Deep Learning–Assisted Vaginal Cytology for Estrus Classification in Dogs and Cats

Muruvvet Kalkan MSc[1], Burak Fatih Yuksel[2], Mert Turanli MSc[2], Muhammed Uz[2], and Cahit Kalkan[2]

[1]Ankara Universitesi
[2]Firat Universitesi

May 21, 2025

Abstract

Vaginal cytology is a diagnostic tool for evaluating estrous cycle stages and reproductive health in female dogs and cats. It involves microscopic examination of vaginal epithelial cells, but subjective interpretation can lead to inconsistencies. This study explores artificial intelligence (AI), specifically deep learning, to enhance accuracy. A total of 1,096 vaginal smear samples were collected, stained, digitized, and analyzed using AI. Several pre-trained convolutional neural networks (CNNs), including MobileNetV2, ResNet152V2, EfficientNetV2L, Xception, VGG-16, InceptionV3, NasNetLarge, InceptionResNetV2, DenseNet201, and ConvNeXtSmall, were evaluated. The Xception model achieved the highest accuracy at 97.65%. These findings demonstrate AI's potential to reduce subjectivity, improve diagnostic consistency, and advance reproductive health assessments in veterinary medicine.

.

On simplifying 'incremental remap'-based transport schemes. *J Comput Phys.* 2021;00(00):1–18.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Deep Learning–Assisted Vaginal Cytology for Estrus Classification in Dogs and Cats

Muruvvet Kalkan MSc[1] | Burak Fatih Yuksel PhD[2] | Mert Turanli MSc[2] | Muhammed Uz[2] | Cahit Kalkan PhD[2]

[1]Department of Computer Engineering, Ankara University, Ankara, Turkey

[2]Department of Obstetrics and Gynecology, Firat University, Elazig, Turkey

Correspondence
Corresponding author Muruvvet Kalkan, This is sample corresponding address.
Email: kalkanm@ankara.edu.tr

Present address
This is sample for present address text this is sample for present address text.

Abstract

Vaginal cytology is a diagnostic tool for evaluating estrous cycle stages and reproductive health in female dogs and cats. It involves microscopic examination of vaginal epithelial cells, but subjective interpretation can lead to inconsistencies. This study explores artificial intelligence (AI), specifically deep learning, to enhance accuracy. A total of 1,096 vaginal smear samples were collected, stained, digitized, and analyzed using AI. Several pre-trained convolutional neural networks (CNNs), including MobileNetV2, ResNet152V2, EfficientNetV2L, Xception, VGG-16, InceptionV3, NasNetLarge, InceptionResNetV2, DenseNet201, and ConvNeXtSmall, were evaluated. The Xception model achieved the highest accuracy at 97.65%. These findings demonstrate AI's potential to reduce subjectivity, improve diagnostic consistency, and advance reproductive health assessments in veterinary medicine.

KEYWORDS
Artificial Intelligence, *Deep Learning*, Classification, Estrus Cycle, Vaginal Cytology

1 | INTRODUCTION

Examining vaginal cells under the microscope—commonly referred to as vaginal cytology—offers a straightforward yet valuable tool for evaluating disorders of the reproductive and urinary tracts in both bitches and queens (Kustritz, 2020). Used alongside a thorough clinical history, physical examination, and other diagnostic procedures, this method facilitates accurate diagnoses and tailored treatments. Because female dogs experience a prolonged heat phase and their behavioral signs do not always align with the precise timing of ovulation, determining the best day for breeding can be challenging (Linde & Karlsson, 1984). Consequently, if mating occurs at the wrong time, one might mistakenly suspect fertility complications even when there are none (Grundy, Feldman, & Davidson, 2002; Moxon, Copley, & England, 2010). By assessing vaginal epithelial cells in conjunction with other diagnostic measures, veterinarians can more precisely identify ovulation and optimize the timing for mating. Additionally, vaginal cytology can help ascertain the sexual cycle stage, given in Figure 1, detect irregularities in the cycle or lack of ovulation, localize abnormal bleeding or discharge, and diagnose issues such as inflammation (e.g., metritis and pyometra), tumors, or vaginal hyperplasia (Kaymaz, Rişvanlı, & Köker, 2019).

Vaginal cytology relies on categorizing epithelial cells according to their morphology to determine the reproductive cycle phase (Johnston, Kustritz, & Olson, 2001). Four primary types of epithelial cells are distinguished. Parabasal cells, which lie close to the basement membrane, are small and round, exhibit a prominent nucleus, and have a narrow band of cytoplasm. These cells are consistently found in canine vaginal samples. Positioned above the parabasal layer, intermediate cells appear slightly bigger, with a larger proportion of cytoplasm relative to the nucleus. Collectively, parabasal and intermediate cells are sometimes grouped under the label "non-cornified." As estrogen levels rise, parabasal cells undergo division, giving rise to superficial cells—often called superficial intermediate cells—that are large and irregularly shaped, with abundant cytoplasm and a relatively small nucleus. Some superficial cells lack visible nuclei after staining; these are referred to as non-nucleated squamous cells. Both superficial cells and non-nucleated squamous cells are frequently called "cornified." Other elements typically observed in vaginal cytology include polymorphonuclear leukocytes (PMNs or neutrophils), red blood cells (RBCs), and bacteria (Kustritz, 2020).

During anestrus in bitches, the vaginal lining consists mostly of a thin layer of parabasal and intermediate cells (Post, 1985). As proestrus

Abbreviations: AI, artificial intelligence; AUC, area under the curve; CNN, convolutional neural networks; ROC AUC, area under the curve of receiver operating characteristic

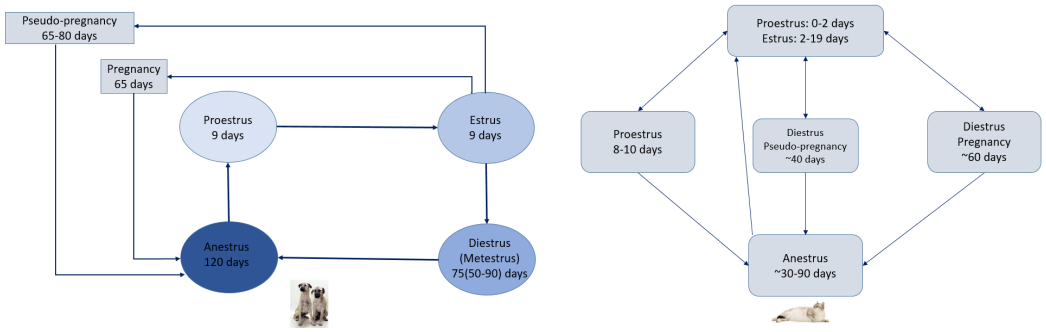

FIGURE 1 Estrus cycle

progresses and transitions into estrus, the tissue thickens and accumulates more layers to prepare for mating. Early smears taken during proestrus often show predominantly intermediate cells with relatively few parabasal or keratinized cells, and red blood cells appear in substantial numbers. As estrus approaches, however, RBCs typically diminish. In estrus itself, the proportion of keratinized cells in the smear climbs, peaking around ovulation as progesterone levels become sufficient for ovulation. By then, both red blood cells and leukocytes generally disappear from the sample, and the smear appears cleaner compared to proestrus or metestrus. When metestrus begins, leukocytes return in greater numbers, and there is a notable increase in parabasal and intermediate cells once again (Moxon et al., 2010). In cats, parabasal, intermediate, nucleated superficial and non-nucleated superficial cells are observed in vaginal cytology. Increased estradiol levels in proestrus cause vaginal cornification, leading to morphologic changes in cytologic cells. In cats in proestrus, intermediate cells are mostly present, while parabasal, neutrophil and cornified cells are present to a lesser extent. In cats, unlike dogs, the peak of vaginal cornification occurs simultaneously with the plasma estradiol peak. In cats in estrus, superficial cells are most abundant, followed by cornified, intermediate, neutrophil and parabasal cells in decreasing order. In diestrus, neutrophils are observed with the majority of parabasal and intermediate cells, while cornified and superficial cells are observed in very small proportions. In anoestrus, intermediate cells are observed at high rates, while parabasal, superficial and neutrophils are observed at low rates (Johnston et al., 2001; Kaymaz et al., 2019). With the advancement of technology, various alternative diagnostic methods are being developed. In vaginal cytology, subjective evaluation of the samples under the microscope may cause differences in interpretation. In order to prevent these subjective evaluations and to make more objective and accurate determinations, computer-aided programs and software are used and continue to be developed. Artificial intelligence, which is a current field of study, is used in different techniques and fields (Matias et al., 2021). By uploading and recognizing images, it has been used in the interpretation of colposcopy, cervical cancers and cervical cytology in humans (Fu et al., 2022; Holmström et al., 2021; G. Liu et al., 2022; Tareef et al., 2017), determining the estrous cycle and stages of rodents (Çeçen et al., 2024; Wolcott et al., 2022), and determining the cycle stage with vaginoscopic images (Rajan, Mooloor Harshan, & Gopinathan, 2024). In this study, it was aimed to help more objective evaluation and diagnosis by determining the cytology images of different cycle stages of cats and dogs using artificial intelligence.

2 | LITERATURE REVIEW

When studies on determining the stage of the estrous cycle of animals with cytological images using deep learning methods are examined, it is possible to say that these methods are much faster than manual methods. In their 2022 review, Hennessey et al. highlight that the application of artificial intelligence (AI) in veterinary medicine remains in its early stages, with fewer than forty academic studies published to date. This emerging approach primarily utilizes machine learning techniques applied to large image datasets for diagnostic purposes. Moreover, further advancements in AI have the potential to enhance areas such as radiology services, workflow optimization, quality control, and image interpretation (Hennessey, DiFazio, Hennessey, & Cassel, 2022).

Calderón and his team propose to automatically identify six cell types in vaginal cytology with 91.6% accuracy to determine the estrous cycle of dogs with a Faster R-CNN-based system. The proposed system reduces the analysis time from approximately 1 hour to a few seconds, speeding up the diagnostic process and making it more efficient. This innovative approach aims to increase accuracy by reducing subjective interpretations in diagnoses and to prevent economic losses (Calderón, Carrillo, Nakano, Acevedo, & Hernández, 2020).

The study conducted by Çeçen and his teammates examines deep learning-based YOLOv5 models to classify the estrous cycle using uterine tissue images taken from female rats. The YOLOv5m model showed the highest performance with 98.3% accuracy and 98% F1 score. The results reveal that the proposed model can support expert pathologists in histological analysis (Çeçen et al., 2024).

In their study, Lodkaew and his colleagues developed a system called CowXNet to automatically detect estrus behavior in cows on farms. CowXNet used YOLOv4 and deep learning methods to analyze camera images and classified the movements of cows with 83% accuracy. This system aims to replace costly electronic devices and enable farmers to

detect estrus more efficiently and effectively (Lodkaew, Pasupa, & Loo, 2023).

In the PhD thesis study conducted by İbrahim Arıkan, he aims to determine the estrus period in farm animals by detecting mounting behavior with deep learning methods. While the ResNet model detected mounting behavior with 99% accuracy, XAI (Explainable Artificial Intelligence) techniques such as Grad-CAM and Gradient Inputs were used to explain the focal points of the models in the decision-making process (the udder and back regions of the cows). The explainability of the models was evaluated with "accuracy," "maximum sensitivity," and "complexity" metrics, providing reliable and understandable results (Arıkan, 2024).

Research conducted by Onishi and colleagues points out that assessing the estrous cycle in adult female mammals is a pivotal step in confirming both the safety and the effectiveness of potential therapeutics. Traditional pathology methods, which often rely on expert assessment, can be time-intensive and prone to variability across different observers. In contrast, deep learning-based image analysis offers significant benefits in streamlining these evaluations. Their findings indicate that two AI-driven models designed to identify estrous cycle phases using cervical and vaginal tissue images achieve accuracies comparable to those of seasoned pathologists, suggesting that this digital approach could expedite the drug research and development pipeline (Onishi et al., 2022).

In Lee's study, it is emphasized that advancements in artificial intelligence are leading to more accurate outcomes in image classification tasks. The researchers conducted a comparative analysis of various machine learning algorithms, including support vector machines (SVM), alongside more advanced deep learning techniques such as convolutional neural networks (CNNs) with varying parameters. These methods were applied to address the classic "Cats vs. Dogs" classification problem (Lee, 2021).

Haghofer and his colleagues introduce a workflow that integrates artificial intelligence (AI) and image processing techniques to classify lymphoma based on nuclear size—categorized as small, medium, or large. Their study highlights the effectiveness of modular segmentation models like Stardist for nuclear segmentation, as well as a Unet++ model trained on labeled nuclear cells from canine lymphoma histological images. Consequently, the proposed workflow achieves classification accuracies of 92% for canine lymphoma data and 84.21% for feline lymphoma data. This system assists pathologists in distinguishing lymphoma subtypes by analyzing nuclear size (Haghofer et al., 2023).

The study conducted by Rajan and his co-researchers presents a contemporary method to determine the phases of the estrous cycle in female dogs. Features were extracted with models of InceptionV3 and ResNet152 and they were optimized with binary gray wolf optimization (BGWO) and classified with extreme gradient boosting (XGBoost) algorithm. The results show that the ResNet152 model performed best with the XGBoost producing 90.37% accuracy in average (Rajan et al., 2024).

The team of Wolcott focuses on that the deep learning-based EstrousNet algorithm was used to classify the estrous stage, and this method achieved expert-level accuracy. EstrousNet uses the time dimension of the hormone cycle to highlight misclassifications and flag anestrus stages (e.g., pseudopregnancy), allowing researchers to quickly assess endocrine status during rodent studies (Wolcott et al., 2022). Pu and his colleagues in their study, an automatic EfficientNet model proposed with deep learning techniques to recognize the estrous cycle of female rats. This model, which provides higher accuracy and efficiency compared to traditional methods, optimizes its performance by adjusting model, layer and input properties, which are depth, width and resolution. The model outcome has high accuracy on predicting stages of the rat estrous cycle, thus increasing the efficiency of experiments and reducing human errors (Pu, Liu, Zhou, & Xu, 2024).

Numerous studies have focused on using deep learning models to analyze the estrus cycle, as highlighted in this section. However, this study is unique in its use of multiple models and a comparative experimental approach, which has not been explored in previous research. By introducing an original dataset and applying this innovative methodology, the study contributes significantly to the literature, offering new insights and advancing the field of estrus cycle analysis. Furthermore, the comparison of different models enhances our understanding of their relative strengths and potential applications, enriching the current body of knowledge in this area.

3 | METHODOLOGY

3.1 | Dataset

A total of 1096 smear images were collected from dogs and cats with healthy genital tract and cycle. Samples were obtained from animals in different phases of the estrous cycle. All samples were collected using a sterile cotton swab moistened with isotonic serum (Pérez, Rodríguez, Dorado, & Hidalgo, 2005). The cotton swab collected cells from the caudodorsal surface of the vagina (Aydin, Sur, Ozaydin, & Dinc, 2011). Swabs were taken by rolling the swab from the dorsal wall of the vagina, removed and spread on a glass slide (Davidson, 2015). The smears were stained using Diff-Quick staining kit (Davidson, 2015; Reckers, Klopfleisch, Belik, & Arlt, 2022). The smears were taken using a camera-integrated microscope (Olympus CX23, LCmicro, Olympus Europa SE & CO. KG, Hamburg, Germany) and digitized, which can be seen in Figure 2.

In the dataset, there are four phases of the estrous cycle of cats and dogs: Anoestrus, Diestrus, Estrus and Proestrus, in Figure 3. In this study, an image processing system was developed and it was aimed to estimate the phase of the estrous cycle of the relevant animal based on an image given to the system.

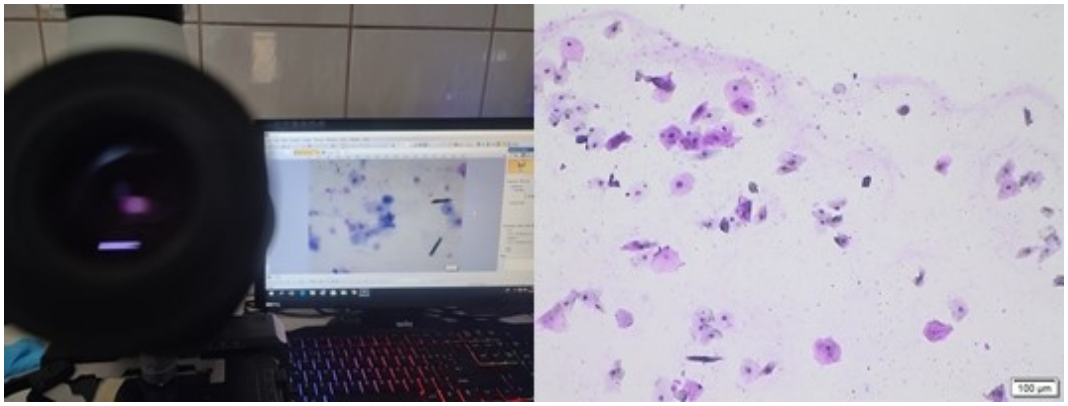

FIGURE 2 Collection of smear images using a camera-assisted microscope.

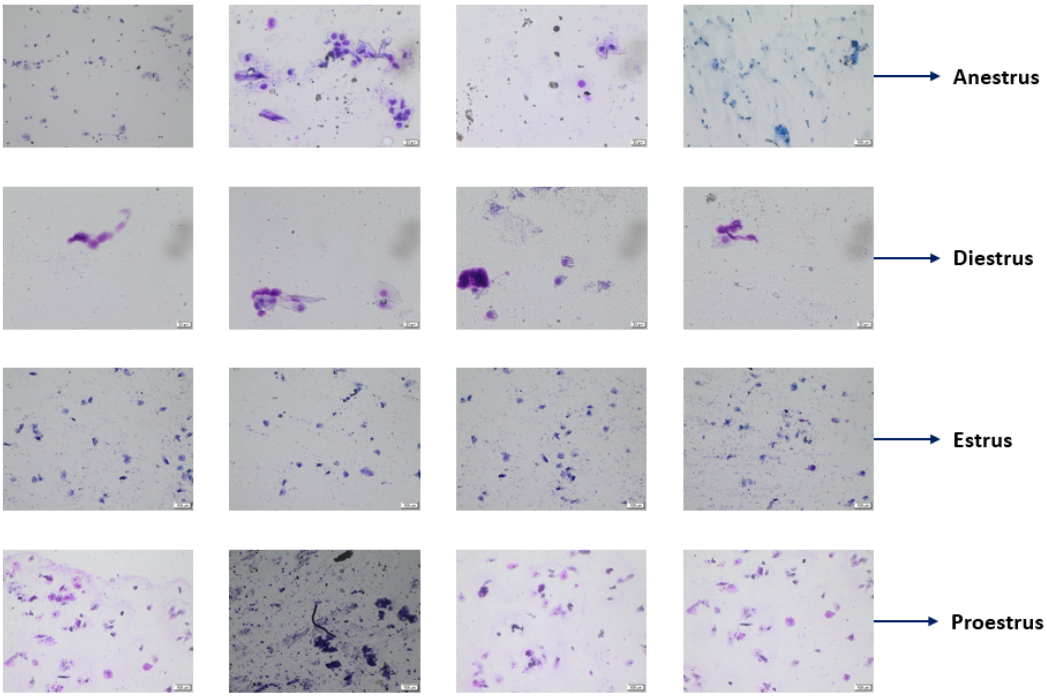

FIGURE 3 Sample images from estrus stages as classes

3.2 | Model Structure

In this experiment, a model is built for the given goal using a pre-trained model as its foundation. The selected base models include MobileNetV2, ResNet-152 V2, EfficientNetV2L, Xception, VGG-16, InceptionV3, NasNet, InceptionResNetV2, DenseNet201, and ConvNextSmall.

MobileNets are specifically designed for robust real-time performance while minimizing accuracy loss (A. G. Howard et al., 2017; A. Howard, Zhmoginov, Chen, Sandler, & Zhu, 2018).

ResNets employ deep architectures with residual mappings to generate reliable predictions (He, Zhang, Ren, & Sun, 2015, 2016).

EfficientNets, derived from MobileNets, achieve greater efficiency by systematically scaling properties such as model depth, layer width etc. (Tan & Le, 2019, 2021).

The Xception model builds upon the Inception model but pushes its design to extremes, earning the name “Extreme Inception” (abbreviated as Xception) (Chollet, 2016).

VGG models, named after the Visual Geometry Group, feature deep architectures with small convolutional layers to address challenges posed by high model depth (Simonyan & Zisserman, 2014).

Inception models employ a unique layer structure consisting of inception blocks, which enable parallel computations followed by concatenation (Christian, Vincent, Sergey, Jonathon, & Zbigniew, 2015).

NasNet (Neural Architecture Search Network) is a model designed to discover the optimal neural network architecture for a specific task. It identifies and optimizes the best-performing model by initially evaluating candidates on a smaller subset of the dataset (Zoph, Vasudevan, Shlens, & Le, 2018).

InceptionResNet models combine the strengths of Inception's architecture with ResNet's residual mappings, often achieving superior results compared to either approach individually (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017).

DenseNet models use densely connected blocks of layers to enhance feature propagation and achieve high accuracy predictions (Huang, Liu, Van Der Maaten, & Weinberger, 2017).

ConvNext models were developed to modernize convolutional neural networks in response to the rise of vision transformers. Designed as next-generation CNNs, ConvNext models come in various sizes, with larger sizes offering higher accuracy at the cost of reduced efficiency. ConvNextSmall was selected for this experiment to balance accuracy and time performance (Z. Liu et al., 2022).

The experiment is carried out for each pre-trained model. In addition to retrieving and utilizing the pre-trained models, several crucial steps are implemented throughout the experiment. The following outlines the experimental process step by step.

The following algorithm describes the key steps involved in the process presented in Algoirthm 1.

1. The dataset is loaded from Google Drive and rescaled to 224x224 size..
2. The dataset of 1,096 images was split into three parts for distinct functions: 70% was designated for training the model, 20% for validating the model during the training process, and the final 10% was reserved as a test set to assess the model's overall performance.
3. A data augmentation layer is implemented to synthetically expand the dataset.By utilizing techniques like flipping, rotation, cropping, and scaling the original collection of 1,096 images can be expanded more than tenfold, greatly increasing the variety of the training dataset.
4. The dataset images are represented as 1D arrays of RGB pixel values, each ranging from 0 to 255. To facilitate faster and more efficient computations, these values are normalized to either the [0,1] or [-1,1] range. The choice of the target range depends on the preprocessing requirements of the specific pre-trained model being used. As a result, an appropriate preprocessing layer is added to ensure compatibility with the model's input format.
5. In this study, transfer learning is utilized by incorporating pre-trained base models into the program through the TensorFlow library, enabling the application to leverage existing knowledge for enhanced performance.
6. Neural network architectures are made up of multiple layers. During the training phase, certain layers are kept inactive, or "frozen," to preserve their pre-trained weights, while the other layers are adjusted. This strategy, commonly called the freeze-out fine-tuning method, was utilized in the experiment by immobilizing approximately the first third of the model's layers. This approach facilitates concentrated training on the active layers while taking advantage of the existing knowledge within the frozen layers.
7. A pooling layer has been implemented in the model. This layer utilizes the global average pooling method, which computes the average value of each feature map. This approach effectively reduces the dimensionality of the data while retaining its essential features.
8. A prediction layer is integrated into the architecture to generate the model's final outputs. To assess the likelihood of input images belonging to each class, the softmax activation function is employed. Softmax is particularly suitable for this task as it transforms the output values into a probability distribution, ensuring that the total probabilities across all classes add up to one.
9. The complete model is built by integrating the base model with additional layers. For each experiment, a different pre-trained model serves as the foundational model, and identical procedural steps are uniformly applied across all seven models. The layer configurations of the primary models are illustrated in Figure 4.

Algorithm 1 Experimental program

```
Fetch the images as dataset
Divide the data into training, validation, and test sets
with approximate proportions of 70%, 20%, and 10%
respectively.

baseModel ← get the pretrained model

model ← define a model

model.layers ← empty layer list

model.layers.insert(inputLayer)
model.layers.insert(dataAugmentationLayer)
model.layers.insert(preprocessLayer)
model.layers.insert(baseModel)
model.layers.insert(globalAveragePoolingLayer)
model.layers.insert(predictionLayer)

baseModel.trainable ← True
fineTuneAt                ← freeazeout last one third of
          layers in the model
for k ∈ {0, ..., fineTuneAt} do
   baseModel.layers[k].trainable ← False
end for
metrics ← [accuracy, loss, precision,
          recall, f1Score, roc]
model.compile(metrics)
Fit the model
Plot the learning curves
Produce results of model on test dataset
```

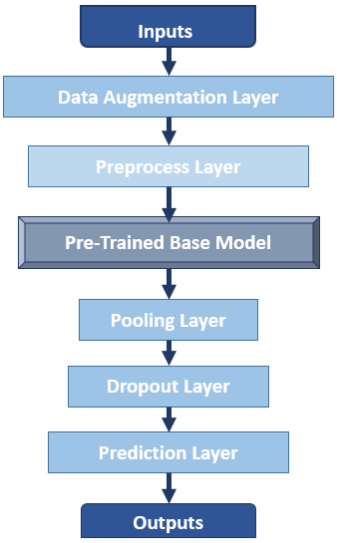

FIGURE 4 Layer structure of main models

10. The model is compiled using categorical cross-entropy as the loss function. Its performance is thoroughly evaluated with metrics including accuracy, precision, recall, F1-score, and the Area Under the Curve of Receiver Operating Characteristic (ROC AUC).
11. The main model training begins with 100 epochs, during which both the training and validation datasets are processed. Throughout the process, the loss and metric outcomes are recorded after each epoch, creating a detailed training history for analysis.
12. Once the training is complete, the recorded metrics and loss values for each epoch are visualized as learning curves. This visualization is highly effective for understanding how the CNN model learns over time, showcasing trends in learning curve and helping identify areas for potential improvement.
13. To assess the performance of the trained primary model, it is evaluated using the test dataset. The loss function and evaluation metrics applied during training are also utilized in this evaluation. The final evaluation results provide a basis for comparing the performance of different main models, offering insight into their effectiveness in handling the task at hand.
14. In the final step, a selection of predictions made by the model is presented alongside the corresponding input images. These predictions are visualized graphically, providing a clear representation of the model's output for comparison with the original images. This provides a clearer insight into the model's performance and prediction accuracy. Examples are illustrated in Figure 5.

4 | EXPERIMENTAL RESULTS

All experiments in this study were conducted within the Google Colab cloud environment. The experiments utilized a Google Colab TPU V2 processor, a specialized unit engineered to accelerate the matrix computations fundamental to neural network processing.

Data from various estrus stages of both cats and dogs were categorized into four distinct classes. These classes were randomly divided into three subsets: training, validation, and testing. The testing subset consisted of unbiased data that was excluded from the training phase, ensuring an objective evaluation of the study's results. In the classification phase, the softmax activation function was employed in the final prediction layer. This function calculates one-dimensional vectors where the length of each vector corresponds to the number of classes—in this case, four. Each element in the vector represents the probability that the input belongs to a specific class, ranging from 0 to 1, with the sum of all probabilities equal to one. The highest value in the vector determines the predicted class. Thus, the two-dimensional input images are ultimately converted into a single numerical value, class number, serving as a label to indicate their respective class.

A total of ten classification models were developed and analyzed for this study. To identify the best-performing model, six evaluation metrics were utilized: accuracy, cross-entropy loss, precision, F1 score, recall, and ROC AUC.

The computed metrics for each model are thoroughly evaluated and further discussed in detail in the Discussion section.

Accuracy refers to the proportion of correct predictions out of the total predictions made by a classification model. As outlined in the equation, it provides insight into how effectively the model assigns labels to the input data, as illustrated in Equation 1. This metric is particularly significant when comparing the performance of different models.

$$Accuracy = \frac{TN + TP}{TP + TN + FN + FP} \tag{1}$$

Precision is the proportion of true positives (TP) accurately identified by the classification model out of all positive predictions (TP + FP). It specifically reflects the model's effectiveness in correctly detecting and classifying positive instances, as demonstrated in Equation 2.

$$Precision = \frac{TP}{FP + TP} \tag{2}$$

Recall quantifies the fraction of true positives (TP) accurately identified by the classification model out of all actual positives (TP + FN) in the dataset, as illustrated in Equation 3.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F-Scores are especially crucial when balancing precision and recall metrics is essential. In this context, the F1 score is introduced, representing the harmonic mean of precision and recall. The F1 score is particularly useful for minimizing incorrect predictions and providing a balanced assessment of the model's performance, as shown in Equation 4.

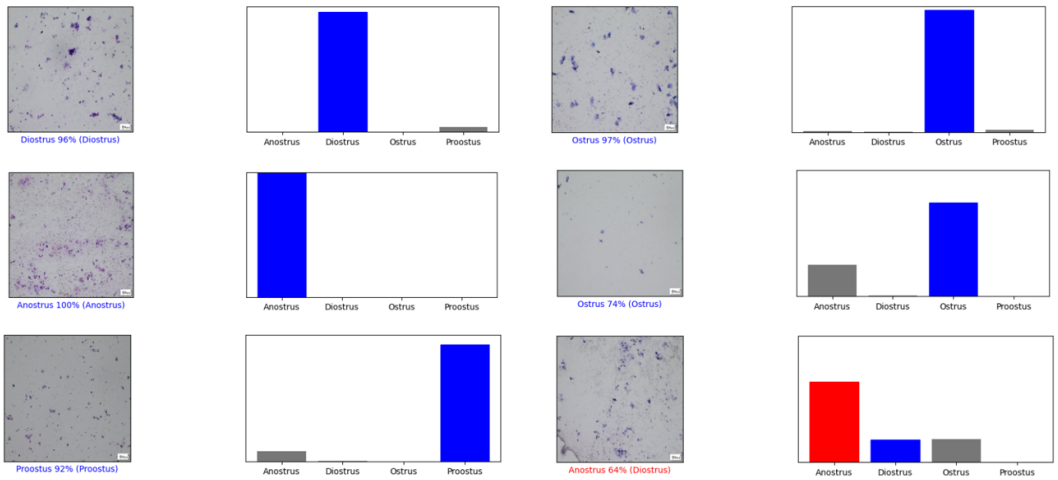

FIGURE 5 Samples of prediction results

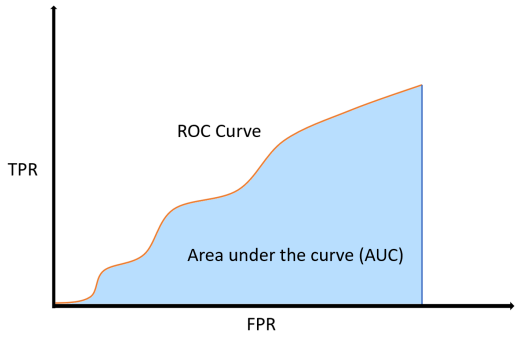

FIGURE 6 ROC curve and its AUC

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4}$$

The AUC ROC is a vital measure for assessing the effectiveness of a classification model. The ROC curve depicts the balance between true positive rates and false positive rates, while the AUC quantifies the area under this curve, as seen in Figure 6. In contrast to accuracy metrics, AUC of ROC provides valuable insights into the model's ability to distinguish between positive and negative classes effectively.

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + FN} \tag{6}$$

Cross-entropy loss is a logarithmically calculated metric that assesses the difference between the predicted probability distribution and the actual distribution of target classes. It measures how well the predicted probabilities align with the true labels by imposing higher penalties for incorrect and confident predictions. A lower cross-entropy loss value signifies that the model's predictions are closer to the true labels, indicating better performance and greater accuracy in its predictions.

Model training is conducted in epochs, with each epoch producing metric results that allow us to observe the learning process. Consequently, we plotted the learning curves for each model on both the training and validation datasets based on these epoch-wise accuracy results, as detailed below.

The learning curves of MobileNetV2 demonstrate rapid improvement on the training dataset up to approximately 60%, after which progress slows significantly. On the validation dataset, performance is inferior, plateauing around 80%, as shown in Figure 7. ResNet152V2 exhibits learning curves similar to MobileNetV2 but with a slower initial phase and slightly better validation performance, reaching nearly 90% accuracy, as depicted in Figure 7. EfficientNetV2L shows a sharp increase in accuracy and is one of the most stable models in terms of curve oscillations, as illustrated in Figure 8. The learning curves of Xception follow a trend similar to EfficientNetV2L. However, the validation curves experience more significant downward fluctuations and do not achieve accuracy levels close to 100%, as shown in Figure 8. The VGG-16-based model appears to be the slowest learner, is the only model with marginal oscillations on the training dataset, and its validation accuracy barely reaches 80%, as presented in Figure 9. InceptionV3, NasNetLarge, and InceptionResNetV2 present roughly similar trends and shapes with minor differences. They exhibit low oscillations on the validation dataset and steady-paced learning on both datasets, with gently curving lines on average. Notably, NasNetLarge's curves reach higher accuracy in earlier epochs, as shown in Figures 9 and 10.

DenseNet-201's learning curves show a reasonable increase during training. However, the validation curve experiences the most oscillations and only barely reaches 90% accuracy, as illustrated in Figure 11. ConvNextSmall achieves the fastest early learning rate during training without a doubt but performs poorly on validation, with an upper bound near 80%. Consequently, it exhibits the greatest discrepancy between training and validation performance, sharing only one common point, as depicted in Figure 11.

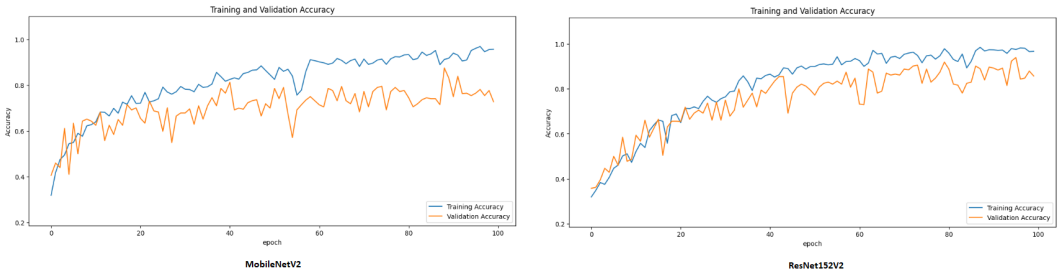

FIGURE 7 Accuracy trends of the MobileNetV2 and ResNet152V2 models across each training epoch, illustrated through learning curves

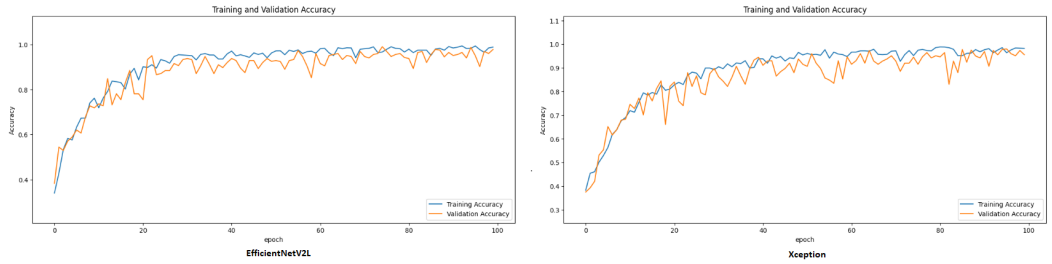

FIGURE 8 Accuracy trends of the EfficientNetV2L and Xception models across each training epoch, illustrated through learning curves

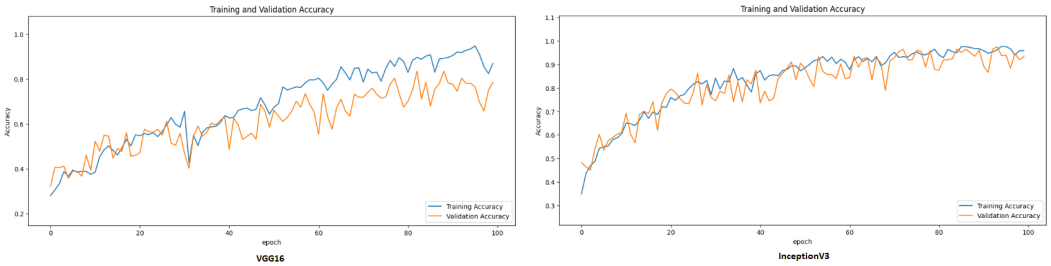

FIGURE 9 Accuracy trends of the VGG-16 and InceptionV3 models across each training epoch, illustrated through learning curves

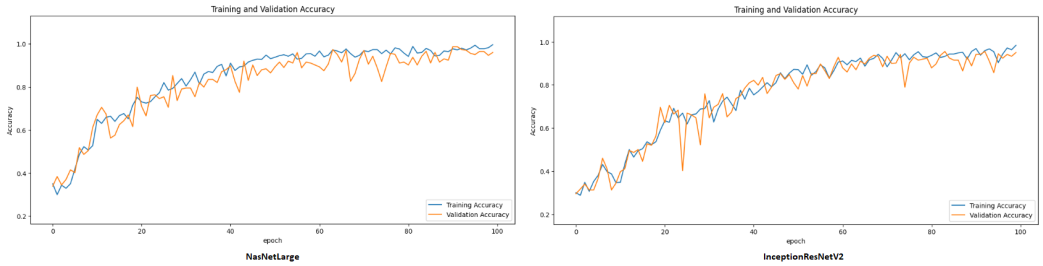

FIGURE 10 Accuracy trends of the NasNetLarge and InceptionResNetV2 models across each training epoch, illustrated through learning curves

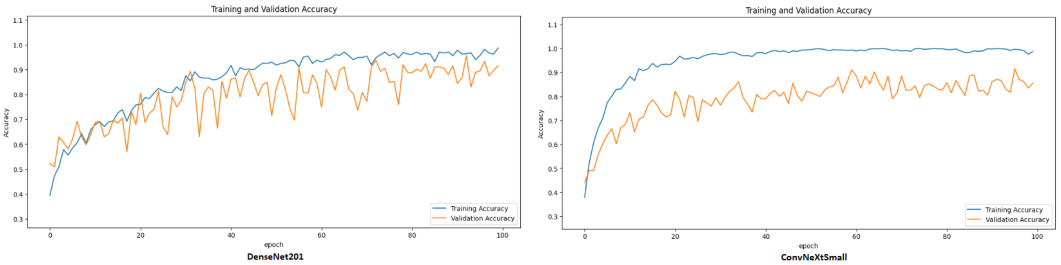

FIGURE 11 Accuracy trends of the DenseNet201 and ConvNeXtSmall models across each training epoch, illustrated through learning curves

TABLE 1 Classification Performance Metrics

Base Model	Accuracy	Loss	Precision	Recall	F1-Score	ROC AUC
MobileNetV2	78.12%	1.09	79.03%	79.16%	79.10%	92.52%
ResNet152V2	91.40%	0.22	93.06%	92.68%	92.87	99.58%
EfficientV2L	94.53%	0.13	93.86%	94.94%	94.40%	99.71%
Xception	**97.65%**	**0.08**	**98.64%**	**97.39%**	**98.01%**	**99.91%**
VGG-16	87.50%	0.40	87.18%	87.32%	87.25%	97.27%
InceptionV3	93.75%	0.11	94.45%	93.64%	94.04%	99.82%
NasNetLarge	95.31%	0.25	96.59%	94.87%	95.72%	98.78%
InceptionResNetV2	96.09%	**0.08**	97.33%	94.29%	95.79%	99.88%
DenseNet201	92.30%	0.20	92.68%	92.44%	92.56%	99.75%
ConvNeXtSmall	89.42%	0.38	90.44%	88.11%	89.26%	97.13%

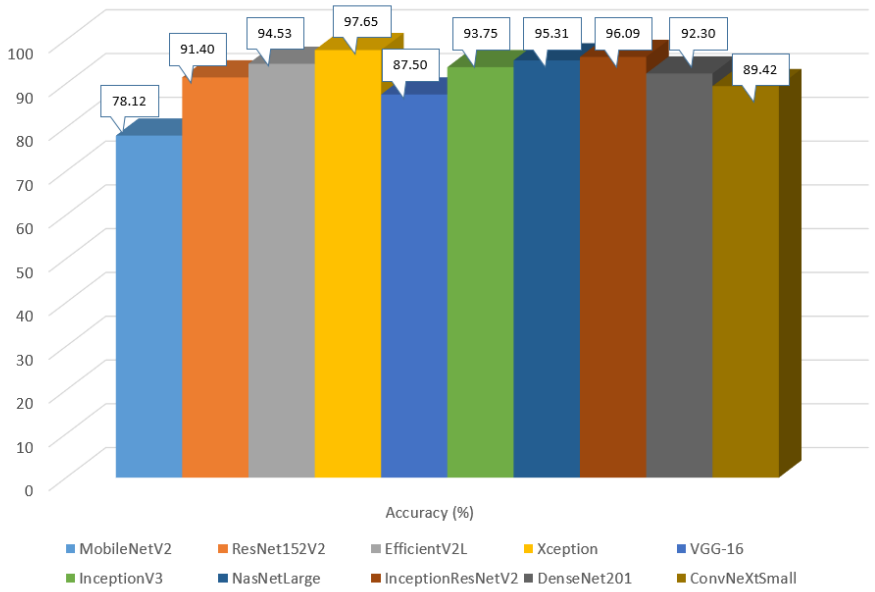

FIGURE 12 Classification accuracy metrics from the final evaluation are displayed in a bar graph

Overall, with the exceptions noted above, all models' learning curves exhibit certain common behaviors. Specifically, the validation curves are consistently more oscillatory and lower in accuracy compared to the training curves, which is expected. Both training and validation curves naturally display an overall increasing trend on average. Unless stated otherwise above, both curves can reach approximately 100% accuracy at some point.

In evaluating the results, all metrics derived from the test dataset were carefully analyzed in the study. The calculated metrics provided objective performance assessments based on the respective models. The Xception model achieved the highest values, with an accuracy of 97.65%, precision of 98.64%, recall of 97.39%, F1 score of 98.01%, and an impressive ROC AUC of 99.91%. In terms of loss, InceptionResNetV2 produced the lowest value, with a result of 0.08. Overall, the Xception model proved to be the most effective, achieving the highest results in this study. For a detailed comparison, refer to Table 1 along with Figures, 12, and 13.

5 | DISCUSSION

As presented in the preceding section on Experimental Results, each pre-trained model yielded distinct outcomes across various evaluation metrics, exhibiting significant disparities. Therefore, it is necessary to explicitly articulate the reasons behind these differences.

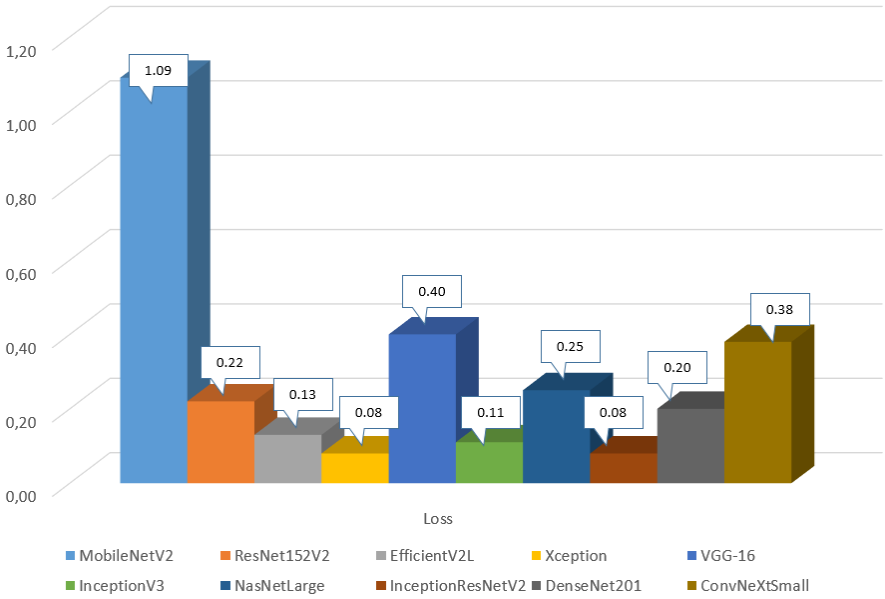

FIGURE 13 Classification loss metrics from the final evaluation are displayed in a bar graph

The MobileNetV2-based model achieved the lowest scores across all metrics. Given the high resource and time demands of training deep learning models, MobileNets are specifically designed to address these issues by sacrificing some predictive accuracy. Although MobileNetV2 is fast, its underperformance is expected, placing it last among the selected models with an accuracy of 78%.

The ResNet152V2 model ranks near the middle, albeit on the lower side across all metrics. Despite its moderate placement, an accuracy of 91.53%, a loss of 0.22, and all other results exceeding 90% indicate that it is a successful model for the task overall. However, superior models are undeniably present in the experiment.

EfficientNetV2L maintains a position comparable to ResNet152V2 in terms of metric rankings, yet it achieves a 3% higher accuracy. While a three percent increase might appear modest, it signifies a substantial difference when accuracy percentages surpass 90%, where even a single percentage point can be impactful. Additionally, EfficientNetV2L boasts the second-highest recall; however, it has lower precision and a less favorable balance between precision and recall, as indicated by its F1-score. Nevertheless, with an accuracy of 94.53%, EfficientNetV2L demonstrates that its model scaling approach yields successful results on this dataset.

Xception, short for eXtreme Inception, outperforms all other models across all metrics, boasting an accuracy of 97.65% and an excellent loss of just 0.08. Extending the Inception architecture to its extreme clearly represents the most effective approach in this study, particularly when compared to InceptionV3 results.

VGG-16 is one of the older models, featuring deeper layers compared to MobileNetV2. Consequently, it ranked second to last across all metrics, including an accuracy of 87.5%, except for ROC AUC, where it was third to last. This performance is attributable to its outdated architecture, especially when contrasted with the more contemporary models employed in this study, resulting in an expected outcome.

The InceptionV3-based model ranks just below EfficientNetV2L, achieving an accuracy of 93.75%. Although its performance in accuracy, precision, recall, and F1-score is positioned near the middle compared to its peers, it secures a place within the top three for loss and ROC AUC, alongside InceptionResNetV2 and Xception. The vanilla Inception architecture exhibits limitations when compared to some non-Inception-based models. However, it demonstrates significant potential when evaluating loss and ROC AUC—metrics that measure the discrepancy between predictions and actual results, and the overall quality of the prediction model, respectively. This potential is validated by the performance of modified Inception models: Xception, which is the best, and InceptionResNetV2, the second best.

NasNetLarge, as a hypermodel architecture, delivers strong performance across key metrics, including accuracy, precision, recall, and F1-score. Despite being the third most accurate model with an accuracy of 95.31%, it shows signs of lagging behind in loss and ROC AUC. Nevertheless, NasNetLarge demonstrates one of the most refined results among model-building architectures.

The InceptionResNetV2 model, achieving an accuracy of 96.09%, is based on the Inception architecture. While it shares a similar foundational structure with Xception, it uniquely integrates ResNet's residual mapping approach, leading to outstanding performance. This excellence is demonstrated by its shared first place with Xception in loss, second place across all other metrics, and a fourth-place ranking in recall. As previously mentioned, modified Inception models emerge as the top performers in this experiment.

DenseNet201 is typically a powerful architecture for image classification tasks; however, in this project, it only secured mediocre rankings, falling into the lower half for all metrics except ROC AUC, where it ranked fifth. Although an accuracy of 92.3% is commendable, it trails behind competing models in this experiment. The dense layer block structures inherent to DenseNets resulted in unexpectedly lower rankings. Therefore, it can be concluded that DenseNet201 is not the most suitable model for classifying the estrus cycle with this dataset.

ConvNextSmall emerged as another underperforming model, with an accuracy of 89.42%. This is particularly notable given its design to keep pace with next-generation deep learning techniques, such as vision transformers. ConvNextSmall typically ranked third to last across all metrics. Consequently, similar to DenseNet201, ConvNextSmall is outperformed by more suitable candidates for the experimental task at hand.

When examining the learning curves, several key observations emerge. First, the learning speed—characterized by a steep increase to higher percentages—does not necessarily influence the test results. For instance, NasNetLarge exhibits a slower learning rate compared to EfficientNetV2L, yet their outcomes are contrasting. Secondly, validation curves offer crucial insights into test performance. Models with validation curves that display fewer oscillations tend to be more successful. Additionally, the validation upper bound serves as an indicator of potential test results; a higher upper bound is associated with greater test accuracy. These patterns are consistently observed across all models, particularly when comparing the top performer, Xception, with the lowest performer, MobileNetV2.

Overall, the experiment can be deemed successful due to the exceptionally high accuracy results, notably the 97.65% achieved by Xception. Excluding three models—one scoring below 80% and two just below 90%—all other models achieved accuracies above 90%, with some even exceeding 95%. By applying contemporary AI technologies to the veterinary domain, the determination of the estrus cycle can be significantly facilitated.

6 | CONCLUSION

In this study, images from four different estrus periods of cats and dogs were classified using various deep learning models. The accuracy values obtained were used to compare the performance of each model in terms of classification. The results indicate that the Xception model achieved the highest accuracy, with a remarkable 97.65%, demonstrating its effectiveness in estrus period classification.

For future research, expanding the dataset with larger and more diverse image collections can significantly enhance the generalization capability of the models. This can be particularly beneficial by incorporating images captured under different environmental and lighting conditions, as well as including a broader range of cat and dog species. Furthermore, applying techniques such as model optimization and transfer learning could improve model accuracy. Transfer learning, in particular, can accelerate the training process and yield better results, even with smaller datasets, by leveraging pre-trained models. Additionally, ensemble learning methods, which combine the strengths of multiple models, could potentially achieve even higher accuracy levels. To increase the practicality of these methods, future work could focus on developing models optimized for real-time classification and tailored for mobile devices, facilitating their implementation in clinical applications.

AUTHOR CONTRIBUTIONS

Muruvvet Kalkan: Conceptualization of this study, Methodology, Software, Writing - Original draft preparation. **Burak Fatih Yuksel:** Conceptualization of this study, Data curation, Writing - Original draft preparation. **Mert Turanli:** Data curation, validation, writing—review and editing. **Muhammed Uz:** Data curation, validation, writing—review and editing. **Cahit Kalkan:** Conceptualization of this study, writing—review and editing, supervision.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

References

Arıkan, İ. (2024). *Estrus detection in cows with deep learning techniques* (Unpublished doctoral dissertation). İzmir Institute of Technology.

Aydin, I., Sur, E., Ozaydin, T., & Dinc, D. A. (2011). Determination of the stages of the sexual cycle of the bitch by direct examination. *Journal of Animal and Veterinary Advances*, *10*(15), 1962–1967.

Calderón, G., Carrillo, C., Nakano, M., Acevedo, J., & Hernández, J. E. (2020). Automatic estrus cycle identification system on female dogs based on deep learning. In *Mexican conference on pattern recognition* (pp. 261–268).

Çeçen, Ş., Çeribaşı, S., Erkuş, M., Özer, A. B., Tuncer, T., & Çınar, A. (2024). Classification of estrus cycles in rats by using deep learning. *Traitement du Signal*, *41*(1).

Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions. *CoRR*, *abs/1610.02357*. doi: 10.48550/arXiv.1610.02357

Christian, S., Vincent, V., Sergey, I., Jonathon, S., & Zbigniew, W. (2015). Rethinking the inception architecture for computer vision. *CoRR*, *abs/1512.00567*. doi: 10.48550/arXiv.1512.00567

Davidson, A. (2015). Determining canine estrus stage via vaginal cytology. *Clinician's Brief*, *13*(5), 19–20.

Fu, L., Xia, W., Shi, W., Cao, G.-x., Ruan, Y.-t., Zhao, X.-y., … Gao, X. (2022). Deep learning based cervical screening by the cross-modal integration of colposcopy, cytology, and hpv test. *International Journal of Medical Informatics*, *159*, 104675.

Grundy, S. A., Feldman, E., & Davidson, A. (2002). Evaluation of infertility in the bitch. *Clinical techniques in small animal practice*, *17*(3), 108–115.

Haghofer, A., Fuchs-Baumgartinger, A., Lipnik, K., Klopfleisch, R., Aubreville, M., Scharinger, J., ... Bertram, C. A. (2023). Histological classification of canine and feline lymphoma using a modular approach based on deep learning and advanced image processing. *Scientific Reports*, *13*(1), 19436.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, *abs/1512.03385*. doi: 10.48550/arXiv.1512.03385

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. *CoRR*, *abs/1603.05027*. doi: 10.48550/arXiv.1603.05027

Hennessey, E., DiFazio, M., Hennessey, R., & Cassel, N. (2022). Artificial intelligence in veterinary diagnostic imaging: A literature review. *Veterinary Radiology & Ultrasound*, *63*, 851–870.

Holmström, O., Linder, N., Kaingu, H., Mbuuko, N., Mbete, J., Kinyua, F., ... others (2021). Point-of-care digital cytology with artificial intelligence for cervical cancer screening in a resource-limited setting. *JAMA network open*, *4*(3), e211740–e211740.

Howard, A., Zhmoginov, A., Chen, L.-C., Sandler, M., & Zhu, M. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. In *Proc. cvpr* (pp. 4510–4520).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. doi: 10.48550/arXiv.1704.04861

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

Johnston, S. D., Kustritz, M. V., & Olson, P. S. (2001). *Canine and feline theriogenology*. Saunders.

Kaymaz, M., Rişvanlı, A., & Köker, A. K. (2019). *Kedilerde doğum ve jinekoloji* (2nd ed.). Malatya: Medipres.

Kustritz, M. V. R. (2020). Vaginal cytology in the bitch and queen. *Veterinary cytology*, 552–558.

Lee, Y. (2021). Image classification with artificial intelligence: cats vs dogs. In *2021 2nd international conference on computing and data science (cds)* (pp. 437–441).

Linde, C., & Karlsson, I. (1984). The correlation between the cytology of the vaginal smear and the time of ovulation in the bitch. *Journal of Small Animal Practice*, *25*(2), 77-82.

Liu, G., Ding, Q., Luo, H., Sha, M., Li, X., & Ju, M. (2022). Cx22: A new publicly available dataset for deep learning-based segmentation of cervical cytology images. *Computers in Biology and Medicine*, *150*, 106194.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *CoRR*, *abs/2201.03545*. doi: 10.48550/arXiv.2201.03545

Lodkaew, T., Pasupa, K., & Loo, C. K. (2023). Cowxnet: An automated cow estrus detection system. *Expert Systems with Applications*, *211*, 118550.

Matias, A. V., Amorim, J. G. A., Macarini, L. A. B., Cerentini, A., Onofre, A. S. C., Onofre, F. B. D. M., ... von Wangenheim, A. (2021). What is the state of the art of computer vision-assisted cytology? a systematic literature review. *Computerized Medical Imaging and Graphics*, *91*, 101934.

Moxon, R., Copley, D., & England, G. (2010). Quality assurance of canine vaginal cytology: A preliminary study. *Theriogenology*, *74*(3), 479-485.

Onishi, S., Egami, R., Nakamura, Y., Nagashima, Y., Nishihara, K., Matsuo, S., ... others (2022). Digital workflows for pathological assessment of rat estrous cycle stage using images of uterine horn and vaginal tissue. *Journal of Pathology Informatics*, *13*, 100120.

Pérez, C., Rodríguez, I., Dorado, J., & Hidalgo, M. (2005). Use of ultrafast papanicolaou stain for exfoliative vaginal cytology in bitches. *The Veterinary Record*, *156*(20), 648–650. doi: 10.1136/vr.156.20.648

Post, K. (1985). Canine vaginal cytology during the estrous cycle. *The Canadian veterinary journal*, *26*(3), 101.

Pu, X., Liu, L., Zhou, Y., & Xu, Z. (2024). Determination of the rat estrous cycle vased on efficientnet. *Frontiers in Veterinary Science*, *11*, 1434991.

Rajan, B. K., Mooloor Harshan, H., & Gopinathan, V. (2024). Automated detection of reproductive stages of female canine from vaginoscopic images. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, 1-13.

Reckers, F., Klopfleisch, R., Belik, V., & Arlt, S. (2022). Canine vaginal cytology: a revised definition of exfoliated vaginal cells. *Frontiers in Veterinary Science*, *9*, 834031.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, *abs/1409.1556*. doi: 10.48550/arXiv.1409.1556

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31, pp. 4278–4284). doi: 10.1609/aaai.v31i1.11231

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, *abs/1905.11946*. doi: 10.48550/arXiv.1905.11946

Tan, M., & Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. arxiv 2021. *CoRR*, *abs/2104.00298*. doi: 10.48550/arXiv.2104.00298

Tareef, A., Song, Y., Huang, H., Wang, Y., Feng, D., Chen, M., & Cai, W. (2017). Optimizing the cervix cytological examination based on deep learning and dynamic shape modeling. *Neurocomputing*, *248*, 28–40.

Wolcott, N. S., Sit, K. K., Raimondi, G., Hodges, T., Shansky, R. M., Galea, L. A., ... Goard, M. J. (2022). Automated classification of estrous stage in rodents using deep learning. *Scientific reports*, *12*(1), 17685.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8697–8710).