Udderly Fit: Efficient Transfer Learning for Dairy Cattle Body Condition Scoring from Tailhead Photos

Paul Creavin
Department of Statistics
Stanford University
Stanford, CA
pcreavin@stanford.edu

Quentin MacFarlane
Department of Electrical Engineering
Stanford University
Stanford, CA
qmac3@stanford.edu

Ishan Seendripu
Department of Electrical Engineering
Stanford University
Stanford, CA
ishanks2@stanford.edu

I. Motivation

I (Paul) grew up on a dairy farm in rural Ireland, where limited veterinary access meant herd health often depended on early visual judgement. Small decisions, like spotting weight loss, had major impacts on welfare, fertility, and productivity. Body Condition Scoring (BCS) is central to managing cow health but remains subjective and inconsistent, especially when time or expertise is limited. This motivates our work on automated, accessible BCS estimation.

We develop a deep learning system that predicts BCS from tailhead photographs. Starting with EfficientNet-B0, we evaluate multiple fine-tuning strategies and will expand to modern convolutional and transformer backbones, alongside ensemble methods for robustness. Our aim is a lightweight, interpretable, and deployable model that supports consistent on-farm scoring, enhancing, not replacing, farmers' expertise.

II. Related Work

Recent work has shown that convolutional neural networks can achieve strong agreement with expert annotators when predicting BCS from rump-based imagery [Nagy et al.(2023)]. This line of research highlights that evaluation strategy and class grouping substantially affect predictive performance. However, Nagy et al. report that varying the size of the rump ROI does not significantly change model accuracy. Detection-based pipelines have also been explored using YOLO-style architectures, demonstrating reliable multi-class BCS classification across breeds and production environments [Dandıl et al.(2024)], though these systems typically rely on high-capacity detectors and full-body or back-view imaging.

Multi-view and rear-view imaging approaches further improve fine-grained BCS detection, with back-view imagery capturing the most informative morphological cues [Lewis et al.(2025)]. Such methods, however, often assume controlled camera setups or continuous monitoring infrastructure. In contrast, our work focuses on a single tailhead view captured under unconstrained farm conditions, using lightweight transfer learning and explainability techniques.

III. Dataset

We use a publicly available dairy cow BCS dataset collected from large-scale farms and annotated by professional veterinarians [Huang et al.(2025)]. The dataset reflects real-world variability in lighting, posture, and background, making it suitable for practical deployment studies. It contains 53,566 JPEG images with Pascal VOC bounding-box annotations around the tailhead region. Each cow was scored by experts into five standard classes (3.25, 3.5, 3.75, 4.0, 4.25). Figure 1 shows representative examples from each BCS class, highlighting the substantial variation in lighting, posture, and viewpoint that the model must handle when learning such fine-grained distinctions. The overall class distribution is shown in Figure 2, which exhibits a moderate imbalance toward mid-range BCS values.

Fig. 1: Representative full-frame images from each BCS class (3.25, 3.5, 3.75, 4.0, 4.25) with annotated tailhead bounding boxes.

IV. Method

A. Data Pipeline

Image file paths and corresponding BCS labels are loaded from a CSV file. When available, Pascal VOC XML annotations are parsed to crop the tailhead region-of-interest (ROI), the anatomical region most associated with visual BCS assessment. If no annotation is present, the full image is used. Images are resized to 224×224 pixels via bilinear interpolation and normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).

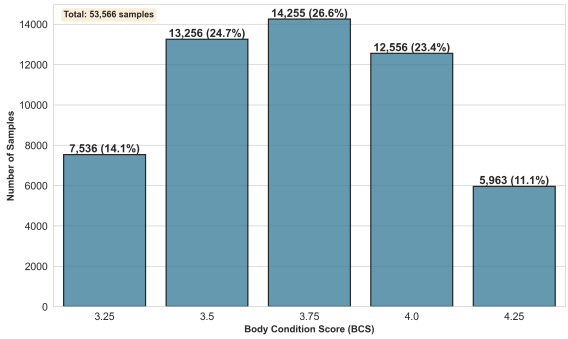

Fig. 2: Distribution of body condition score (BCS) classes in the dataset.

B. Data Splits

We use a stratified $70/15/15$ split to preserve class balance across training, validation, and test sets. The resulting split contains 37,496 training images, 8,035 validation images, and 8,035 test images. A fixed random seed of 42 ensures reproducibility. The validation set is used exclusively for model selection and early stopping, while the test set is reserved for final model evaluation.

C. Model Architecture and Transfer Learning

Models are instantiated via a unified `timm`-based model factory, with EfficientNet-B0 chosen for its strong accuracy–efficiency trade-off. We evaluate four transfer-learning strategies: (i) *Head-only* (freeze backbone; train classifier), (ii) *Last-block* (unfreeze blocks 6–7), (iii) *Full* (fine-tune all layers from ImageNet weights), and (iv) *Scratch* (train all layers from random initialization). All models predict five BCS classes (3.25–4.25), and trainable parameter counts are recorded for each configuration.

D. Training Procedure

Training is configured through YAML files for reproducibility. Unless otherwise stated, we use AdamW (lr 3×10^{-4}, weight decay 10^{-4}), cross-entropy loss with label smoothing ($\epsilon = 0.05$), early stopping with patience 5, and fixed random seeds for Python, NumPy, and PyTorch. The best model is selected by validation macro-F1.

E. Evaluation Metrics

We report overall accuracy, macro-F1, weighted F1, per-class precision/recall, underweight recall (BCS 3.25), and confusion-matrix heatmaps.

V. Experiments and Results

A. Transfer Learning Ablation (EfficientNet-B0)

We conducted preliminary experiments to establish baselines before further tuning. As a classical reference, we first used EfficientNet-B0 as a frozen feature extractor and trained a multinomial logistic regression model on its embeddings. We then trained the four transfer-learning configurations described in Section IV-C using identical data splits and optimisation settings to enable a fair comparison.

TABLE I: Preliminary validation results.

Mode	Acc	Macro-F1	W-F1	UW-Rec
Logistic Reg.	0.504	0.510	0.505	0.498
Head-only	0.407	0.410	0.407	0.436
Last-block	0.722	0.728	0.722	0.730
Full	0.923	0.925	0.923	0.916
Scratch	0.698	0.704	0.696	0.672

These results show that full fine-tuning clearly performs best, exceeding 0.92 accuracy and macro-F1. Partial unfreezing (last-block) also performs well, indicating that pretrained features remain valuable with limited adaptation. In contrast, head-only and from-scratch training generalise poorly, and the logistic baseline demonstrates that ImageNet features alone are not sufficient for fine-grained BCS classification.

B. Enhanced Baseline: Data Augmentation and Cosine Annealing

After confirming full fine-tuning as the strongest baseline, we introduced two enhancements to improve generalisation and convergence.

a) Data Augmentation: Using Albumentations, each training image had a 50% chance of a horizontal flip and a 50% chance of a random rotation within $\pm 10°$. These mild transformations increased viewpoint diversity without distorting BCS cues, reducing overfitting compared with fixed 224×224 crops.

b) Cosine Scheduler: A cosine annealing schedule decayed the learning rate smoothly from 3×10^{-4} to zero across 20 epochs, promoting stable convergence and reducing late-epoch oscillations.

Together, these enhancements increased validation macro-F1 from roughly 0.94 to 0.9621, with balanced per-class recall (0.95–0.97) and underweight recall improving to 0.9513.

C. ROI Cropping, Jitter, and Robustness to Context

To ensure the model learns robust anatomical cues rather than overfitting to tightly aligned bounding boxes, we introduce controlled variation around the annotated tailhead region. Figure 3 illustrates the resulting crops.

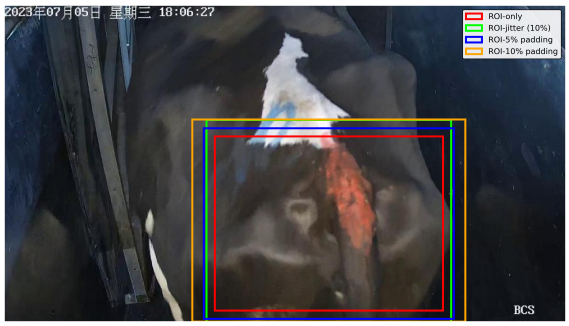

Fig. 3: Illustration of the different cropping strategies applied to the tailhead region.

First, we apply mild ROI jitter during training, randomly shifting and rescaling the bounding box by up to 10%. This simulates natural localisation noise from an upstream detector and encourages the classifier to remain stable under small annotation errors. Jitter is used only during training. Validation and test crops remain fixed.

Second, we evaluate fixed padding of 5% and 10% around the ROI to test whether the model relies on background cues. As shown in Table II, both padding levels maintain performance comparable to the ROI-only baseline, indicating limited sensitivity to modest background variation and confirming that the model focuses on the intended anatomical region.

TABLE II: ROI cropping and robustness to context.

Configuration	Accuracy	Macro-F1	UW Recall
ROI-only	96.13%	96.21%	95.13%
ROI + jitter	96.25%	96.37%	94.51%
ROI + 5% padding	96.42%	96.52%	95.49%
ROI + 10% padding	96.44%	96.50%	95.22%

D. Input Resolution and Model Scaling

To assess whether increased spatial resolution or model capacity improves performance, we conducted a sequence of experiments varying both input size (224 px vs. 320 px) and EfficientNet depth (B0–B3). All models used the enhanced training setup from Section V-B, with ROI cropping applied to the B0 variants.

TABLE III: Model size and resolution experiments (validation set).

Model / Resolution	Acc	Macro-F1	UW-Rec
B0 @ 224px	**0.9625**	**0.9637**	0.9451
B0 @ 320px	0.9598	0.9611	0.9407
B1 @ 320px	0.9450	0.9470	0.9372
B2 @ 320px	0.9493	0.9509	0.9593
B3 @ 320px	0.9540	0.9552	0.9549

Across these settings, EfficientNet-B0 at 224 px consistently performed best. Increasing resolution to 320 px did not yield improvements for B0, and larger models (B1–B3), tested at 320 px due to their higher capacity, failed to match the performance of the lighter B0 backbone. These results suggest that the ROI-based B0 @ 224 px model offers the best balance of accuracy and efficiency, and we adopt it as the primary architecture for all subsequent experiments.

E. Ordinal Regression Experiments

Because body condition scores form an ordered scale, we explored CORAL-style ordinal regression as an alternative to multiclass classification. Unlike a classifier that treats classes independently, the ordinal approach models BCS as a sequence of ordered thresholds, encouraging the network to learn the monotonic structure of the scoring scale.

The unweighted ordinal model achieved lower exact accuracy than the classifier but showed strong ordinal behaviour, with over 99% of predictions within ± 1 class of the ground truth. This indicates that the ordinal head captures the smooth progression between neighbouring BCS values even when exact matches are weaker.

To improve sensitivity to the underweight class, we increased the weight on the first cumulative threshold, which boosted underweight recall with minimal effect on accuracy or macro-F1. We also compared threshold-count and expected-value decoding strategies and found threshold-count gave the best overall performance.

Finally, we evaluated a simple classifier–ordinal ensemble that relied on the ordinal model only when the two predictions differed by more than one class. Although competitive, the ensemble provided only marginal gains and added unnecessary complexity. Given our focus on robustness and efficiency, we retain the simpler classification model for final evaluation.

Approach	Accuracy	Macro-F1	Underweight Recall	Ord. Acc. ($\pm$1)
Class. (baseline)	96.25%	96.37%	94.51%	99.08%
Ordinal (unweighted)	94.09%	94.32%	91.50%	99.24%
Ordinal (weighted)	94.16%	94.53%	93.10%	99.18%
Class.–ordinal ensemble	96.09%	96.22%	94.07%	98.97%

TABLE IV: Classification vs. ordinal regression approaches.

F. Further Ensemble Approaches

Building on the ordinal–classification hybrid introduced in Section V-E, we evaluate broader ensemble configurations here, including mixtures of multiple EfficientNet backbones and the ordinal model.

1) Multi-Model Ensembles: We evaluated ensembles with 3, 4, and 5 models using majority voting.

- **3-way ensemble:** EfficientNet-B0, B1, and B2 classification models
- **4-way ensemble:** EfficientNet-B0, B1, B2 classification models + ordinal regression model
- **5-way ensemble:** EfficientNet-B0, B1, B2, B3 classification models + ordinal regression model

Table V shows the performance of multi-model ensembles on the validation set:

Ensemble	Accuracy	Macro-F1	Underweight Recall
Best Individual (B0)	96.25%	96.37%	94.51%
3-way	92.83%	93.11%	94.07%
4-way	96.28%	96.45%	95.13%
5-way	95.49%	95.72%	94.25%

TABLE V: Performance of Multi-Model Ensembles (Validation Set)

The 4-way ensemble performed best among the multi-model variants, reaching 96.28% accuracy and 96.45% macro-F1, only slightly above the single B0 model. Both the 3-way and 5-way ensembles showed diminished returns, with the 3-way version performing worse due to the inclusion of weaker models such as B1.

Agreement statistics reflected this variation: the 4-way ensemble achieved 57.8% full agreement and the 5-way ensemble 52.9%, indicating more diverse predictions as additional models were added.

2) Multi-Scale Ensemble Results: A weighted ensemble of EfficientNet-B0 and B2 reached 90.33% accuracy and 90.13% macro-F1, improving on B2 but still below B0. The performance gap between the two models was too large for the ensemble to provide meaningful gains.

3) Conclusion: Overall, ensembles offered only marginal improvements over the EfficientNet-B0 baseline, and the added complexity and size make them less suitable for practical on-farm deployment.

G. Alternative Architectures

In addition to the EfficientNet family, we evaluated several modern lightweight convolutional and transformer-based architectures to assess whether larger or more recent backbones offer improved performance on fine-grained tailhead BCS classification. Specifically, we tested ConvNeXt-Tiny, Swin-Tiny, and RegNetY-8GF, each fine-tuned end-to-end under the same training setup as our EfficientNet models. Table VI summarises the results. RegNetY-8GF achieved the strongest performance among the three, reaching 93.63% accuracy and a macro-F1 of 0.9385, though still below our EfficientNet-B0 baseline. ConvNeXt-Tiny and Swin-Tiny underperformed relative to all EfficientNet variants, suggesting that the dataset's fine-grained morphological cues favour lighter, highly optimised CNN backbones over heavier transformer-style models in this setting.

Model	Accuracy	Macro-F1	Underweight Recall
Swin-Tiny	0.8326	0.8389	0.8071
ConvNeXt-Tiny	0.8877	0.8913	0.8876
RegNetY-8GF	0.9363	0.9385	0.9195

TABLE VI: Performance of alternative backbone architectures on the validation set.

VI. Final Evaluation on Test Data

A. Final Model Selection

For the final test evaluation, we selected the ROI-based EfficientNet-B0 model with 5% padding. This configuration achieved the highest macro-F1 among all single-model variants, offering strong performance across the full BCS range while remaining lightweight at 224×224 resolution. The added context improved robustness to minor ROI misalignment without increasing computational cost, making the model well suited for practical on-farm deployment.

B. Test-Set Performance

The selected ROI-based EfficientNet-B0 model achieves strong performance on the held-out test set, reaching 95.7% accuracy and a macro-F1 score of 0.9591. As shown in Table VII and Figure 4, per-class precision and recall remain consistently high across the full BCS range, with the clinically important underweight class (3.25) attaining 0.95 recall. Overall, these results indicate that the model generalises well and provides reliable, stable predictions suitable for practical on-farm deployment.

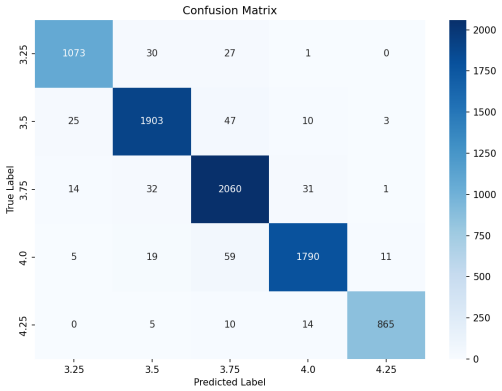

Fig. 4: Test-set confusion matrix for the selected ROI-based model, showing per-class BCS predictions.

TABLE VII: Test Set Per-Class Performance Metrics

BCS Class	Recall	Precision	F1-Score
3.25	0.9487	0.9606	0.9546
3.5	0.9572	0.9568	0.9570
3.75	0.9635	0.9351	0.9491
4.0	0.9501	0.9697	0.9598
4.25	0.9676	0.9830	0.9752

C. Error Analysis

To understand the failure modes of our model, we conducted a comprehensive error analysis.

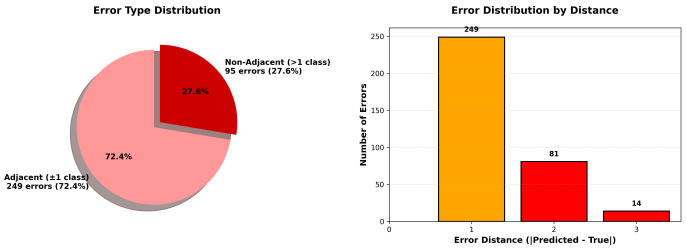

Fig. 5: Adjacent vs. non-adjacent error analysis. **Left:** Pie chart showing error distribution. **Right:** Distribution of errors by distance.

Figure 5 presents a critical finding: **72.4% of all misclassifications are adjacent-class errors**, meaning the model's prediction differs from the true label by only one class (± 0.25 BCS points). This is highly desirable for ordinal classification tasks, as it indicates the model has learned the inherent ordering of BCS values. Furthermore, 95.9% of errors are within ± 0.50 BCS and no errors are more than 3 classes away.

Figure 6 shows the distribution of model confidence for correct versus incorrect predictions. Correct predictions exhibit a strong right-skewed distribution with most predictions having confidence >0.8, while incorrect predictions show lower confidence scores centered around 0.5–0.7. This separation suggests the model is well-calibrated because it is more confident when correct and less confident when wrong. When deployed on the farm, the low confidence errors can be reviewed by humans.

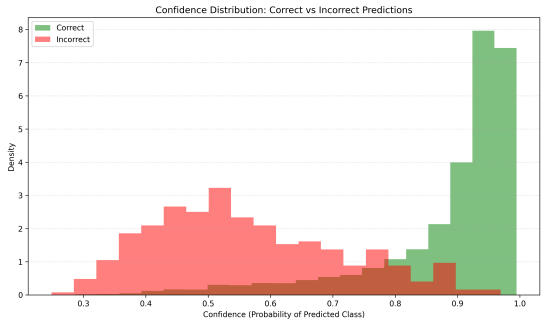

Fig. 6: Confidence distribution for correct vs. incorrect predictions.

Qualitative inspection of several high-confidence misclassifications further suggested that many errors arise from visually ambiguous, low-quality images, consistent with the inherent subjectivity of manual BCS labeling.

D. Model Interpretability via Grad-CAM

To understand what the model attends to when predicting BCS, we use Grad-CAM visualizations. This method highlights the image regions most influential for the model's decisions, offering interpretability for an otherwise black-box network. In our setting, these visualizations help confirm that the model focuses on anatomically relevant tailhead features rather than background artifacts.

a) Correct Predictions Analysis: The correctly predicted samples show that the model consistently attends to the anatomical structures that define tailhead body condition, particularly the ridge, fat deposits, and surrounding musculature. Across a range of BCS values, the Grad-CAM overlays remain focused on the central tailhead region with little activation on irrelevant background areas. Even under variations in lighting, coat pattern, or minor occlusion, localisation remains stable, indicating that the model is using meaningful physiological cues rather than relying on shortcuts.

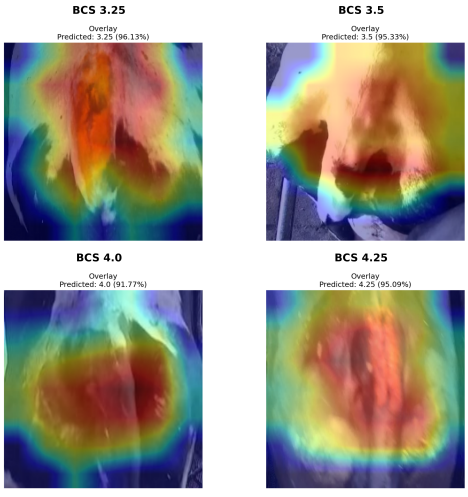

Fig. 7: Grad-CAM overlays for correctly predicted BCS samples, illustrating that the model focuses on anatomically relevant tailhead features.

b) Incorrect Predictions Analysis: We also examined several misclassified samples to understand where the model struggles (Figure 8). Sometimes mud or objects block the view of the camera, making the tailhead difficult to identify by the model. Other cases show unstable heatmaps when texture contrast is low, or overly loose crops that cause background regions to be treated as part of the cow. In high-angle shots, the network attends to the animal's back rather than the tailhead.

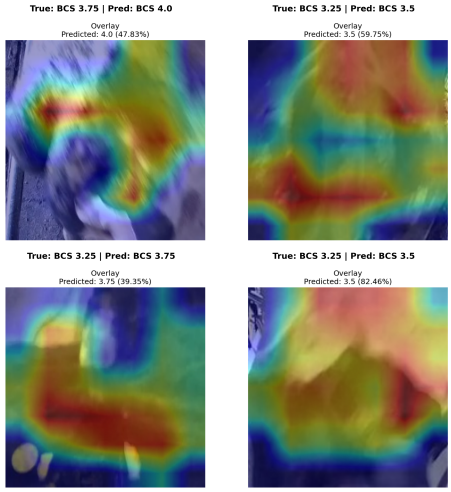

Fig. 8: Grad-CAM overlays from four misclassified samples highlighting errors caused by occlusion, weak localisation, loose cropping, and high camera angle.

VII. Conclusion and Future Work

A. Conclusion

This work shows that fine-grained body condition scoring from tailhead images is feasible using a lightweight EfficientNet-B0 model suitable for on-farm use. Full fine-tuning with an ROI-based pipeline delivered strong performance across all BCS classes, including the underweight category. Robustness tests confirmed stability under crop variation, and Grad-CAM visualisations showed anatomically meaningful attention. On the held-out test set, accuracy and macro-F1 remained high, with most errors confined to adjacent classes, consistent with the subjectivity of manual scoring. These results demonstrate a practical, low-cost approach to automated BCS estimation and provide a basis for broader deployment.

B. Future Work

Several directions remain for extending this system. Broader training across farms, breeds, and lighting conditions would improve generalisation, and evaluating transfer to younger animals is an open question. Mobile and edge deployment could be enabled through pruning, quantisation, or distillation. Adding a lightweight tailhead detector would allow end-to-end operation without external bounding boxes. Finally, testing under extreme poses, occlusion, mud, and challenging viewpoints would help characterise failure modes and guide future data collection.

VIII. Team Contributions

- **Paul Creavin:** Implemented the data pipeline, ROI-cropping system, training configuration files, EfficientNet models, and all associated experiments. Ran the full set of ablations, model evaluations, and test-set analyses, and wrote the main body of the report.
- **Ishan Seendripu:** Implemented the alternative architecture backbones and the logistic regression baseline. Contributed supporting code for model configuration and training utilities.
- **Quentin MacFarlane:** Implemented the ensemble experiments and the test-set error analysis. Wrote the corresponding ensemble and error-analysis sections of the report.

References

[Dandıl et al.(2024)] Emre Dandıl, Kerim Kürşat Çevik, and Mustafa Boğa. 2024. Automated Classification System Based on YOLO Architecture for Body Condition Score in Dairy Cows. *Veterinary Sciences* 11, 9 (2024). doi:10.3390/vetsci11090399

[Huang et al.(2025)] Xiaoping Huang, Zihao Dou, Fei Huang, Huanyu Zheng, Xiankun Hou, Chenyang Wang, Tao Feng, and Yuan Rao. 2025. Dairy cow body condition score target detection data set. doi:10.57760/sciencedb.16704 Accessed: 2025-10-05.

[Lewis et al.(2025)] Reagan Lewis, Teun Kostermans, Jan Wilhelm Brovold, Talha Laique, and Marko Ocepek. 2025. Automated Body Condition Scoring in Dairy Cows Using 2D Imaging and Deep Learning. *AgriEngineering* 7, 7 (2025). doi:10.3390/agriengineering7070241

[Nagy et al.(2023)] Sára Ágnes Nagy, Oz Kilim, István Csabai, György Gábor, and Norbert Solymosi. 2023. Impact Evaluation of Score Classes and Annotation Regions in Deep Learning-Based Dairy Cow Body Condition Prediction. *Animals* 13, 2 (2023), 194. doi:10.3390/ani13020194