From Pixels to Pregnancies: AI-Driven Oocyte Grading for Scalable Livestock Breeding

Parkash Singh
The Ohio State University
Computer Science & Engineering
singh.2068@osu.edu

Grace Koppelman
The Ohio State University
Food, Agricultural & Biological Engineering
koppelman.9@osu.edu

Tanya Berger-Wolf
The Ohio State University
Computer Science & Engineering
berger-wolf.1@osu.edu

John Fulton
The Ohio State University
Food, Agricultural & Biological Engineering
fulton.20@osu.edu

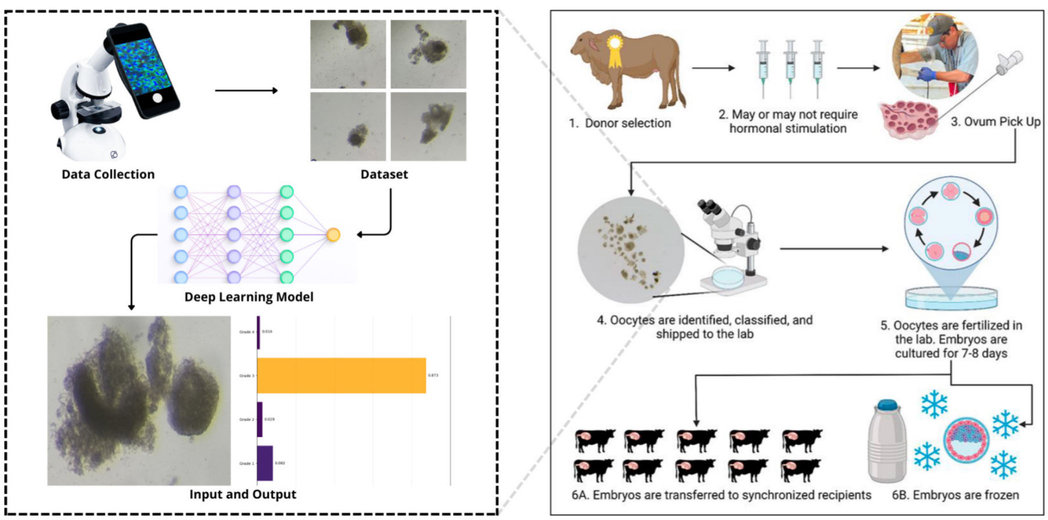

Figure 1: Overview of the in vitro embryo production (IVP) workflow (right, adapted from Gonella-Diaza, 2023 [noa, b]) and AI-based oocyte grading pipeline (left). The right panel shows each stage of IVP, with Oocyte grading highlighted as a key step. Oocyte grading is a critical step where our deep learning model integrates into IVP to enable objective and robust classification.

Abstract

Sustainable livestock breeding is essential to meeting the food demands of a growing global population. Assisted reproductive technologies (ART), such as in vitro fertilization(IVF), are increasingly used to enhance reproductive efficiency. A key determinant of IVF success is the quality of the oocyte, which directly affects fertilization, embryo development, and blastocyst yield. However, oocyte grading today remains a subjective and inconsistent process, creating variability that affects the entire in vitro embryo production pipeline. We introduce a deep learning framework for automated oocyte grading, built on the first dataset of its kind: 1,140 bovine cumulus–oocyte complex (COC) images labeled according to the International Embryo Transfer Society (IETS) scale. Our models achieve up to 65%

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

accuracy across four grades, improving to over 80% when grouped into industry relevant quality categories. By aligning with IETS guidelines, we establish the first benchmark for standardized oocyte grading in livestock IVF. This work provides a strong foundation for AI-based assisted livestock breeding, offering consistency, reduced human variability, and increased throughput in livestock breeding.

1 Introduction

Feeding a projected population of nearly 10 billion people by 2050 [Nations] requires more than incremental improvements in agriculture and livestock farming. Livestock contributes a major share of the global food supply through meat and dairy, yet their productivity remains limited by the pace of natural breeding. To meet this demand, Assisted reproductive technologies (ART) are being utilized over traditional natural breeding methods to enhance both reproductive efficiency and genetic selection [Ferré et al., 2020]. Among ART procedures, in vitro fertilization (IVF) plays an important role [Hansen, 2023]. IVF is a process where oocytes (eggs) are collected from the ovaries and fertilized by sperm in a laboratory setting.

The success of IVF depends heavily on selecting high quality oocytes, as oocyte quality directly influences fertilization rates, embryo development, and pregnancy outcomes [Demetrio et al., 2022]. Currently, grading is performed manually by experienced technicians who assess subtle morphological traits under a microscope. This process is slow, requires years of training, and is inherently subjective, leading to inconsistency across observers and laboratories [Fjeldstad et al., 2024, Farin et al., 1995]. The variability in oocyte grading reduces reproducibility and throughput, creating a bottleneck that limits the scalability and success of IVF in livestock breeding [Boni, 2018]. These are challenges that AI can uniquely address through automation and standardization of oocyte evaluation.

While AI based methods have been applied to human IVF, they lack publicly available datasets and benchmarks due to the proprietary nature of clinical data. Livestock applications face the same challenge, with no open datasets or standardized evaluation protocols [Iannone et al., 2024]. Additionally, within livestock, grading practices also vary: many farms and laboratories collapse oocytes into two or three categories, while the International Embryo Transfer Society (IETS) recommends four Boni [2018]. This lack of standardization underscores the need for reproducible, open, and standarized solutions.

We present the first large scale dataset and benchmark for bovine oocyte grading aligned with IETS guidelines. This work is based on the graduate Master thesis work by the co-author Grace Koppleman [Koppelman, 2025], who collected the data, framed the problem, and the initial evaluation of the computational solution. The dataset contains 1,140 expert annotated images of bovine cumulus–oocyte complexes (COCs) commonly referred as oocytes in this paper, and we evaluate both object detection and classification models. Baseline results show competitive performance on 4-class IETS grading and improved accuracy under industry relevant three class and binary schemes. As a step toward automating IVF workflows in livestock farming, our framework addresses the foundational task of oocyte grading, reducing grading time, minimizing human variability, and providing high clinical value for IVF practice. This work lays the groundwork for scalable, reproducible AI benchmarks that support both academic research and industry adoption.

2 Related Work

AI has shown promise in assisted reproductive technologies (ART), particularly for oocyte and embryo assessment. Most work has focused on human IVF, where deep learning pipelines have achieved high accuracy in classifying oocytes by meiotic stage and predicting fertilization potential. For instance, [Targosz et al., 2023] combined DeepLabV3Plus segmentation with a refined SqueezeNet classifier, reaching 96% validation accuracy for human oocyte classification. While encouraging, these studies remain constrained by small, imbalanced datasets and limited reproducibility, as clinical data are rarely shared.

In livestock, research has largely targeted blastocyst prediction rather than direct oocyte grading. Costa et al. (2023) proposed a semi automatic CNN based system for bovine oocyte competence, labeling images post hoc by blastocyst outcomes. More recently, [Raes et al., 2025] reported that

Table 1: Number of cumulus–oocyte complex (COC) images across IETS grading categories (Grades 1–4).

COC Grades	Count
Grade 1	258
Grade 2	307
Grade 3	316
Grade 4	259
Total	1140

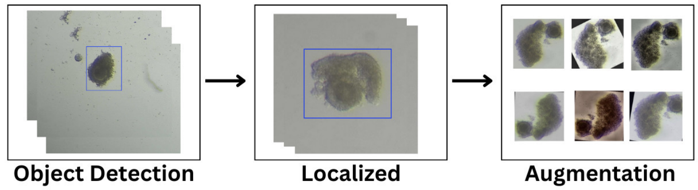

Figure 2: Preprocessing pipeline for oocyte images. Object detection identifies the region of interest; localized crops are extracted, and augmentations are applied to improve training diversity. We ablate the effects of detection and augmentation are evaluated in Table 4.

neural networks and random forest models outperformed embryologists in predicting blastocyst potential from COC images, with balanced accuracy >70% compared to <45% for humans, which validates the potential of ML models in this field. However, blastocyst formation depends on many factors beyond oocyte morphology, making such predictions an unreliable proxy for oocyte quality.

Despite widespread use of morphological grading in practice, few AI studies address oocyte grading directly, and none provide standardized, open benchmarks for livestock IVF. Existing approaches lack alignment with International Embryo Transfer Society (IETS) guidelines and vary across farms and labs [Boni, 2018]. Our work addresses this gap by introducing the first large scale dataset and benchmark for bovine oocyte grading, comprising 1,140 expert annotated oocyte images labeled according to IETS standards. This benchmark reframes oocyte grading as a machine learning task, enabling reproducibility, standardization, and practical integration into livestock breeding.

3 Methodology

3.1 Data Collection

Bovine cumulus–oocyte complexes (COCs) were collected from slaughterhouse cattle and transferred to the laboratory for imaging and labeling. Two acquisition setups were used: (i) Leica MC120 HD camera mounted on a Leica M80 microscope. (ii) iPhone 15 Pro Max attached to the Leica M80 microscope using a custom adapter.

Each image contained a single COC centered in the frame, captured under consistent magnification and lighting to minimize variability. Images were labeled by experienced technicians into one of four categories (Grades 1–4) following the International Embryo Transfer Society (IETS) guidelines, which define morphological criteria for oocyte quality. In total, 1,140 COC images were collected (see table 1). To our knowledge, this represents the first dataset of bovine COCs labeled into four categories according to IETS guidelines, providing a foundation for reproducible machine learning benchmarks in livestock IVF.

3.2 Data Preprocessing

To ensure consistent inputs, we first localized oocytes using a lightweight YOLO based (Yollos)[Ultralytics] object detector. The motivation was that raw images frequently contained off center oocytes along with extraneous cells and background noise, which could bias the downstream classification task. The detector was trained on 50 manually annotated images and achieved mAP@0.5 of 0.51 on 10 val images. Once trained, it was applied to the full dataset to crop regions of interest (ROIs), producing standardized oocyte-centered images.

Over cropped images, we further applied data augmentation on dataset for robustness. Augmentations included (i) geometric: flips, small rotations, and scale/shift; (ii) photometric: brightness/contrast and color shifts; and (iii) imaging artifacts: Gaussian blur and local contrast enhancement (CLAHE) [Marimuthu, 2022]. Augmentations were applied probabilistically to the training set only, ensuring realistic variability while preserving morphology.

Table 2: Comparison of DINOv2+KNN, YOLOv8, and InceptionV3 on 4-class IETS grading. YOLOv8 achieves higher mean accuracy and stronger recall/F1 across most grades, with notable gains on Grade 1.

	DINOv2+KNN				YOLOv8				InceptionV3			
Grade	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**
1	0.62	0.63	0.62	0.63	0.64	0.84	0.86	0.84	0.67	0.63	0.65	0.63
2	0.52	0.67	0.58	0.67	0.62	0.45	0.52	0.45	0.44	0.43	0.44	0.43
3	0.55	0.45	0.49	0.45	0.58	0.63	0.60	0.62	0.52	0.60	0.55	0.60
4	0.84	0.71	0.77	0.71	0.69	0.64	0.67	0.64	0.82	0.74	0.78	0.74
Mean	0.63	0.61	0.61	**0.61**	0.63	0.64	0.66	**0.63**	0.61	0.60	0.61	**0.59**

Table 3: Generalization on test and independent set which is outside main data corpus. Δ_{Gen} is the change from Test to Independent accuracy for each model.

	DINOv2+KNN			YOLOv8			InceptionV3		
Grade	**Test**	**Indep.**	$\boldsymbol{\Delta}_{\text{Gen}}$	**Test**	**Indep.**	$\boldsymbol{\Delta}_{\text{Gen}}$	**Test**	**Indep.**	$\boldsymbol{\Delta}_{\text{Gen}}$
1	0.63	0.10	**-0.53**↓	0.84	0.30	**-0.54**↓	0.63	0.30	**-0.33**↓
2	0.67	0.70	**+0.03**↑	0.45	0.90	**+0.45**↑	0.43	0.50	**+0.07**↑
3	0.45	0.40	**-0.05**↓	0.62	0.40	**-0.22**↓	0.60	0.50	**-0.10**↓
4	0.71	0.67	**-0.04**↓	0.64	0.55	**-0.09**↓	0.74	0.45	**-0.29**↓
Mean	**0.61**	**0.46**	–	**0.63**	**0.53**	–	**0.59**	**0.43**	–

3.3 Modeling

We evaluated both transfer learning and end-to-end fine tuning for oocyte image classification after preprocessing. Following CNN and Transformer based pipelines were tested on 1,140 images (70/15/15 split):

DINOv2 with KNN. Cropped oocyte images were embedded using DINOv2 (ViT-B/14) [Oquab et al., 2023], a self supervised Vision Transformer, and classified with a K-Nearest Neighbors (KNN) model. This pipeline achieved 61% test accuracy on the 4-class IETS grading task. When categories were grouped into binary or three-class schemes, performance improved to 70–80%, highlighting the difficulty of fine-grained classification.

YOLOv8 classifier. Using YOLO based cropping, we fine tuned YOLOv8[noa, c] as a 4-class classifier. YOLOv8 achieved 63% test accuracy, outperforming DINOv2+KNN and generalizing more robustly across conditions. Training used AdamW with cross entropy loss, cosine learning rate decay, and ran for 100 epochs on a single GPU.

Inception V3. Using YOLO-based cropping, we fine-tuned InceptionV3[noa, a] as a 4-class classifier by unfreezing the last 50 layers. InceptionV3 achieved 59% test accuracy on the held-out test set. Training used Adam optimizer, categorical crossentropy loss, dropout, and L2 regularization. The model was trained for up to 100 epochs with early stopping and learning rate reduction on plateau, running on a single GPU.

Summary. YOLO based cropping proved critical for improving input quality. YOLOv8 delivered the strongest baseline for 4-class IETS grading, while grouped classification achieved higher accuracy. This suggests a tradeoff between fine grained precision and practical deployment.

3.4 Error Analysis

We compared the classification performance of DINOv2+KNN, YOLOv8, and IncpetionV3 on the 4-class IETS grading task (Table 2). A consistent trend is that models separate extreme classes (Grade 1 and Grade 4) more effectively, while performing only moderately on the middle grades (2 and 3). This difficulty in distinguishing intermediate categories has also been observed by [Rocha et al., 2017] and reflects the inherent ambiguity in human grading. Notably, many farms collapse Grades 2 and 3 into a single class, which aligns with our findings.

Table 4: Ablation study on DINOv2+KNN for 4-class oocyte grading.Results highlight that preprocessing with object detection is critical, improving mean accuracy from 0.52 to 0.61.

	With Preprocessing				Without Preprocessing			
Grade	**Precision**	**Recall**	**F1**	**Acc.**	**Precision**	**Recall**	**F1**	**Acc.**
1	**0.62**	**0.63**	**0.62**	**0.63**	0.38	0.33	0.35	0.33
2	0.52	**0.67**	**0.58**	**0.67**	0.50	0.57	0.53	0.57
3	0.55	0.45	0.49	0.45	**0.56**	**0.52**	**0.54**	**0.52**
4	**0.84**	**0.71**	**0.77**	**0.71**	0.66	0.64	0.65	0.64
Mean	0.63	0.62	0.62	**0.61**	0.53	0.52	0.52	0.52

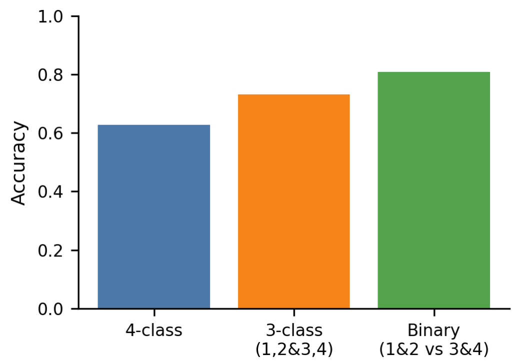

Figure 3: YOLOv8 accuracy on test set under different grading schemes.The grading reflects industry practices where many labs use three-class grading or binary grading to reduce subjectivity.

Generalization analysis (Table 3) using an independent test set of oocytes images cropped from group COC images (multiple oocytes with differnt grades taken in a single image) shows a consistent performance drop across model families. This highlights the challenges of domain shift differences in imaging setups and data distribution. This underscores the need for larger, more diverse datasets to improve robustness.

Finally, it is important to place these results in the context of human performance. [Farin et al., 1995] reported only 69% agreement among six expert technicians when ranking oocyte quality, illustrating the subjectivity of this task. Our results mirror this challenge: AI models face the same ambiguities as humans, particularly in borderline grades, but offer the promise of reproducibility and standardization once larger datasets and improved imaging protocols are available.

3.5 Generalization

We further evaluated model robustness on an independent test set of 40 oocyte images extracted from group COC samples that are not part of main corpus. Despite the smaller and more challenging dataset (because group COC sample are multiple oocytes captured in a single image and then cropped out for evaluation) YOLOv8 maintained competitive performance on the 4-class IETS task. Both DINOv2+KNN and YOLOv8 achieve even higher accuracy under practical grading schemes (three-class and binary), consistent with industry practice. These results reinforce that while fine grained grading remains difficult, groupings provide a reliable and scalable path for deployment. (see table 3)

4 Discussion & Conclusion

Our study demonstrates the feasibility of automating bovine oocyte grading with deep learning. We introduce the first dataset and benchmark aligned with IETS guidelines. Among the evaluated pipelines, YOLOv8 consistently outperformed other CNN and transfer based model under the strict 4-class IETS grading scheme, providing a strong baseline for future work. Grouped grading (three-class that merged classes 2 and 3, as well as binary 1-2 vs 3-4) further improved accuracy, highlighting both the challenge of fine-grained distinctions and the practical utility of coarser schemes in industry settings. Object detection-based preprocessing also helped ensure data consistency and improved downstream classification.

Our work highlights the importance of oocyte grading as an early stage step in *in vitro* embryo production (IVP), an area less explored compared to AI applications on blastocyst stage embryos [Iannone et al., 2024]. Unlike blastocyst prediction, which depends on multiple biological and environmental factors such as genetics, sperm quality, and culture conditions, oocyte grading provides a standardized, immediate, and widely adopted metric in IVF. Future research should explore direct linking oocyte grading to later developmental outcomes such as blastocyst formation and pregnancy rates.

While our baseline models establish feasibility, further improvements in model architectures, imaging technologies, and domain-relevant performance metrics are needed to capture subtle morphological traits. Further work should also include multiple experts grading same set of oocytes and do inter- and intra annotator consistency evaluations for label reliability. Our experiments showed that even low cost setups, such as an iPhone mounted to a microscope, provide a starting point, but robust deployment will require higher resolution imaging and standardized acquisition protocols.

References

Inception V3 — Torchvision main documentation, a. URL https://docs.pytorch.org/vision/main/models/inception.html.

Key Concepts to Consider Before Starting an Embryo Transfer Program for Your Herd | Panhandle Agriculture, b. URL https://nwdistrict.ifas.ufl.edu/phag/2023/03/31/key-concepts-to-consider-before-starting-an-embryo-transfer-program-for-your-herd/.

YOLOv8: State-of-the-Art Computer Vision Model, c. URL https://yolov8.com/.

R. Boni. Origins and effects of oocyte quality in cattle. *Origins and effects of oocyte quality in cattle*, 9(3):333–340, July 2018. URL http://www.animal-reproduction.org/article/5b5a6059f7783717068b46ef.

D. Demetrio, M. Oliveira, R. Reis Silva, D. Amorim, C. Demetrio, and R. Santos. 12 Relationship between Angus oocyte quality and embryo production. *Reproduction, Fertility and Development*, 35(2):131–131, December 2022. ISSN 1031-3613, 1448-5990. doi: 10.1071/RDv35n2Ab12. URL https://www.publish.csiro.au/RD/RDv35n2Ab12.

P. W. Farin, J. H. Britt, D. W. Shaw, and B. D. Slenning. Agreement among evaluators of bovine embryos produced in vivo or in vitro. *Theriogenology*, 44(3):339–349, August 1995. ISSN 0093-691X. doi: 10.1016/0093-691X(95)00189-F. URL https://www.sciencedirect.com/science/article/pii/0093691X9500189F.

L. B. Ferré, M. E. Kjelland, L. B. Strøbech, P. Hyttel, P. Mermillod, and P. J. Ross. Review: Recent advances in bovine in vitro embryo production: reproductive biotechnology history and methods. *Animal: An International Journal of Animal Bioscience*, 14(5):991–1004, May 2020. ISSN 1751-732X. doi: 10.1017/S1751731119002775.

Jullin Fjeldstad, Weikai Qi, Nadia Siddique, Natalie Mercuri, Dan Nayot, and Alex Krivoi. Segmentation of mature human oocytes provides interpretable and improved blastocyst outcome predictions by a machine learning model. *Scientific Reports*, 14(1):10569, May 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-60901-1.

Peter J. Hansen. Review: Some challenges and unrealized opportunities toward widespread use of the in vitro-produced embryo in cattle production. *animal*, 17:100745, May 2023. ISSN 17517311. doi: 10.1016/j.animal.2023.100745. URL https://linkinghub.elsevier.com/retrieve/pii/S1751731123000411.

Antonio Iannone, Alessandro Carfì, Fulvio Mastrogiovanni, Renato Zaccaria, and Claudio Manna. On the role of artificial intelligence in analysing oocytes during *in vitro* fertilisation procedures. *Artificial Intelligence in Medicine*, 157:102997, November 2024. ISSN 0933-3657. doi: 10.1016/j.artmed.2024.102997. URL https://www.sciencedirect.com/science/article/pii/S0933365724002392.

Grace M. Koppelman. *Utilizing Machine Learning for Grading Bovine Cumulus-Oocyte Complexes*. PhD thesis, The Ohio State University, 2025. URL https://etd.ohiolink.edu/acprod/odb_etd/etd/r/1501/10?clear=10&p10_accession_num=osu1744909834572529.

Parthiban Marimuthu. Image Contrast Enhancement Using CLAHE, August 2022. URL https://www.analyticsvidhya.com/blog/2022/08/image-contrast-enhancement-using-clahe/.

United Nations. World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100. URL https://www.un.org/en/desa/world-population-projected-reach-98-billion-2050-and-112-billion-2100.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Annelies Raes, Danilo Babin, Osvaldo Bogado Pascottini, Geert Opsomer, Ann Van Soom, and Katrien Smits. Artificial intelligence outperforms humans in morphology-based oocyte selection in cattle. *Scientific Reports*, 15(1):21829, July 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-09019-6. URL `https://www.nature.com/articles/s41598-025-09019-6`.

José Celso Rocha, Felipe José Passalia, Felipe Delestro Matos, Maria Beatriz Takahashi, Marc Peter Maserati Jr, Mayra Fernanda Alves, Tamie Guibu de Almeida, Bruna Lopes Cardoso, Andrea Cristina Basso, and Marcelo Fábio Gouveia Nogueira. Automatized image processing of bovine blastocysts produced in vitro for quantitative variable determination. *Scientific Data*, 4(1):170192, December 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.192. URL `https://www.nature.com/articles/sdata2017192`.

Anna Targosz, Dariusz Myszor, and Grzegorz Mrugacz. Human oocytes image classification method based on deep neural networks. *BioMedical Engineering OnLine*, 22(1):92, September 2023. ISSN 1475-925X. doi: 10.1186/s12938-023-01153-4. URL `https://doi.org/10.1186/s12938-023-01153-4`.

Ultralytics. YOLO11. URL `https://docs.ultralytics.com/models/yolo11`.