





## Article

# Impact Evaluation of Score Classes and Annotation Regions in Deep Learning-Based Dairy Cow Body Condition Prediction

Sára Ágnes Nagy <sup>1</sup>, Oz Kilim <sup>2</sup>, István Csabai <sup>2</sup>, György Gábor <sup>3</sup> and Norbert Solymosi <sup>1,2,\*</sup><sup>1</sup> Centre for Bioinformatics, University of Veterinary Medicine, 1078 Budapest, Hungary<sup>2</sup> Department of Physics of Complex Systems, Eötvös Loránd University, 1117 Budapest, Hungary<sup>3</sup> Androvet Ltd., 1182 Budapest, Hungary

\* Correspondence: solymosi.norbert@gmail.com; Tel.: +36-30-9347-069

**Simple Summary:** The body condition of dairy cattle is an essential indicator of the energy supply of the animals. Various scoring systems are used in practice to quantify body condition. These systems rely on visual observation of different body parts and sometimes on collecting tactile data. In all cases, scoring requires expert knowledge and practice and is time-consuming. Therefore, it is rarely carried out on livestock farms. However, for animal husbandry and veterinary practice, it would be meaningful to have data on the condition of the animals continuously or even daily, which is not feasible with expert scoring. We investigated how computer vision-based supervised deep learning, specifically neural networks, can automate body condition scoring. To execute this, we have used video recordings of the rumps of a large number of animals. We have trained and tested various convolutional neural networks with this collected data. Scoring by trained networks yielded results that met or exceeded the agreement among experts. We have made our trained neural networks freely available, using these as pretrained models. Those working on similar developments can achieve even better results with less data collection required with their own fine-tuning.

**Abstract:** Body condition scoring is a simple method to estimate the energy supply of dairy cattle. Our study aims to investigate the accuracy with which supervised machine learning, specifically a deep convolutional neural network (CNN), can be used to retrieve body condition score (BCS) classes estimated by an expert. We recorded images of animals' rumps in three large-scale farms using a simple action camera. The images were annotated with classes and three different-sized bounding boxes by an expert. A CNN pretrained model was fine-tuned on 12 and 3 BCS classes. Training in 12 classes with a 0 error range, the Cohen's kappa value yielded minimal agreement between the model predictions and ground truth. Allowing an error range of 0.25, we obtained minimum or weak agreement. With an error range of 0.5, we had strong or almost perfect agreement. The kappa values for the approach trained on three classes show that we can classify all animals into BCS categories with at least moderate agreement. Furthermore, CNNs trained on 3 BCS classes showed a remarkably higher proportion of strong agreement than those trained in 12 classes. The prediction precision when training with various annotation region sizes showed no meaningful differences. The weights of our trained CNNs are freely available, supporting similar works.

**Keywords:** deep learning; dairy cow; body score; prediction; accuracy



**Citation:** Nagy, S.Á.; Kilim, O.; Csabai, I.; Gábor, G.; Solymosi, N. Impact Evaluation of Score Classes and Annotation Regions in Deep Learning-Based Dairy Cow Body Condition Prediction. *Animals* **2023**, *13*, 194. <https://doi.org/10.3390/ani13020194>

Academic Editor: Andrea Pezzuolo

Received: 31 October 2022

Revised: 19 December 2022

Accepted: 29 December 2022

Published: 4 January 2023



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Body condition scoring of cattle is a widespread, noninvasive and easy-to-use but subjective and time-consuming method for estimating an animal's saturation of subcutaneous fat deposits [1,2]. The different scoring systems infer the animal's energy supply from the coverage of the lumbar, pelvic and tail head regions [2]. The accumulation of body stores is quantified using a numerical scale, where lean individuals have low values and overweight individuals have high values [3]. The saturation of fat and energy stores

can provide important guidance for farm management. Several studies have shown how a shift in body condition score (BCS) away from what is ideal is associated with changes in production [4]. Condition scores that are too high or low may indicate several health issues or mismanagement [4]. The association between a high condition score and the risk of ketosis is well known [4,5]. Furthermore, it is also associated with other metabolic problems (e.g., fatty liver) [6] or placenta retention [5]. Too low a condition score can be associated with lameness and reduced milk production [4]. Several pathological processes correlate with a decrease in BCS (e.g., metritis, inactive ovaries, displaced abomasum or more days open) [4]. Since pathological changes are often closely related to changes in BCS [4] rather than a specific condition score, it is understandable that continuous and reliable herd-level condition scoring would be an essential aid to dairy herd management [7–11]. This is hampered by the fact that scoring requires trained staff [2], and herd-level scoring is time-consuming. The volatility of intraobserver and interobserver agreements makes the BCS data generated challenging to use [12]. Mullis et al. [13] showed that the agreement between two experts' BCS values is moderate, while Song et al. [14] found that inter- and intraobserver agreement is weak and moderate, respectively.

Our study aims to investigate the accuracy with which supervised machine learning, specifically deep convolutional neural network (CNN)-based Detectron2 models, can be used to recover the BCS classes estimated by an expert using images of cows taken with a simple RGB camera. As a first approach, we investigate the quality of the predictions of the CNNs trained on our 12-level BCS scoring. In the following approach, we investigate the quality of the predictions of CNNs trained on three BCS classes corresponding to four different target intervals: T1 for calving (days in milk (DIM): 0) with dry period (DIM from –60 to –1) and dry off period (DIM > 300 and DIM < –60), T2 for early (DIM: 1–30) and mid-lactation (DIM: 101–200), T3 for peak lactation (DIM: 31–100) and T4 for late lactation (DIM: 201–300). Furthermore, we study how different region of interest (ROI) rectangles of the rump cause variation in the predictive power of our models.

## 2. Materials and Methods

### 2.1. Data Collection

Digital video recordings were taken over the course of 2 years with an SJCAM 4000 RGB camera at three large-scale dairy cattle farms in Hungary (farm F1: 1150 cows; F2: 880 cows; F3: 960 cows), with the camera positioned in the rotary milking parlor pointing at the animals' rumps. Not all of the milking cows were filmed in each video. To avoid overrepresentation of any cow with a given body scoring in the data set, we skipped at least a month between any two videos taken at a single site to ensure the conditions of the animals had changed.

### 2.2. Data Preprocessing

The recorded videos were annotated later by an expert remotely using the Visual Object Tagging Tool (VoTT, v2.2.0) [15], assigning both a bounding rectangle to the animal's rump as well as an estimated BCS. Scoring was performed in the range of 1–5. In this range, 12 levels were defined: 1–2.5 and 4–5 were divided into 0.5 score intervals, while 2.5–4.00 was divided more finely with 0.25 score intervals [2]. The scoring was performed continuously with movie footage, drawing the bounding boxes and making the BCS inference from only the image captured where the rump was at the closest point to the camera. These were the final images used to build the training, validation and test sets. After annotating the videos produced in the study, the same expert rechecked and adjusted the annotations with Label Studio [16]. The F1 and F2 farm images were split into the training and validation sets. This was performed with stratification of the BCS scores. By randomly selecting 80% of the images within each score level for the train set, we created the validation set with the remaining pictures. This ensured that the training and validation sets contained the same class distributions, thereby making the validation set loss an appropriate metric for model choice decisions. The annotated images from site

F3 were retained as an independent test set. The number of images per score in each set is summarized in Table 1. Three size bounding boxes were generated for each image automatically from the initial annotations (See Supplementary Material: Automated 3-box size annotation).

**Table 1.** The number of annotated images included in the study per score. Images from sites F1 and F2 were used to create the training and validation sets, while the held-out test set only consisted of images from the F3 site.

BCS	F1 & F2 Farm		F3 Farm
	Training	Validation	Test
1.00	162	41	22
1.50	215	54	31
2.00	363	91	53
2.50	370	93	65
2.75	182	46	42
3.00	323	81	41
3.25	381	95	50
3.50	659	165	66
3.75	74	18	9
4.00	58	14	11
4.50	46	12	6
5.00	89	22	11
Total:	2922	732	407

### 2.3. Choice of Model Architecture

Body scoring is an object detection and classification problem. For this reason, we chose an object detection model that could leverage both bounding box and class ground truth annotations. The Faster R-CNN architecture [17] has a shared model internal representation for the joint localization and classification prediction tasks. This joint learning task is apparent by inspecting the form of the loss function used for training  $\ell = \ell_{cls} + \ell_{bbox}$ . Due to the state-of-the-art results, the Detectron2 [18] implementation of the Faster R-CNN architecture was chosen. All models were downloaded from the Model Zoo code repository with network parameters pretrained on the COCO Dataset.

### 2.4. Evaluation Metrics

The performance of our model's localization could be reviewed in terms of the bounding box prediction average precision at IoU = 0.50 (AP50). This value increases when the predicted bounding boxes overlap more with the ground truth annotation bounding boxes. In addition to detecting an object in the image, Detectron2 estimates a class probability distribution over all classes. As the final classification, it assigns the object to the class for which the probability is the highest among all possible classes. The quality of the predictions was quantified using Cohen's kappa [19] and accuracy. Kappa is defined as  $kappa = (P_0 - P_e) / (1 - P_e)$ , where  $P_0$  is the observed agreement between the ground truth and predicted classes and  $P_e$  is the probability change agreement between the model prediction and annotation ground truth. Cohen's kappa values can be interpreted as follows: 0–0.20 = no, 0.21–0.39 = minimal, 0.40–0.59 = weak, 0.60–0.79 = moderate, 0.80–0.90 = strong and above 0.90 = almost perfect agreement [20]. Following the “one-versus-all” scheme, the accuracy was calculated by comparing each class to the remaining levels with the formula  $accuracy = (TP + TN) / (TP + TN + FP + FN)$ , and from these results, the overall accuracy was reported as the mean.

### 2.5. Model Screening

To pre-screen for the most appropriate pretrained model for the BCS task, 10 pretrained models of Detectron2 [18] were further trained and validated on our data sets, each with

identical hyperparameters for 15 epochs. In this model selection phase, the raw BCS scores were classified into three classes (BCS classes: <2.5, between 2.5 and 3.75 and >3.75). The R\_50\_FPN\_3x model gave the lowest validation loss, so we selected and used this pretrained model in all further experiments (Table 2).

**Table 2.** Model selection. Ten pre-trained models of Detectron2 were run on the same data set with the same settings (number of epochs: 15). The model with the lowest validation loss was chosen for all further experiments.

Pretrained Faster R-CNN Model	Validation Loss
R_50_FPN_3x	0.0612
R_101_FPN_3x	0.0628
R_50_FPN_1x	0.0637
X_101_32x8d_FPN_3x	0.0662
R_50_DC5_1x	0.0796
R_50_DC5_3x	0.0840
R_101_C4_3x	0.0848
R_101_DC5_3x	0.0848
R_50_C4_1x	0.1019
R_50_C4_3x	0.1040

## 2.6. Model Training and Prediction

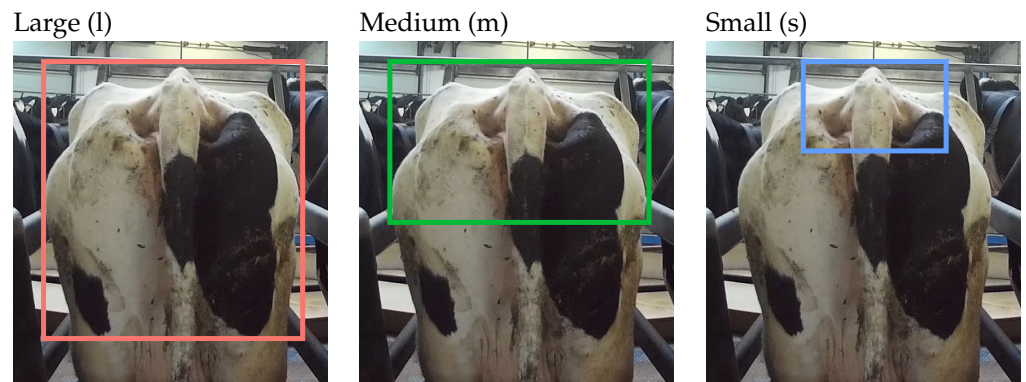
### 2.6.1. With 12 BCS Classes

Using the selected pretrained model, the first experiment was run by training, validating and testing of the 12-level ordinal scoring annotations. This was repeated for each of the three bounding box sizes (see Figure 1) separately. Using the model's optimal weights at checkpoints with the lowest validation losses and AP50s, we made predictions for the validation and test sets. The class where the probability distribution (output by the model) had its maximum was taken as the estimated body score. We also kept the "predicted class probabilities", as they are a measure of model confidence for a given prediction.

Following the approach of Yukun et al. (2019) [21], we evaluated the model predictions with 0, 0.25 and 0.5 error thresholds to account for the ordinal nature of the body condition scores and thereby allowed "near miss" predictions to be defined as correct, reducing metric stringency.

### 2.6.2. With Three BCS Classes for Four Practical Target Intervals

In addition to the 12-point BCS annotation, we also assessed the quality of the predictions according to more broad BCS classes of practical relevance. Different BCS ranges are considered optimal at different stages of lactation. These target intervals are summarized in Table 3. We relabeled all the original BCSs according to four different threshold regime (T1, T2, T3 and T4) target ranges: T1, the optimal target interval for calving (days in milk (DIM): 0) and the dry period (DIM from -60 to -1) and dry off period (DIM > 300 and DIM: < -60); T2 for early (DIM: 1-30) and mid-lactation (DIM: 101-200); T3 for peak lactation (DIM: 31-100) and T4 for late lactation (DIM: 201-300). We created three new BCS classes for each type of thresholding: below the target interval, the target interval and above the target interval. The training, validating and testing sessions were re-run with these three new classes for each of the four thresholding regimes and for each of the three bounding box sizes. Each experiment used the same input images, where the label definitions varied between each. In Table 3, the linked figure shows the proportion of images after reclassification in the training, validation and test sets. The proportion of classes was the same in the training and validation set (farm F1 and F2) and the test (farm F3) set. This allowed for fair testing, as the class imbalance was the same in training and validation and test sets.

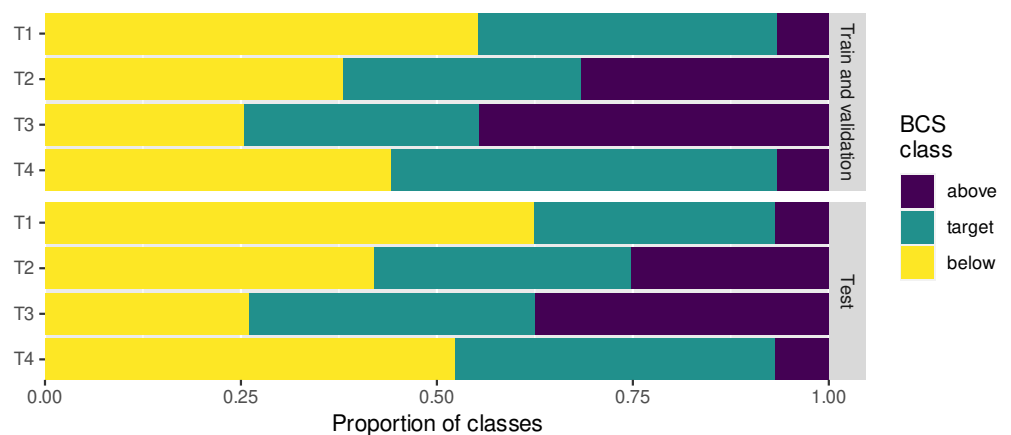


**Figure 1.** Annotation rectangles. The large (l) box was placed across the width of the two *tuber coxae* from the *anterior coccygeal vertebrae* (tail head) to mid-thigh. The medium (m) box was the entire width between the *coxal tuberosities* from the tail head to the *symphysis pelvis*. The small (s) box only framed the *ischial tuberosities* and the tail head, containing the depression between the *tuber ischii* and the *anterior coccygeal vertebrae*, and the area between the *tuber coxae* and the *sacral spinous processes*.

To compare the training process between training on 3 classes and training on 12 classes, we repeated the analysis in a way where 12 classes were predicted, but they were then reclassified to the new 3 classes at the inference time (see Figure S1).

**Table 3.** Body condition score target intervals: T1 for calving (DIM: 0) and the dry period (DIM from −60 to −1), and dry off period (DIM > 300 and DIM < −60); T2 for early (DIM: 1–30) and mid-lactation (DIM: 101–200), T3 for peak lactation (DIM: 31–100) and T4 for late lactation (DIM: 201–300). Below, the relabeling for each interval is shown. For each regime, the original ordinal labels were relabeled according to the given threshold of that regime. For example, with the T1 thresholding, there were more images with cows in the “below” class, whereas if we relabeled the data with the T2 thresholding, then the three classes were more evenly split. Class distributions in the training and validation sets match respective test sets.

Mark	Target BCS Interval	
	Min	Max
T1	3.25	3.75
T2	2.75	3.25
T3	2.50	3.00
T4	3.00	3.75

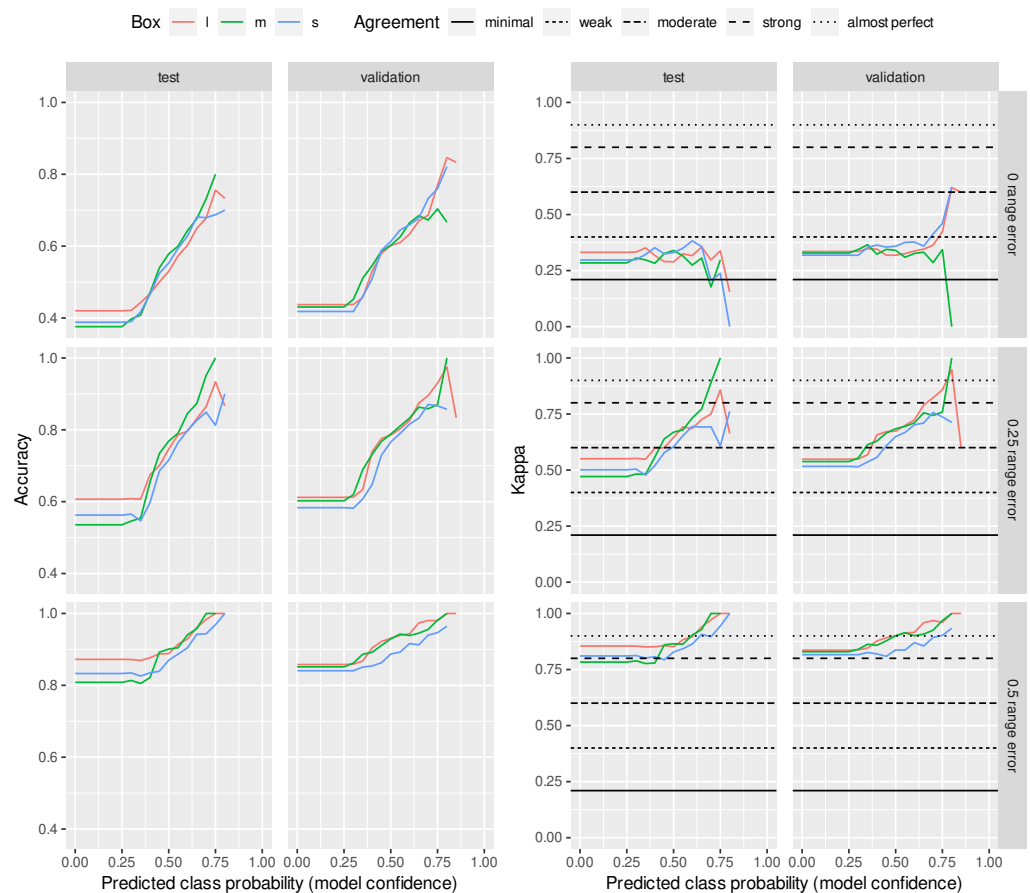


### 3. Results

#### 3.1. Prediction with 12 BCS Classes

The quality of the predictions for the 12 classes, based on the training with 12 classes, is shown in Figure 2. The  $x$  axis represents the results thresholded by predicted class probabilities. As the threshold increased, only images classified with high model confidence were used to calculate kappa and the accuracy. Traversing the  $x$  axis of each plot gives an idea of how the data clustered near the learned decision boundary in high dimensional space. Images near the decision boundary had a low predicted class probability (model confidence). Images with high confidence were to be classified better.

When training and evaluating in 12 classes and allowing for an error range of 0, the kappa value on the test set yielded minimal agreement but worse agreement above a class prediction probability of about 75%. For the validation set, the agreement was similar to below 75% class prediction probability. In contrast, above this level, the agreement was weak for the l and s boxes and even weaker than the minimum for the m box. If we allowed an error range of 0.25, then we obtained minimum agreement below the 50% class prediction probability on both the test and the validation set, and above that, the maps fell into the weak agreement range. When allowing an error range of 0.5, we had curves running above or close to the strong cut point with a class prediction probability of about 60% on both the test and validation sets. However, from 65% to 70%, we found almost perfect agreement.



**Figure 2.** Model trained and evaluated with 12 BCS classes. Prediction confidence values (accuracy and Cohen’s kappa) were estimated on test and validation sets as a function of predicted class probability. Three error ranges (0, 0.25 and 0.5) were allowed for agreement analysis of the expert-given and predicted scores. The horizontal lines represent the thresholds of McHugh [20] to interpret Cohen’s kappa values.



### 3.2. Prediction with Three BCS Classes

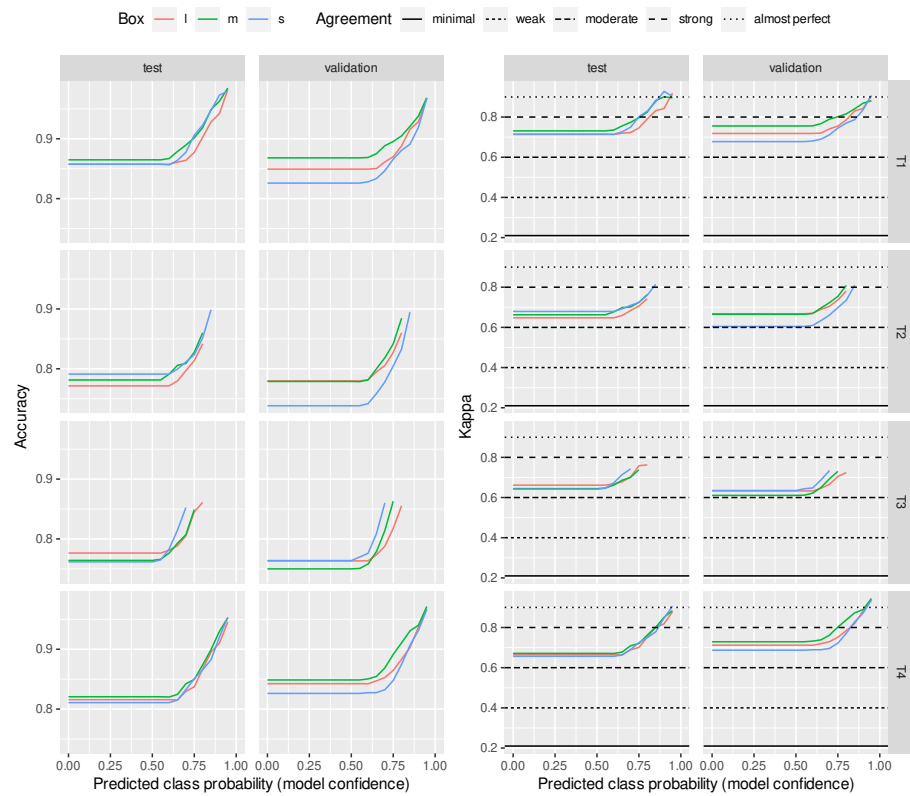
Figure 3 shows the quality of the predictions made on only 3 BCS classes instead of 12 BCS classes. The three classes were created according to the four target ranges (T1–T4) presented in Table 3. The expert's scores assigned to the images were reassigned to one of the three BCS classes and then used to perform the training, with the prediction also being made for three classes. In this three-class regime, the network made high-quality predictions. This type of classifier may be more relevant for farms to give lower-resolution but reliable predictions and act more as a screening tool. If animals are predicted to be out of the target, then they can be further investigated.

In a second “control” approach, we trained the network on 12 BCS classes and returned 12 classes with the predictors as in Figure 2. We then reclassified the 12 classes into the 3 BCS classes. This was also repeated for each T1–4 regime. These results can be seen in Figure S1. The traversal of model confidence thresholding showed more noise. This is understandable, as the body scoring classes are ordinal but not continuous [14]. To further compare the two approaches, we examined the difference in the kappa and accuracy values for the joint class classification probabilities. For this, we subtracted the values of the trained set on 12 classes from the values of the trained set on 3 classes. In Figure S2a, the mean and the standard deviation of the differences are plotted as outlined in the Supplementary Materials. The methods showed similar performances.

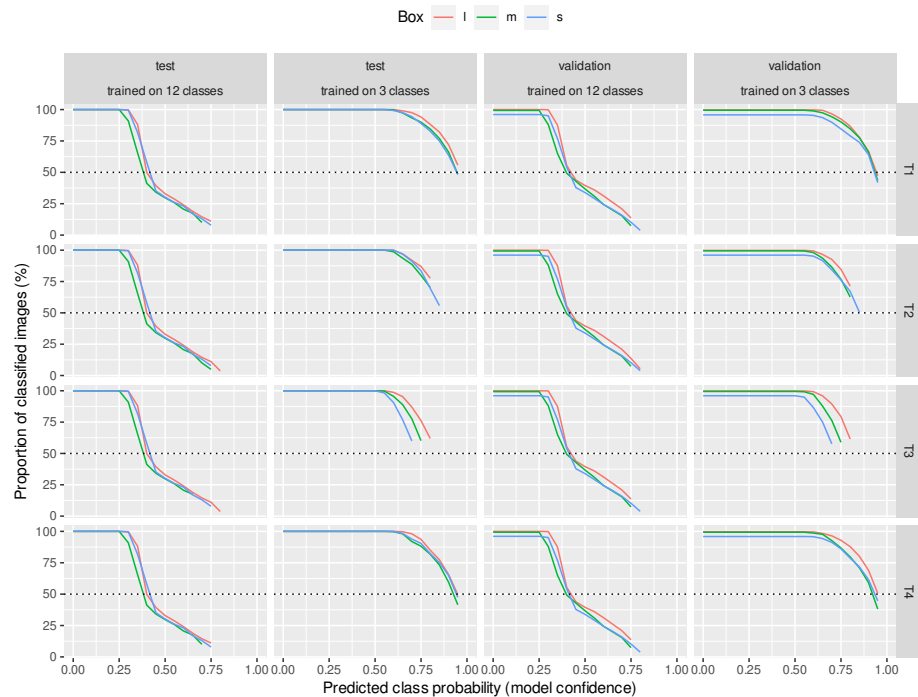
The differences between the prediction goodness of annotation boxes obtained from the networks taught by the three classes are summarized in Figure S2b. We may observe the same trends for accuracy and kappa for all target ranges in the box prediction differences. It can also be seen that the differences between the boxes were an order of magnitude smaller than those obtained from the networks taught by classes of 12 and 3. Based on the test set, the prediction precision per box in descending order for the target ranges was as follows: T1: m, s, l; T2: s, m, l; T3: l, s, m; T4: m, l, s. In the validation set, the prediction precision per box in descending order for the target ranges was as follows: T1: m, l, s; T2: m, l, s; T3: s, l, m; T4: m, l, s.

As we traversed the  $x$  axis for each result, we had fewer and fewer predicted images to evaluate the metric. Figure 4 summarizes the proportion of the initial (test or validation) data set corresponding to each predicted class probability value. For the neural networks trained on 12 classes, at a predicted class probability value of 40%, half of the images were already dropped. However, for the networks taught by 3 classes, even at the highest predicted class probability value, half of the images or slightly less than half of the images were still part of the analysis. The three-class classification could be considered more of an “easy” task for the network, so higher model confidence for the predicted classes was reasonable. We found variation between the results of the three box sizes to not be very large, indicating that the supervised signal was mainly found within the medium box area.

Regardless of the target range (T1–T4), the kappa values of the approach trained and evaluated in the three classes showed that we could classify all animals into BCS categories with at least moderate agreement. Figure 5 shows the proportion of predictions based on training for different target ranges with different box sizes that resulted in at least strong agreement.

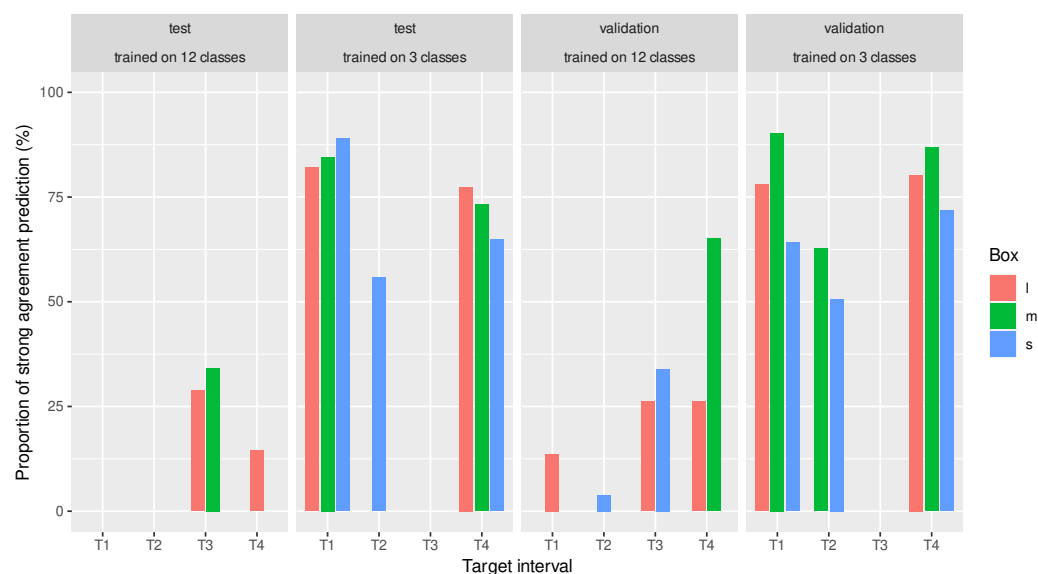


**Figure 3.** Model trained and evaluated with three BCS classes. Prediction confidence values (accuracy and Cohen’s kappa) were estimated on test and validation sets as a function of predicted class probability. The horizontal lines represent the thresholds of McHugh [20] to interpret Cohen’s kappa values.



**Figure 4.** Proportion of classified images. As the threshold cutoff is more stringent, the number of images left gets smaller. The reduction occurred to a larger extent for the 12 class experiments. This gives an insight into the distribution of images concerning the learned decision boundary in the model representation.





**Figure 5.** Proportion of predictions with strong agreement (Cohen's kappa  $\geq 0.8$ ). For predictions over different target ranges (T1–T4). CNNs trained in three BCS classes showed a remarkably higher proportion of strong agreement than those trained in 12 classes and then reclassified.

#### 4. Discussion

Several possible approaches to estimating cattle body conditions using neural networks exist [21–23]. The more common approach in the literature is to use recordings of animals scored with high-resolution scoring, such as in 0.25 or 0.5 unit increments, to train the neural network and to test how the trained network performs in terms of prediction reliability at the same scale. Assume no discrepancy is allowed between the observed and predicted scores. In that case, these approaches yield weak agreement, as indicated by the  $kappa = 0.45$  value of Yukun et al. [21]. If we allowed some variation between the observed and predicted scores, both the kappa and accuracy values improved. As the first step in our investigation, we followed the approach, and similar results were obtained around the reliability values presented by other authors [21,22,24–28].

However, in addition to these high-resolution score classes, we felt it was worthwhile to investigate the prediction quality that could be achieved for practically important condition score classes. We conducted this investigation in two forms. In one, we trained a neural network on the high-resolution, detailed 12-point level data set, and then both the predictions and original expert scores were classified into three condition categories. Thus, there were a target range category (T1, T2, T3 and T4) and categories below and above the target range. In the second approach, we trained the neural network with the data set already divided into three categories and made predictions corresponding to the three classes. The test and validation set's predictions showed that the latter approach gave better results with lower noise. However, after splitting into three categories, the networks trained on the 12 categories also gave good results, but these results were noisier and could be evaluated to be a less robust approach.

We also considered the probability that the neural network assigned a BCS class to the object detected in each image. The kappa and accuracy values presented were evaluated as a function of this class assignment probability. It can be seen that the precision of the predictions improved with an increasing class assignment probability. Nevertheless, as the class ranking probability increased, fewer images could be considered in evaluating the reliability of the predictions. When comparing the networks trained on 12 classes and 3 classes, it can be seen that the class assignment probabilities were lower for the former than for the latter. For this reason, as we increased the threshold of the classification probability of the images included in the prediction precision analysis, the number of usable images predicted by the neural network trained on the 12 classes decreased rapidly, while for the

networks trained on 3 classes, the number of usable images decreased much more slowly as the classification probability threshold increased. Even at the highest threshold, roughly half of the images were retained. This approach has not been found in the literature, where the prediction precision was analyzed in conjunction with the classification probability. However, the results show that it significantly affects the prediction quality and thus the practical efficacy of body condition prediction based on neural networks. Nevertheless, it is still problematic that the number of images with a higher classification probability was less than the number of animals in the set used for prediction. The results show that at the most reliable classification probability threshold, we lost half of the animals, which means we obtained reliable conditioning information for half of the animals on a given day. However, we aim to obtain daily information on all individuals in a given herd. We see several possibilities to address this problem, which further studies could clarify. One approach could be based on the fact that animals are snapped not only once but several times during milking in the carousel systems. We could identify the highest probability class from their class distribution if we predicted each of these. Nevertheless, we can conclude from Figure 5 that the prediction of CNNs trained in class 3 showed a significant proportion of strong agreement which was better than the inter-rater agreement found in the literature.

The use of practical thresholds in the 3-grade approach trained in 12 classes and assessed in 3 classes is problematic. Several authors showed that high prediction reliability can be obtained in error ranges of 0.25 or 0.5 [21,22,24–28]. However, when we tried to assign this to the practical condition categories we used, it was impossible for many to decide which practical condition interval to place an animal in with an error range of 0.25 or 0.5. It is important to emphasize that our study aimed to investigate the reliability with which a neural network can reproduce the scores and score categories of animals scored by an expert. It is also possible to construct a ground of truth from the scores of several experts rather than one. However, it is worth considering that the agreement between the scores of two independent experts is weak or moderate based on the literature. The results reported by Mullis et al. [13] show that Cohen's kappa of the agreement between two experts' BCS values is 0.62 and 0.66, while Song et al. [14] found that an inter-assessor agreement kappa = 0.48, while the intra-assessor agreement kappa is 0.52 and 0.72. Thus, the prediction precision of a neural network built on this basis could easily be worse, not better.

In our study, we used a test set from an independent site in addition to the validation set to see the robustness of the neural network prediction. After all, it was expected that if animals from the same farm were given the training and validation sets, the prediction for the validation set would be better than the predictions for an utterly independent farm. Surprisingly, the predictions of the networks trained on the three classes differed little for the test and validation sets. A further interesting feature of our results is that the prediction precision based on the three types of annotation boxes (large, medium and small) also differed very little. We chose three different-sized annotation boxes because we might have thought a large box contained more information than the algorithm could capture. Conversely, it seems that a medium-sized box contained the most usable information, which still had little noise. Next in line was the small box, which gave slightly better results than the large box. Here, we can think of it as containing less information and less noise. In contrast, the big one had more noise with more information. Thus, the order was that the middle one came after the small box, followed by the large box in terms of prediction performance.

In our work, we deliberately did not use complementary tactile examinations in the generation of expert scores because when teaching a neural network based purely on images, this information is not available to the algorithm, so the expert's information in scoring is richer than what we can offer the neural network.

The results show that the quality of training and prediction from two-dimensional images taken with a simple sports camera using Detectron2 is not inferior to the prediction

results based on three-dimensional cameras or on scoring with tactile detection [29,30]. An additional option to consider to improve the prediction quality could be to use ensemble prediction of different trained networks as the final output.

The results for the weights generated for the CNNs are publicly available. Others can use this as a pretrained model for training neural networks on similar images. Thus, presumably, they can create their neural networks while using fewer images to predict BCS categories. The presented results suggest that similar outcomes can be expected in condition scoring for other breeds and types of utilization. However, this assumes that similar CNNs should be trained on data sets generated by a scoring system applied to given breeds and types of utilization. Thus, the trained algorithm presented here cannot be used one-to-one for other breeds and utilization types.

## 5. Conclusions

Our results conclude that CNN training on classes corresponding to practically relevant target ranges gives more robust and precise predictions than training on high-resolution classes. With predictions based on target interval training, we obtained similar or even better results than the agreement between experts. The prediction precision based on training with various annotation regions showed no meaningful differences.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ani13020194/s1>, Figures S1 and S2. Automated 3 box size annotation, 12 class training and testing after re-labeling test sets into T1-4 classes.

**Author Contributions:** N.S. takes responsibility for the integrity of the data and the accuracy of the data analysis. N.S., S.Á.N. and G.G. conceived the concept of the study. S.Á.N. performed the data annotation. N.S. and S.Á.N. participated in the model training, predictions and statistical analysis. N.S., S.Á.N. and O.K. participated in the drafting of the manuscript. N.S., S.Á.N., O.K., I.C. and G.G. carried out the critical revision of the manuscript for important intellectual content. All authors read and approved the final manuscript.

**Funding:** This study was supported by European Union project RRF-2.3.1-21-2022-00004 within the framework of the MILAB Artificial Intelligence National Laboratory.

**Institutional Review Board Statement:** At no point in the study did the data collection affect the welfare of the cows, as the video recordings had no effect on the animals.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The weights used for the predictions from training with each BCS class and annotation box combination can be downloaded from <https://doi.org/10.6084/m9.figshare.21372000.v1> (accessed on 3 January 2023).

**Acknowledgments:** We thank Alex Olár for his suggestions to help us in our work.

**Conflicts of Interest:** The authors declare no conflict of interest. The funder had no role in the study design; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Edmonson, A.; Lean, I.; Weaver, L.; Farver, T.; Webster, G. A body condition scoring chart for Holstein dairy cows. *J. Dairy Sci.* **1989**, *72*, 68–78. [[CrossRef](#)]
2. Ferguson, J.D.; Galligan, D.T.; Thomsen, N. Principal descriptors of body condition score in Holstein cows. *J. Dairy Sci.* **1994**, *77*, 2695–2703. [[CrossRef](#)]
3. Roche, J.; Dillon, P.; Stockdale, C.; Baumgard, L.; VanBaale, M. Relationships among international body condition scoring systems. *J. Dairy Sci.* **2004**, *87*, 3076–3079. [[CrossRef](#)] [[PubMed](#)]
4. Bewley, J.; Schutz, M. An interdisciplinary review of body condition scoring for dairy cattle. *Prof. Anim. Sci.* **2008**, *24*, 507–529. [[CrossRef](#)]
5. Morrow, D.A. Fat cow syndrome. *J. Dairy Sci.* **1976**, *59*, 1625–1629. [[CrossRef](#)]
6. Roche, J.R.; Kay, J.K.; Friggens, N.C.; Looor, J.J.; Berry, D.P. Assessing and managing body condition score for the prevention of metabolic disease in dairy cows. *Vet. Clin. Food Anim. Pract.* **2013**, *29*, 323–336. [[CrossRef](#)]

7. Silva, S.R.; Araujo, J.P.; Guedes, C.; Silva, F.; Almeida, M.; Cerqueira, J.L. Precision technologies to address dairy cattle welfare: Focus on lameness, mastitis and body condition. *Animals* **2021**, *11*, 2253. [CrossRef] [PubMed]
8. Albornoz, R.I.; Giri, K.; Hannah, M.C.; Wales, W.J. An improved approach to automated measurement of body condition score in dairy cows using a three-dimensional camera system. *Animals* **2022**, *12*, 72. [CrossRef] [PubMed]
9. Tao, Y.; Li, F.; Sun, Y. Development and implementation of a training dataset to ensure clear boundary value of body condition score classification of dairy cows in automatic system. *Livest. Sci.* **2022**, *259*, 104901. [CrossRef]
10. Truman, C.M.; Campler, M.R.; Costa, J.H. Body condition score change throughout lactation utilizing an automated BCS system: A descriptive study. *Animals* **2022**, *12*, 601. [CrossRef]
11. Zhao, K.; Zhang, M.; Shen, W.; Liu, X.; Ji, J.; Dai, B.; Zhang, R. Automatic body condition scoring for dairy cows based on efficient net and convex hull features of point clouds. *Comput. Electron. Agric.* **2023**, *205*, 107588. [CrossRef]
12. Kristensen, E.; Dueholm, L.; Vink, D.; Andersen, J.; Jakobsen, E.; Illum-Nielsen, S.; Petersen, F.; Enevoldsen, C. Within-and across-person uniformity of body condition scoring in Danish Holstein cattle. *J. Dairy Sci.* **2006**, *89*, 3721–3728. [CrossRef] [PubMed]
13. Mullins, I.L.; Truman, C.M.; Campler, M.R.; Bewley, J.M.; Costa, J.H. Validation of a commercial automated body condition scoring system on a commercial dairy farm. *Animals* **2019**, *9*, 287. [CrossRef] [PubMed]
14. Song, X.; Bokkers, E.; Van Mourik, S.; Koerkamp, P.G.; Van Der Tol, P. Automated body condition scoring of dairy cows using 3-dimensional feature extraction from multiple body regions. *J. Dairy Sci.* **2019**, *102*, 4294–4308. [CrossRef] [PubMed]
15. Visual Object Tagging Tool (VoTT). 2020 Available online: <https://github.com/microsoft/VoTT> (accessed on 3 January 2023).
16. Tkachenko, M.; Malyuk, M.; Holmanyuk, A.; Liubimov, N. Label Studio: Data labeling software, 2020–2022. Open source software. Available online: <https://github.com/heartexlabs/label-studio> (accessed on 3 January 2023).
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 1, Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; NIPS’15, pp. 91–99.
18. Yuxin Wu and Alexander Kirillov and Francisco Massa and Wan-Yen Lo and Ross Girshick. Detectron2. 2019. Available online: <https://github.com/facebookresearch/detectron2> (accessed on 3 January 2023).
19. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
20. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Medica* **2012**, *22*, 276–282. [CrossRef]
21. Yukun, S.; Pengju, H.; Yujie, W.; Ziqi, C.; Yang, L.; Baisheng, D.; Runze, L.; Yonggen, Z. Automatic monitoring system for individual dairy cows based on a deep learning framework that provides identification via body parts and estimation of body condition score. *J. Dairy Sci.* **2019**, *102*, 10140–10151. [CrossRef]
22. Alvarez, J.R.; Arroqui, M.; Mangudo, P.; Toloza, J.; Jatip, D.; Rodríguez, J.M.; Teyseyre, A.; Sanz, C.; Zunino, A.; Machado, C.; et al. Body condition estimation on cows from depth images using Convolutional Neural Networks. *Comput. Electron. Agric.* **2018**, *155*, 12–22. [CrossRef]
23. Çevik, K.K. Deep Learning Based Real-Time Body Condition Score Classification System. *IEEE Access* **2020**, *8*, 213950–213957. [CrossRef]
24. Krukowski, M. *Automatic Determination of Body Condition Score of Dairy Cows from 3D Images*; Skolan för datavetenskap och kommunikation; Kungliga Tekniska Högskolan: Stockholm, Sweden, 2009.
25. Bercovich, A.; Edan, Y.; Alchanatis, V.; Moallem, U.; Parmet, Y.; Honig, H.; Maltz, E.; Antler, A.; Halachmi, I. Development of an automatic cow body condition scoring using body shape signature and Fourier descriptors. *J. Dairy Sci.* **2013**, *96*, 8047–8059. [CrossRef]
26. Anglart, D. *Automatic Estimation of Body Weight and Body Condition Score in Dairy Cows Using 3D Imaging Technique*; SLU, Department of Animal Nutrition and Management: Uppsala, Sweden, 2014.
27. Shelley, A.N. *Incorporating Machine Vision in Precision Dairy Farming Technologies*; University of Kentucky: Lexington, KY, USA, 2016.
28. Spoliansky, R.; Edan, Y.; Parmet, Y.; Halachmi, I. Development of automatic body condition scoring using a low-cost 3-dimensional Kinect camera. *J. Dairy Sci.* **2016**, *99*, 7714–7725. [CrossRef] [PubMed]
29. Shigeta, M.; Ike, R.; Takemura, H.; Ohwada, H. Automatic measurement and determination of body condition score of cows based on 3D images using CNN. *J. Robot. Mechatronics* **2018**, *30*, 206–213. [CrossRef]
30. Yu, J.; Yang, B.; Wang, J.; Leader, J.K.; Wilson, D.O.; Pu, J. 2D CNN versus 3D CNN for false-positive reduction in lung cancer screening. *J. Med. Imaging* **2020**, *7*, 051202. [CrossRef] [PubMed]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.