Research Article / *Araştırma Makalesi*

Ege Journal of Medicine / *Ege Tıp Dergisi* 2025; 64 (4): 653-664

Artificial intelligence in assisted reproductive techniques: comparative evaluation of deep learning architectures for bovine cumulus-oocyte complexes classification

Yapay zekâ destekli üreme tekniklerinde: sığır kumulus-oosit komplekslerinin sınıflandırılmasına yönelik derin öğrenme mimarilerinin karşılaştırmalı değerlendirmesi

Aylin Gökhan[1] Seda Çetinkaya Karabekir[2] Emre Ölmez[3] Saadet Özen Akarca Dizakar[2] Mert Can Tekel[4] Orhan Er[5] Mehmet Kemal Güllü[6] Türker Çavuşoğlu[2]

[1]Department of Histology and Embryology, Faculty of Medicine, Ege University, İzmir, Türkiye

[2]Department of Histology and Embryology, Faculty of Medicine, İzmir Bakırçay University, İzmir, Türkiye

[3]Department of Biomedical Engineering, Faculty of Engineering and Architecture, İzmir Bakırçay University, İzmir, Türkiye

[4]Pınar Integrated Meat and Flour Industry Inc., İzmir, Türkiye

[5]Department of Computers Engineering, Faculty of Engineering and Architecture, İzmir Bakırçay University, İzmir, Türkiye

[6]Department of Electrical-Electronics Engineering, Faculty of Engineering and Architecture, İzmir Bakırçay University, İzmir, Türkiye

ABSTRACT

Aim: Follicular quality is a key determinant of success in assisted reproductive technologies, directly affecting outcomes such as fertilization, embryo development, implantation, and live birth rates. However, conventional assessment of cumulus-oocyte complexes relies on subjective morphological evaluation, introducing variability and reducing consistency in clinical decision-making.

Materials and Methods: A comparative evaluation of various pre-trained deep learning architectures—including both convolutional neural networks and transformer-based models—was conducted for the automated morphological grading of bovine cumulus-oocyte complexes into four quality categories (Grade A–D). A dataset of 1,400 annotated images of cumulus-oocyte complexes, enhanced through data augmentation techniques to increase image diversity, was used for model training and validation.

Results: Among the tested architectures, Xception41 (a variant of convolutional neural networks) and Swin Transformer (a transformer-based model) achieved the highest performance, with test accuracies of 74.75% and 73.25%, and macro F1-scores of 0.75 and 0.74, respectively. While both models performed well in grading cumulus-oocyte complexes with distinct morphological features (Grades 3 and 4), classification accuracy decreased for the more subtle differences between Grades 1 and 2. Furthermore, most models exhibited signs of overfitting under the current training configuration.

Conclusion: This study demonstrates the potential of deep learning-based approaches to standardize and enhance the efficiency of cumulus-oocyte complexes evaluation in assisted reproductive technologies. Further optimization is needed to improve model generalization and to address challenges in grading morphologically similar follicular structures.

Corresponding author: Türker Çavuşoğlu

İzmir Bakırçay University, Faculty of Medicine, Department of Histology and Embryology, İzmir, Türkiye.

E-mail: *turker.cavusoglu@bakircay.edu.tr.*

Application date: 08.07.2025 Accepted: 11.08.2025

Keywords: Artificial intelligence, convolutional neural networks, xception41, deep learning, swin transformer, cumulus-oocyte complexes, oocyte classification, assisted reproductive technologies

ÖZ

Amaç: *Folikül kalitesi, yardımcı üreme teknolojilerindeki başarıyı belirleyen temel bir faktördür ve döllenme, embriyo gelişimi, implantasyon ve canlı doğum oranları gibi önemli sonuçları doğrudan etkiler. Ancak, kumulus-oosit komplekslerinin geleneksel değerlendirmesi, öznel morfolojik gözlemlere dayandığı için klinik karar süreçlerinde değişkenliğe ve tutarsızlığa neden olmaktadır.*

Gereç ve Yöntem: *Sığır kumulus-oosit komplekslerinin otomatik olarak morfolojik kaliteye göre dört kategoriye (A–D) ayrılması amacıyla, önceden eğitilmiş çeşitli derin öğrenme mimarileri—evrişimsel sinir ağları ve dönüştürücü tabanlı modeller dahil—karşılaştırmalı olarak değerlendirilmiştir. Görsel çeşitliliği artırmak amacıyla veri artırma teknikleri uygulanarak oluşturulan 1.400 etiketlenmiş kumulus-oosit kompleksi görüntülerinden oluşan veri seti, modellerin eğitimi ve doğrulaması için kullanılmıştır.*

Bulgular: *Test edilen mimariler arasında Xception41 (evrişimsel sinir ağları varyantı) ve Swin Transformer (dönüştürücü tabanlı bir model), sırasıyla %74,75 ve %73,25 test doğrulukları ile 0,75 ve 0,74 makro F1 skorlarına ulaşarak en yüksek performansı göstermiştir. Bu modeller, belirgin morfolojik özelliklere sahip 3. ve 4. derece kumulus-oosit komplekslerinde yüksek sınıflandırma başarısı gösterirken, 1. ve 2. dereceler arasındaki daha ince farkların ayırt edilmesinde zorlanmıştır. Ayrıca, mevcut eğitim konfigürasyonu altında çoğu modelde aşırı öğrenme eğilimi gözlemlenmiştir.*

Sonuç: *Bu çalışma, derin öğrenme tabanlı yaklaşımların yardımcı üreme teknolojileri kapsamında kumulus-oosit kompleksi değerlendirmesini standartlaştırma ve değerlendirme süreçlerinin etkinliğini artırma potansiyelini ortaya koymaktadır. Bununla birlikte, morfolojik olarak benzer foliküler yapılar arasındaki sınıflandırma zorluklarının aşılması ve model genelleme yeteneğinin artırılması için ilave optimizasyon gereklidir.*

Anahtar Sözcükler*: Yapay zekâ, evrişimli sinir ağları, xception41, derin öğrenme, swin transformer, kumulus-oosit kompleksi, oosit sınıflandırması, yardımcı üreme teknikleri*

INTRODUCTION

Infertility affects approximately 9% of reproductive-age couples globally, presenting substantial medical and socioeconomic challenges (1). Assisted reproductive technologies (ART), particularly in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI), have significantly advanced infertility treatment. However, success rates remain inconsistent, influenced by patient demographics, clinic protocols, and treatment variability (2). These discrepancies underscore the need to identify reliable biomarkers, particularly those related to oocyte quality, to optimize ART outcomes.

One of the most critical determinants of ART success is oocyte quality, which directly influences fertilization potential, embryonic development, and implantation success (3,4). Evaluating oocyte quality traditionally involves a visual morphological assessment conducted by embryologists, who examine key indicators such as cumulus-oocyte complex (COC) integrity, the structural integrity of the zona pellucida (ZP), and the uniformity of cytoplasmic features (4,5). This approach is prone to inter-observer variability, underscoring the need for objective, standardized methods (6). Artificial intelligence (AI), particularly deep learning models such as convolutional neural networks (CNNs) and transformers, has shown promise in automating medical image analysis and enhancing diagnostic consistency (7,8).

Deep learning, particularly through the use of CNNs and transformer architectures, emerges as a transformative tool that has shown remarkable accuracy in medical imaging applications. These AI-driven techniques are increasingly explored for their potential to revolutionize the classification processes within the realm of fertility and reproductive medicine (3,7). Recent studies have indicated that deep learning algorithms can vastly improve the standardization of oocyte assessments, facilitate automation of the classification processes, and enhance clinical efficiencies. The applications of these models are particularly relevant as they may not only streamline the evaluation procedures but also minimize biases introduced through subjective human judgments (8,9).

Despite existing research focusing on the application of AI in grading embryos, its utilization for COC evaluation is less explored, indicating a significant gap that this study aims to address. This research evaluates the performance of various CNN and transformer-based models to automate the classification of COCs according to four distinct morphological quality grades with the objective of surmounting limitations posed by manual assessments (3,7). By employing algorithms coupled with data augmentation techniques, the investigation seeks to optimize the efficiency of ART workflows, potentially leading to improved clinical practices surrounding oocyte selection, and subsequently enhanced success rates in ART treatments (8,9).

MATERIALS AND METHODS

1. Ethics Approval

This study received ethical approval in January 2022 and was funded by the İzmir Bakırçay University Scientific Research Projects Commission (Project No: BBAP.2022.006). Ovaries were ethically obtained as by-products from animals slaughtered for human consumption. The study adhered to internationally recognized ethical standards, including the National Institutes of Health (NIH) Guide for the Care and Use of Laboratory Animals. All procedures were conducted in accordance with institutional regulations and ethical guidelines pertaining to the collection and use of animal tissues for research purposes.

2. Ovarian Collection and Transport

Ovaries intended for COC retrieval were collected at commercial slaughterhouses and transported to the laboratory under controlled conditions. They were placed in a transportation medium consisting of RPMI 1640 (#21875034, Thermo Fisher Scientific, UK) supplemented with 100 mg/L of Penicillin-Streptomycin, maintained at 20–25°C in steel thermos containers to preserve cellular viability. Upon arrival, the ovaries were rinsed twice with a pre-warmed phosphate-buffered saline (PBS) solution supplemented with 5% fetal bovine serum (FBS) (#FBS-HI-22A, Capricorn Scientific GmbH, Germany) to remove blood and contaminants before processing **(**Figure-1A**)**.

3. Follicular Aspiration

COCs were retrieved from follicles measuring 2–10 mm in diameter using a 21-gauge needle attached to a sterile syringe (Beybi, Istanbul, Türkiye). The aspirated follicular fluid (5–10 mL per ovary) was collected in 50 mL Falcon tubes and left undisturbed at room temperature for 20 minutes to allow sedimentation. The COCs were carefully extracted from the sediment using a Pasteur pipette and transferred to cell culture dishes for microscopic evaluation (Figure-1B–C). Under a stereomicroscope, embryologists identified and retrieved COCs using a 10 µL Transferpettor (Brand, 701807), capturing high-resolution digital images for further analysis (Figure-1D).

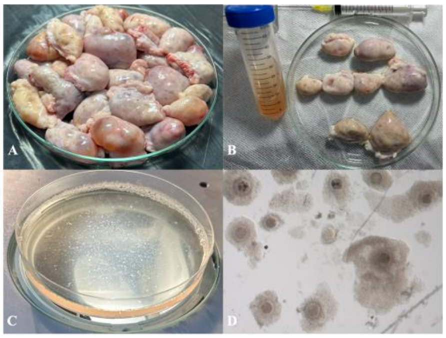

Figure-1. Sequential steps in Cumulus-Oocyte Complex (COC) collection and evaluation.
(A) excised ovaries; (B) follicular aspiration; (C) follicular fluid with COCs; (D) microscopic view showing varying cumulus expansion and structural integrity.

4. Classification of Cumulus-Oocyte Complexes

COCs were classified into four grades based on microscopic morphological characteristics, following established criteria described in previous studies (8,9):

- **Grade-1**. Homogeneous cytoplasm, intact zona pellucida (ZP), and compact, multilayered cumulus cells.
- **Grade-2**. Homogeneous cytoplasm with minor pigmentation, intact ZP, and cumulus layers exceeding five but fewer than in Grade 1.
- **Grade-3**. Heterogeneous cytoplasm with vacuoles, sparse cumulus layers (3–5 layers), and a partially intact ZP.

• **Grade-4**. Pigmented, heterogeneous cytoplasm, disrupted or absent cumulus cells, and compromised ZP integrity.

All COCs were photographed using an Olympus DP72 digital camera mounted on an Olympus BX51 microscope and processed with cellSens Entry software for subsequent machine learning analysis (**Figure 2, left panel**). Images were captured at 400× total magnification to ensure adequate morphological detail. Morphological classification was performed by a single experienced embryologist to ensure consistency, using a similar evaluation approach to that applied in our previous study (9).

5. Data Collection and Preparation

A dataset of 1,400 annotated COC images was systematically curated and divided into training (n=1,000) and test (n=400) sets. Each morphological grade (Grades 1–4) contained 350 images, of which 250 were allocated for training and 100 were reserved for independent testing.

To ensure robust performance evaluation, a separate test dataset of 400 COCs (100 per grade) was used exclusively for validation, preventing overfitting and enhancing the reliability of the CNN's classification capabilities (Figure-2, right panel).

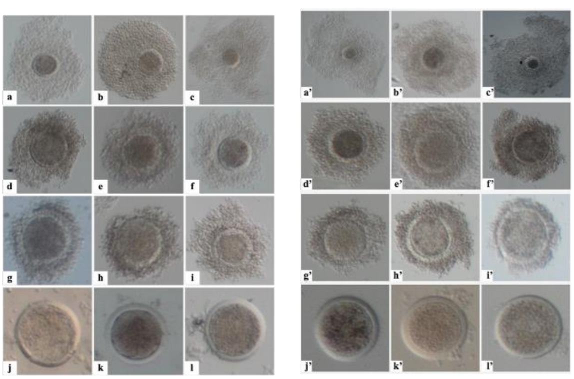

Figure-2. Representative Images of Cumulus-Oocyte Complexes (COCs). Left Panel: Training set images categorized into four morphological grades: (a–c) Grade 1, (d–f) Grade 2, (g–i) Grade 3, (j–l) Grade 4. Right Panel: Independent test set images, similarly categorized (a′–c′ to j′–l′), used for model validation.

6. Deep Learning Model Architecture

This study evaluated a range of state-of-the-art deep learning models for the automated classification of bovine COCs into four distinct morphological quality grades. The selected architecture encompassed both CNN-based and Transformer-based models, all sourced from the PyTorch Image Models (TIMM) library due to their proven efficacy in biomedical image analysis.

The CNN-based architectures included Xception (10), Xception41 (11), ConvNeXt (convnext_base) (12), ResNetRS50 (13,14), and DenseNet201 (15,16). The original Xception (10) model, inspired by Google's Inception architecture, is built on depthwise separable convolutions that decouple spatial and cross-channel filtering operations. This design substantially reduces computational cost while preserving high classification performance.

Xception41 (11), a more advanced variant frequently utilized within the DeepLab v3+ semantic segmentation framework, extends the original architecture by increasing the number of middle flow blocks and replacing MaxPooling operations with strided depthwise convolutions. It also integrates Batch Normalization and ReLU activation after each 3×3 depthwise convolution, thereby enhancing model expressiveness and training stability.

ConvNeXt (12) represents a contemporary reinterpretation of the traditional CNN, integrating several architectural principles from Transformer models, such as expanded kernel sizes, GELU activation functions, and layer normalization. These refinements enable ConvNeXt to narrow the performance gap between CNNs and Transformers while maintaining computational efficiency.

ResNetRS50 (ResNet ReScaled) (14) is a re-scaled and optimized variant of the original ResNet architecture (17,18). It revisits the baseline design by integrating improved training and scaling strategies, including the ResNet-D stem and Squeeze-and-Excitation (SE) blocks (19–21). These enhancements facilitate more effective feature recalibration and enrich representational capacity across channels, ultimately enhancing performance in visual recognition tasks.

DenseNet201 (15), a member of the DenseNet family, employs a densely connected architecture in which each layer receives input from all preceding layers. This design maximizes feature reuse, improves gradient propagation, and facilitates the training of deeper networks with fewer parameters. The specific instance used in this study, DenseNet201, consists of 201 layers and is well-regarded for its compactness and high efficiency in complex image classification tasks, particularly in medical imaging domains.

The Transformer-based models evaluated in this study included the Vision Transformer (ViT, vit_base_patch16_224) (17,18), BERT Pre-Training of Image Transformers (BEiT, beit_base_patch16_224) (19–21), and the Shifted Window Transformer (Swin Transformer, Swin, swin_base_patch4_window7_224) (22,23). The ViT reformulates image classification as a sequence modeling task by dividing each image into fixed-size patches and processing them through standard Transformer encoders using global self-attention mechanisms (17,18). The BEiT model builds on this framework by employing a masked image modeling (MIM) strategy during pretraining, analogous to the Bidirectional Encoder Representations from Transformers (BERT) framework in natural language processing. This method enables the model to learn contextualized visual representations by reconstructing masked image patches, improving generalization across image domains (19–21). The Swin Transformer introduces a hierarchical architecture that computes self-attention within locally shifted windows. This design reduces computational complexity to linear with respect to image size while preserving the model's ability to capture both local and global contextual information, making it particularly effective for high-resolution image classification tasks (22,23).

All models were configured to process input images with dimensions of 224 × 224 × 3. Each architecture incorporates distinct structural components reflective of its design paradigm—for instance, convolutional and pooling layers in CNN-based models, and attention mechanisms with Transformer blocks in Transformer-based architectures. Standard activation functions were employed according to architectural norms, such as ReLU for CNNs and GeLU for Transformer models. Training was conducted using the ADAM optimizer with a fixed learning rate of 0.0001, over 50 epochs, and with a batch size of 32. The deep learning models selected for this study are widely established in the field and are well-regarded for their strong performance across a range of image classification benchmarks.

7. Data Augmentation

Enhancing the diversity of the training dataset and improving the model's generalization capability were addressed through a targeted data augmentation and preprocessing strategy, applied exclusively to the training images. At each training iteration, as batches were loaded, all images were uniformly resized to 224 × 224 pixels. Following resizing, a series of random augmentation techniques were applied, including random horizontal flipping with a 50% probability, random rotation within a range of ±15 degrees, and color jittering with brightness and contrast adjustments up to a factor of 0.2.

After augmentation, each image was converted into a PyTorch tensor to ensure compatibility with the model's input format. Standard normalization was then performed using the mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225], based on the ImageNet dataset from which the pretrained model weights were derived.

This dynamic augmentation and standardized preprocessing pipeline ensured that the model encountered a variety of visual representations of the same image across training epochs. Exposure to this variability encouraged the learning of robust and invariant morphological features, reduced overfitting, and ultimately enhanced the model's ability to generalize effectively to unseen data.

8. Performance Metrics

The performance of the deep learning models utilized in this study was evaluated using fundamental classification metrics derived from the confusion matrix. These metrics included accuracy, precision, recall (also known as sensitivity), and F1-score. These evaluation criteria enabled a comprehensive assessment of the models' capacity to distinguish between different oocyte grades, considering both class-specific performance and overall predictive capability (Table-1) (24).

Table-1. Calculation of Performance Metrics (24).

Metric	Formula	Purpose
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Measures the proportion of correctly classified samples among the total samples.
Precision	TP / (TP + FP)	Evaluates the proportion of true positive predictions among all instances predicted as positive.
Recall	TP / (TP + FN)	Assesses the model's ability to correctly identify true positive cases (sensitivity).
F1-Score	2 x (Precision x Recall) / (Precision + Recal)	Provides a harmonic mean between precision and recall, offering a balanced performance measure.

TP: True Positives. TN: True Negatives. FP: False Positives. FN: False Negatives.

RESULTS

1. Overall Model Performance Comparison

A comparative evaluation of the overall performance of the deep learning models developed for COC classification was conducted, emphasizing training and test (validation) accuracy and loss metrics recorded after 50 training epochs. The overall performance metrics of the deep learning models are summarized in the table (Table-2).

The evaluated models demonstrated varying levels of performance on the independent test dataset, with test accuracy ranging from 67.00% to 74.75%. The highest test accuracy was achieved by the Xception41 model at 74.75%, followed closely by the Swin Transformer at 73.25%. Other high-performing models included ConvNeXt and ResNetRS50, which achieved test accuracy of 71.25% and 70.75%, respectively. In contrast, the baseline Xception model yielded the lowest test accuracy among the architectures evaluated, at 67.00%. Models such as BEiT, ViT, and DenseNet201 exhibited intermediate performance, with test accuracy falling between 69.50% and 69.75%.

An examination of the training dynamics revealed a notable disparity between training and test performance across most models. Architectures such as ConvNeXt, Swin Transformer, Xception41, Xception, ResNetRS50, and DenseNet201 achieved exceptionally high training accuracy, often exceeding 99% and in some instances reaching 100%. Despite these near-perfect training results, test accuracies were consistently lower, indicating limited generalization to previously unseen data. This discrepancy was also reflected in the loss values, where models reported minimal training losses but substantially higher test losses. These patterns collectively suggest that, under the current training configuration, most models exhibited signs of overfitting—performing exceptionally well on the training data while struggling to maintain similar accuracy levels during testing.

Table-2. Overall Performance Metrics of Deep Learning Models

Model Name	Train_Loss	Train_Accuracy	Test_Loss	Test_Accuracy
xception	0.0239	0.9940	1.5491	0.6700
xception41	0.0069	0.9990	1.4960	0.7475
convnext_base	0.0001	1.0000	1.9743	0.7125
vit_base_patch16_224	0.1021	0.9640	1.5022	0.6950
beit_base_patch16_224	0.1185	0.9550	1.2025	0.6975
swin_base_patch4_window7_224	0.0013	1.0000	2.0096	0.7325
resnetrs50	0.0288	0.9930	1.3963	0.7075
densenet201	0.0308	0.9930	1.4180	0.6950

2. Class-Specific Performance Analysis via Confusion Matrices

A detailed examination of the classification performance across the four COC grades was conducted using confusion matrices generated from the test dataset. As illustrated in Figure-3, these matrices provide a comprehensive breakdown of correct and incorrect predictions for each grade, offering insights into specific error trends and patterns of inter-grade misclassification (Figure-3).

The evaluated models exhibited varying levels of accuracy in distinguishing between COC grades. Among them, the Xception41 model, which achieved the highest overall test accuracy (74.75%) as detailed in the overall model performance comparison, exhibited strong predictive capability in recognizing morphologically distinct categories, correctly classifying 89 instances of Grade 4 and 61 instances of Grade 1. Despite this high accuracy, the model demonstrated substantial misclassification between neighboring grades, notably mislabeling 33 instances of Grade 1 as Grade 2, and 28 instances of Grade 2 as Grade 1. A similar trend was observed in the Swin Transformer, which also displayed high classification performance by correctly identifying 87 instances of Grade 4 and 82 of Grade 3. However, notable classification ambiguities remained prevalent between morphologically adjacent grades, particularly Grades 1 and 2.

Similar trends were observed in other competitive models such as ConvNeXt and ResNetRS50. The ConvNeXt model accurately classified 87 instances of Grade 4 and 75 of Grade 3, while ResNetRS50 correctly identified 80 instances of Grade 4 and 82 of Grade 3, further confirming robust detection of higher-quality oocytes with more pronounced morphological features.

Across all models, Grades 1 and 2 consistently appeared as the most frequently misclassified categories. In contrast, better-performing models like Xception41 and Swin Transformer showed improved differentiation between Grades 3 and 2, as well as between Grades 4 and 3. Misclassifications between the morphological extremes-Grade 1 and Grade 4-were minimal, indicating strong recognition of highly divergent phenotypes.

Collectively, these findings suggest that while the models are effective in distinguishing oocytes with clear morphological differences, challenges remain in resolving borderline classifications. Enhancing model sensitivity to subtle, yet clinically relevant, morphological nuances is essential for improving the reliability of automated COC assessment in the context of ART.

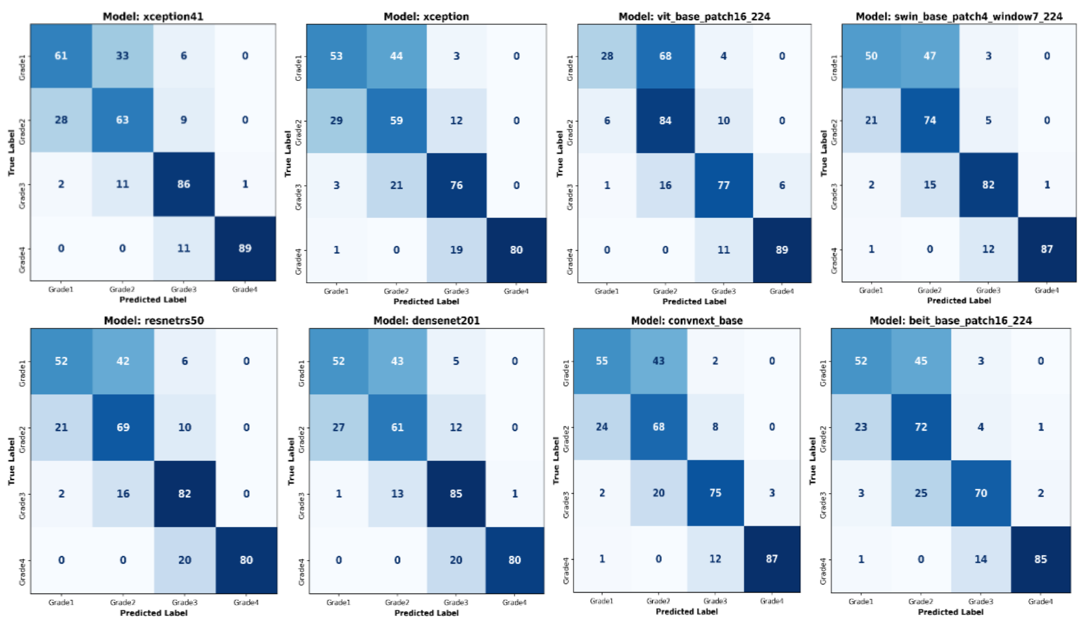

Figure-3. Confusion Matrices of the Evaluated Deep Learning Models for COC Grade Classification.

3. Class-Based and Macro-Averaged Performance Metrics of Deep Learning Architectures in COC Grading

The classification performance of the evaluated deep learning models in COC grading was assessed using class-specific precision, recall, and F1-score metrics across the four standard morphological grades (Grades 1 through 4). Besides these grade-specific evaluations, macro-averaged values for these metrics were also calculated to provide an overall performance summary that considers each class equally, regardless of class imbalance. The detailed results are shown in the tables summarizing class-wise and macro-averaged precision, recall, and F1-score values for the COC grades across deep learning models (Table-3 and Table-4).

Performance varied significantly across the four COC grades. The models consistently showed the highest discriminative accuracy for Grade 4, with the Xception41 achieving an F1-score of 0.94 and the Swin Transformer reaching 0.93. Precision scores for this grade were consistently high, indicating a strong match between predictions and ground truth labels. Grade 3 also performed well; both Xception41 and Swin Transformer recorded F1-scores of 0.81, demonstrating reliable recognition of its morphological features.

In contrast, the classification results for Grades 1 and 2 were less favorable. The Xception41 model achieved F1-scores of 0.64 and 0.61 for Grades 1 and 2, respectively, while the Swin Transformer recorded scores of 0.57 and 0.63. Notably, the performance of the ViT architecture for Grade 1 is especially significant. Although it achieved a precision of 0.80, its recall dropped to 0.28, indicating a tendency to overlook many relevant cases. This imbalance resulted in a modest F1-score of 0.41, demonstrating limited sensitivity despite high confidence in positive predictions. To evaluate the overall model performance more fairly, macro-averaged precision, recall, and F1-score values were analyzed. These metrics offer a comprehensive view of classification performance by treating all grades as equally important, regardless of their frequency in the dataset.

Among the evaluated architectures, Xception41 demonstrated the strongest performance, achieving a macro-averaged F1-score of 0.75, with Swin Transformer close behind at 0.74. These results correspond with the test accuracy rankings in the Overall Model Performance Comparison. ConvNeXt and ResNetRS50 scored macro F1-scores of 0.72 and 0.71, respectively, while Beit and DenseNet201 were around 0.70. Xception and ViT had the lowest combined scores, each with a macro-average F1 of 0.68. Additionally, Xception41 outperformed others in macro-averaged precision (0.76) and recall (0.75), emphasizing its overall effectiveness. This thorough analysis highlights the comparative strengths and limitations of modern deep learning frameworks in the context of COC classification. By showcasing performance trends across morphological grades, it provides key insights into the suitability of architectures for potential deployment in diagnostic workflows and related biomedical applications.

Table-3. Class-Wise Precision, Recall, and F1-Score Values for Each COC Grade Across Deep Learning Models

	Grade 1			Grade 2			Grade 3			Grade 4		
Model	**Precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**	**precision**	**recall**	**f1**
xception	0.62	0.53	0.57	0.48	0.59	0.53	0.69	0.76	0.72	1.00	0.80	0.89
xception41	0.67	0.61	0.64	0.59	0.63	0.61	0.77	0.86	0.81	0.99	0.89	0.94
Convnext	0.67	0.55	0.6	0.52	0.68	0.59	0.77	0.75	0.76	0.97	0.87	0.92
Vit	0.8	0.28	0.41	0.50	0.84	0.63	0.75	0.77	0.76	0.94	0.89	0.91
Beit	0.66	0.52	0.58	0.51	0.72	0.60	0.77	0.70	0.73	0.97	0.85	0.90
Swin	0.68	0.5	0.57	0.54	0.74	0.63	0.80	0.82	0.81	0.99	0.87	0.93
resnetrs50	0.69	0.52	0.59	0.54	0.69	0.61	0.69	0.82	0.75	1.00	0.80	0.89
densenet201	0.65	0.52	0.58	0.52	0.61	0.56	0.70	0.85	0.77	0.99	0.80	0.88

Table-4. Macro-Averaged Precision, Recall, and F1-Score for All COC Grades

	Macro Averaged Metrics		
Model	**Precision**	**Recall**	**f1**
xception	0.70	0.67	0.68
xception41	0.76	0.75	0.75
Convnext	0.73	0.71	0.72
Vit	0.75	0.70	0.68
Beit	0.73	0.70	0.70
Swin	0.75	0.73	0.74
resnetrs50	0.73	0.71	0.71
densenet201	0.72	0.70	0.70

DISCUSSION

The morphological assessment of COCs represents a critical step in ART, as it directly influences embryo developmental potential and pregnancy outcomes. In this study, the performance of various CNN and Transformer-based deep learning architectures was comprehensively evaluated for the automatic classification of bovine COC images into four distinct quality grades. The findings highlight both the potential of these technologies to reduce subjectivity and improve standardization in COC assessment, as well as some important challenges that remain to be addressed (25).

Our results demonstrated that among the evaluated models, the Xception41 and Swin Transformer architectures yielded the most promising outcomes in COC classification. The fact that both a modern CNN derivative (Xception41) and a hierarchical Transformer model (Swin Transformer) achieved top performances suggests that different architectural approaches can be successful in tackling this complex image classification problem. The ability of these models to effectively extract and learn intricate image features likely contributed to their relatively high classification performance (26).

Class-based performance analyses revealed that all models generally achieved higher classification accuracy for Grade 4 and Grade 3 COCs, which are characterized by more pronounced morphological features. This trend can be attributed to the ease with which models can distinguish these grades based on overt structural abnormalities (26,27). Conversely, the classification of Grade 1 and Grade 2 COCs proved more challenging for most models, with frequent misclassifications observed between these two adjacent grades. This difficulty likely stems from the subtle morphological transitions between Grade 1 and Grade 2, which are inherently ambiguous and have been associated with inter-observer variability even in manual assessments. Particularly, the ViT model demonstrated high precision but very low recall for Grade 1, indicating that while its predictions for this class were largely correct, it failed to detect the majority of Grade 1 samples —ultimately leading to a modest F1-score for this category (27).

One of the most notable findings of this study was the significant discrepancy between training and test performance observed across the majority of models. While training accuracies often exceeded 99%, test accuracies remained within a relatively modest range of 67–75% (28). This pattern, coupled with low training losses and considerably higher test losses, strongly suggests that the models exhibited a tendency toward overfitting under the current training configuration (50 epochs, specific learning rates, and data augmentation strategies). Overfitting results in models memorizing patterns within the training data while failing to generalize effectively to unseen samples. Possible contributing factors to this issue include insufficient dataset size and diversity relative to model complexity, as well as limitations in the data augmentation strategies employed, which may not have been adequate to fully mitigate overfitting (29).

The strengths of this study include the comparative evaluation of eight contemporary and diverse deep learning architectures (both CNN and Transformer-based) using a common dataset and standardized performance metrics. This approach provides valuable insights into the relative effectiveness of different modeling strategies for COC classification. However, several limitations should also be acknowledged. Chief among these is the tendency for overfitting, as discussed above. Additionally, the size and diversity of the dataset could be further expanded to enhance model generalizability. Improving performance for the more challenging Grade 1 and Grade 2 classifications remains a critical area for future development. Another limitation lies in the use of two-dimensional images, which may not fully capture the complex three-dimensional morphology of COCs.

In future research, efforts should focus on addressing overfitting through the implementation of different regularization techniques (e.g., adjusting dropout rates, applying weight decay), as well as optimizing advanced data augmentation strategies and transferring learning approaches. Balancing model complexity with dataset size may involve exploring fewer complex models or utilizing model pruning techniques. Training and evaluating models on larger, multi-center, and diverse datasets obtained under various imaging conditions could substantially improve generalizability (30). Furthermore, integrating additional biomarkers — such as metabolomic data or gene expression profiles — alongside morphological images in multimodal artificial intelligence models holds potential to enhance the accuracy of COC selection. Lastly, prospective clinical validation of the most promising models would represent a crucial step toward the integration of these technologies into routine ART practice.

CONCLUSION

In this study, the performance of various pre-trained CNN and Transformer-based deep learning architectures was comparatively evaluated for the classification of bovine COCs into four distinct morphological quality grades. The results demonstrated that the Xception41 and Swin Transformer models achieved the highest performances, with test accuracies of 74.75% and 73.25%, and macro F1-scores of 0.75 and 0.74, respectively. Overall, the models exhibited higher classification success for Grade 4 and Grade 3 COCs, which display more prominent morphological features, while greater challenges were encountered in distinguishing the more subtle differences between Grade 1 and Grade 2.

An important observation was the tendency for most models to overfit under the current training configuration, as evidenced by the notable gap between training and test performances. This finding underscores the need for addressing overfitting and further optimizing the models to improve their generalization ability.

This study highlights the potential of deep learning models to offer a more objective, standardized, and potentially more efficient alternative to traditional subjective morphological COC assessment. Addressing challenges such as overfitting and enhancing classification accuracy for difficult COC grades remain essential priorities. Future work involving larger and more diverse datasets, advanced modeling techniques, and multimodal data integration is expected to make significant contributions to improving the success rates of artificial intelligence applications in ART.

Conflict of Interest

The supply of waste material, ovarian tissues, was provided by the Author M.C.T., who works at Pınar Integrated Meat and Flour Industry Inc. (Izmir, Türkiye). This affiliation did not affect the current study. No financial or material support was provided by the company. The authors declare that they have no conflict of interest.

Financial Disclosures:

This study was supported by the İzmir Bakırçay University Scientific Research Projects Coordination Unit under Project ID: BBAP.2022.006.

Acknowledgments:

The authors gratefully acknowledge the support of the İzmir Bakırçay University Rectorate and the Scientific Research Projects Coordination Unit.

References

1. Fainberg J, Kashanian JA. Recent advances in understanding and managing male infertility. F1000Res. 2019;8:670. doi:10.12688/f1000research.17076.1
2. Crawford S, Fussman C, Bailey M, Bernson D, Jamieson DJ, Murray-Jordan M, et al. Estimates of lifetime infertility from three states: The Behavioral Risk Factor Surveillance System. J Womens Health (Larchmt). 2015;24(7):578-86. doi:10.1089/jwh.2014.5102
3. Čegar B, Šipetić Grujičić S, Bjekić J, Vuksanović A, Bojanić N, Bartolović D, et al. Understanding the male perspective: evaluating quality of life and psychological distress in Serbian men undergoing infertility treatment. Life (Basel). 2023;13(9):1894. doi:10.3390/life13091894
4. Uhde K, van Tol HTA, Stout TAE, Roelen BAJ. Metabolomic profiles of bovine cumulus cells and cumulus-oocyte-complex-conditioned medium during maturation in vitro. Sci Rep. 2018;8:9477. doi:10.1038/s41598-8-27829-9
5. Richani D, Dunning KR, Thompson JG, Gilchrist RB. Metabolic co-dependence of the oocyte and cumulus cells: essential role in determining oocyte developmental competence. Hum Reprod Update. 2021;27(1):27-47. doi:10.1093/humupd/dmaa025
6. Choi Y, Moon SH. Types and characteristics of stress coping in women undergoing infertility treatment in Korea. Int J Environ Res Public Health. 2023;20(3):2648. doi:10.3390/ijerph20032648
7. Stojkovic M, Machado SA, Stojkovic P, Zakhartchenko V, Hutzler P, Gonçalves PB, Wolf E Mitochondrial distribution and adenosine triphosphate content of bovine oocytes before and after in vitro maturation: correlation with morphological criteria and developmental capacity after in vitro fertilization and culture. Biol Reprod. 2001;64(3):904-909
8. Hartmann W, Pereira JFS. Biotechnics applied to bovine female. In: Bergstein-Galan TG, editor. Reproduction Biotechnology in Farm Animals. Brazil: AvidScience; 2018. p. 155-180
9. Cavusoglu T, Gokhan A, Sirin C, Tomruk C, Kilic KD, Olmez E, et al. Classification of bovine cumulus-oocyte complexes with convolutional neural networks. Med Records. 2023;5(3):489-5. doi:10.37990/medr.1292782
10. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway (NJ): IEEE; 2017. p. 1800-7. doi:10.1109/CVPR.2017.195
11. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation [Internet]. arXiv; 2018 Feb 7 [cited 2025 Aug 10]. Available from: http://arxiv.org/abs/1802.02611.
12. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s [Internet]. arXiv; 2022 Jan 10 [cited 2025 Aug 10]. Available from: http://arxiv.org/abs/2201.03545.
13. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11211. Cham: Springer; 2018. p. 833-51. doi:10.1007/978-3-030-01234-2_49
14. Park W, Schwendicke F, Krois J, Huh JK, Lee JH. Identification of dental implant systems using a large-scale multicenter data set. J Dent Res. 2023;102(7):727-33. doi:10.1177/00220345231160750
15. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway (NJ): IEEE; 2017. p. 2261-9. doi:10.1109/CVPR.2017.243
16. Liao Q, Zhang Q, Feng X, Huang H, Xu H, Tian B, et al. Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring. Commun Biol. 2021;4:415. doi:10.1038/s42003-021-01937-1

17. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale [Internet]. arXiv; 2020 Oct 23 [cited 2025 Aug 10]. Available from: http://arxiv.org/abs/2010.11929

18. Savaş S. Application of deep ensemble learning for palm disease detection in smart agriculture. Heliyon. 2024;10(17):e37141. doi:10.1016/j.heliyon.2024.e37141

19. Bao H, Dong L, Piao S, Wei F. BEiT: BERT pre-training of image transformers [Internet]. arXiv; 2021 Jun 15 [cited 2025 Aug 10]. Available from: http://arxiv.org/abs/2106.08254

20. ElMoaqet H, Janini R, Ryalat M, Al-Refai G, Abdulbaki Alshirbaji T, Jalal NA, et al. Using masked image modelling transformer architecture for laparoscopic surgical tool classification and localization. Sensors (Basel). 2025;25(10):3017. doi:10.3390/s25103017

21. Kong F, Shi Z, Cao H, Hao Y, Cao Q. Self-supervised U-transformer network with mask reconstruction for metal artifact reduction. Phys Med Biol. 2025;70(6):065009. doi:10.1088/1361-6560/adbaae

22. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway (NJ): IEEE; 2021. p. 9992-10002. doi:10.1109/ICCV48922.2021.00986

23. Omer AAM. Image classification based on vision transformer. J Comput Commun. 2024;12(4):49-59. doi:10.4236/jcc.2024.124005

24. Li Y, Riganello F, Yu J, Vatrano M, Shen M, Cheng L, et al. The autonomic response following taVNS predicts changes in level of consciousness in DoC patients. Sci Rep. 2025;15:7317. doi:10.1038/s41598-024-84029-4

25. Ishaq M, Raza S, Rehar H, Abadeen SeZu, Hussain D, Naqvi RA, et al. Assisting the human embryo viability assessment by deep learning for in vitro fertilization. Mathematics. 2023;11(9):2023. doi:10.3390/math11092023

26. Arsalan M, Haider A, Choi J, Park KR. Detecting blastocyst components by artificial intelligence for human embryological analysis to improve success rate of in vitro fertilization. J Pers Med. 2022;12(2):124. doi:10.3390/jpm12020124

27. Mushtaq A, Mumtaz M, Raza A, Salem N, Yasir MN. Artificial intelligence-based detection of human embryo components for assisted reproduction by in vitro fertilization. Sensors (Basel). 2022;22(19):7418. doi:10.3390/s22197418

28. Jayashree P, Mitra S. Facilitating a deep approach to learning: an innovative case assessment technique. J Manag Organ. 2012;18(4):555-572. doi:10.5172/jmo.2012.18.4.555

29. Allahbadia GN. Embryo transfer is the last frontier for deep machine learning and artificial intelligence in medically assisted reproduction (MAR). J Reprod. 2023;2(1):28-38. doi:10.58779/issn.2954-467X.tjor2023.v2.n1.18.

30. Athanasiou G, Cerquides J, Raes A, Azari-Dolatabad N, Angel-Velez D, Van Soom A, et al. Detecting the area of bovine cumulus oocyte complexes using deep learning and semantic segmentation. In: Frontiers in Artificial Intelligence and Applications. Vol. 356, Artificial Intelligence Research and Development. Amsterdam: IOS Press; 2022. p. 249-258. doi:10.3233/FAIA220346