig viser testen med det andet kamera, at man med fordel kan tage flere billeder af det samme spaltegulv og kombinere prædiktionerne på de enkelte billeder.

Hvis hovedmålet er at estimere proportionen af fæces med scoren 4, kan det være en fordel at bruge detektionsgrænsen 0.05, da det giver bedre præcision (Tabel 3).

Diskussion: Hvor god modellen skal blive

I dette projekt har vi afsøgt hvor god computer vision modellen kan blive uden et egentlig mål for hvad der er godt. Inden man kaster mere energi i at forbedre modellen yderligere, er det en god ide at overveje om der også er en gevinst i praksis.

Formålet med modellen er at sige hvordan grisenses helbredstilstand er. De fæces, man kan se på et staldgulvsbillede er en indikator på grisenes helbredstilstand, men ikke en perfekt repræsentation. Vi kan derfor forestille os tre forskellige, hypotetiske estimatorer på grisenes helbredstilstand

a. Den perfekte estimator hvis vi kunne se præcis hvordan grisene har det
b. Vores ekspert(er)s bedømmelse af staldgulvsbilledet
c. Computer vision modellens bedømmelse af staldgulvsbilledet

Den bedste estimator ville selvfølgelig være a., og dernæst b. og så c. Ved at arbejde videre med computer vision modellen gør vi sådan at vores estimator c. bliver mere lig med estimator b. Men hvad hvis der er stor forskel på estimator a. og estimator b.? Hvis det er tilfældet, kan de forbedringer vi laver på computer vision modellen jo være stort set ligegyldige da det kun bringer os marginalt tættere på den perfekte estimator a.

Men hvad er forskellen på det som staldgulvsbilledet viser, og grisenes faktiske helbredstilstand? Først og fremmest er de fæces, som er synlige på et billede, kun et snapshot af grisenes helbredstilstand i fortiden. En oplagt model for at modellere sammenhængen kunne være at man forestiller sig at staldgulvsbilledet viser et sandfærdigt men uendeligt lille udsnit af det samlede antal fæces grisene kunne have lavet. Matematisk betyder det, at antallene af fæces med hhv. score 1, 2 og 4 på staldgulvsbilledet alle er poisson fordelt med en middelværdi, der svarer til grisenes egentlige helbredstilstand. Computersimulationer af denne hypotetiske poisson-model viser at den gennemsnitlige forskel mellem staldgulvsbilledernes fæcesscoregennemsnit og det sande fæcesscoregennemsnit er 0.21 (R^2 gennemsnitligt på 0.83 og spearmans rho på 0.91). Det skal så sammenlignes med forskellen mellem de annoterede og prædikterede fæces som var på 0.24.

Med andre ord er den hypotetiske forskel mellem estimator a. og b. 0.21, mens afstanden mellem estimator b. og c. er 0.24. Det betyder at afstanden mellem estimator b. og c. er ca

$$\sqrt{\frac{2}{\pi}}\sqrt{\left(\frac{0.21}{\sqrt{\frac{2}{\pi}}}\right)^2+\left(\frac{0.24}{\sqrt{\frac{2}{\pi}}}\right)^2}\approx 0.32$$

Dermed kan yderligere forbedringer af modellen kun bringe den gennemsnitlige afvigelse fra 0.32 ned til 0.21.

Man kan dog godt argumentere imod denne teoretiske nedre grænse på 0.21 ved at sige at alle grisenes fæces bliver lagt i det afbildede område og at det altså ikke er tilfældigt om man ser 5 eller 6 kategori 4 fæces. På den anden side ved man ikke hvad der sker lige efter billedet bliver taget. Derfor er sandheden nok at det er tilfældigt om man lige ser 5 eller 6 af en bestemt type fæces, men ikke så tilfældigt som hvis billedet kun dækkede en lille del af det samlede fæcesområde.

Konklusion

Vi har udviklet en computer vision model, der kan estimere den gennemsnitlige fæcesscore på en skala fra 1 til 4 med en gennemsnitlig fejl på 0.24, når vi tester det op mod vores testsæt. Man kan diskutere om det er muligt at udvikle en computer vision model, der er mange gange bedre, i og med at billedet i sig selv kun er et estimat af virkeligheden. For at forbedre modellen kan man lave et annoteret træningssæt med mindre varians ved at inkludere flere bedømmelser, evt. af forskellige eksperter. Modellen viser en acceptabel performance på billeder taget med et andet kamera af lavere opløsning, men dens performance kunne forbedres ved at inkludere dens billeder i træningssættet også.

TEKNOLOGISK
INSTITUT