 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

CONSISTENT AND EFFECTIVE METHOD TO DEFINE THE MOUSE ESTROUS CYCLE STAGE BY DEEP LEARNING BASED MODEL

Strauss L[1], Junnila A[1], Wärri A[1], Manti M[2], Jiang Y[3], Löyttyniemi E[4], Stener-Victorin E[2], Lagerquist MK[3], Kukoricza K[1], Heinosalo T[1], Blom S[5] and Poutanen M[1,3]

[1] Institute of Biomedicine, Research Centre for Integrative Physiology and Pharmacology, and Turku Center for Disease Modeling, University of Turku, Turku, Finland

[2] Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden

[3]Sahlgrenska Osteoporosis Centre, Centre for Bone and Arthritis Research at Institute of Medicine, Sahlgrenska Academy at University of Gothenburg, Sweden

[4] Department of Biostatistics, University of Turku, Turku, Finland

[5] Aiforia Technologies Oyj, Pursimiehenkatu 29-31 D, 00150 Helsinki, Finland

Corresponding author:

Matti Poutanen, University of Turku, Institute of Biomedicine, Kiinamyllynkatu 10, 20720 Turku, Finland, matpou@utu.fi

A short title: Defining mouse estrous cycle by deep learning

Keywords: estrogen, progesterone, female reproduction, ovarian function

Word count: 3575

Abstract

The mouse estrous cycle is divided into four stages: proestrus (P), estrus (E), metestrus (M) and diestrus (D). The estrous cycle affects reproductive hormone levels in a wide variety of tissues. Therefore, to obtain reliable results from female mice, it is important to know the estrous cycle stage during sampling. The stage can be analyzed from a vaginal smear under a microscope. However, it is time-consuming, and the results vary between evaluators. Here, we present an accurate and reproducible method for staging the mouse estrous cycle in digital whole slide images (WSIs) of vaginal smears. We developed a model using a deep convolutional neural network (CNN) in a cloud-based platform, Aiforia Create. The CNN was trained by supervised pixel-level multiclass semantic segmentation of image features from 171 hematoxylin-stained samples. The model was validated by comparing the results obtained by CNN with those of four independent researchers. The validation data included three separate studies comprising altogether 148 slides. The total agreement attested by the Fleiss kappa value between the validators and the CNN was excellent (0.75), and when D, E and P were analyzed separately, the kappa values were 0.89, 0.79 and 0.74, respectively. The M stage is short and not well defined by the researchers. Thus, identification of the M stage by the CNN was challenging due to the lack of proper ground truth, and the kappa value was 0.26. We conclude that our model is reliable and effective for classifying the estrous cycle stages in female mice.

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

Introduction

The estrous cycle in female mice, like a menstrual cycle in women, is induced by reproductive hormones whose levels vary greatly during the cycle (Nilsson et al., 2015). Recording the stage of the estrous cycle is an effective way to study the functional status of the hypothalamus-pituitary-ovarian axis, as the estrous cycle is vulnerable to any hormonal or other changes affecting reproduction (Kemiläinen et al., 2016, Roa et al., 2022, Calanni-Pileri et al., 2022). It is also well-established that sex steroids have a crucial role in reproductive functions. However, sex steroids have a wide range of effects also in non-reproductive tissues, for example, on the brain, bone, liver, lung, adipose tissue, and skeletal muscle, as well as on the immune system (Vanderschueren et al., 2014, Fuentes et al., 2018, Elderman et al., 2018, Jaric et al., 2019, Konstandi et al., 2020, Mayneris-Perxachs et al., 2020). Therefore, it is crucial to know the estrous cycle stage of the female mouse, e.g., during tissue sampling, regardless of the organ system to be studied.

The estrous cycle has been shown to modulate several different biological functions, such as white adipose tissue browning (Hua et al., 2018) and the impact of multiple acute stresses on spatial memory (Hokenson et al., 2021). It is also well documented that in some breast cancer models, the stage of the estrous cycle at the time of carcinogen administration or at the time of cancer transplantation is critical for tumor incidence, latency, and growth rate (Wood and Hrushesky, 2005). Interestingly, the dependency of the tumor growth on the estrous cycle is not limited to breast cancer but also occurs in a chemically induced, non-hormonally dependent sarcoma model (Wood and Hrushesky, 2005). A need for a feasible and reliable method to define the estrous cycle stages is, thus, evident.

The mouse estrous cycle is divided into four stages: proestrus (P), estrus (E), metestrus (M) and diestrus (D). One cycle lasts 4-5 days, whereas diestrus is the longest stage, about 2 days (Byers et al., 2012). The most accurate method to study the stage of the estrous cycle in mice

is carried out by microscopic evaluation of the cells present in a vaginal smear spread on a microscopic slide. Although the smears can be evaluated as unstained wet mount preparations, the most exact results are obtained from samples that are fixed and stained with histological stains such as hematoxylin or Giemsa. The evaluation of the stage is based on the determination of the relative amount of three different cell types in the vaginal smears: neutrophils, immature nucleated squamous epithelial cells, and anucleated cornified epithelial cells. The relative amount of different cell types reflects the changes in cell typology within the murine vaginal canal (Cora et al., 2015) that is cycle-dependent.

The visual evaluation of the estrous cycle stage suffers from intra- and inter-observer variation. This well-known issue occurs because humans tend to focus on a small area that may not be representative, and quantifying all cell types on a slide is practically impossible. Other recent publications have also described the problem (Sano et al., 2020; Wolcott et al., 2022). Furthermore, the categorical classification of the estrous cycle stage by human observers typically reflects poorly the continuous progression of the estrous cycle from stage to stage. Therefore, a more quantitative determination of all the cells present in the vaginal smear by an automated computer-assisted method would provide extra benefit and would better identify the transitional stages. Automating the analysis utilizing artificial intelligence (AI) would speed up projects with large cohorts and increase the results' reliability and reproducibility by abolishing the ambiguity caused by different evaluators. In clinical diagnostics, AI-based digital pathology has already shown its power, for example, in extracting disease-specific information that humans cannot quantify. The AI-based method for selecting breast cancer patients who could benefit from immunotherapy from histological samples is an excellent example. (Corredor et al., 2023).

Here, we report an accurate and reproducible AI-based model for estrous cycle determination. The model is based on a deep convolutional neural network (CNN) developed with the Aiforia software (Aiforia Technologies, Helsinki, Finland). The developed CNN determines the specific estrous cycle stage in digital whole slide images (WSIs) of fixed and hematoxylin-

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

stained vaginal smears. The model produces quantitative data on the stage-specific histological features of the vaginal smears at the time of sampling. Based on data provided by CNN, investigators can accurately follow the changes in the mouse estrous cycle and, consequently, track possible aberrations of the normal cycle.

Material and Methods

Mouse models

For training and validating the CNN, vaginal smears were collected from fertile female mice of C57BL/6N and NMRI genetic background. The mice were given SDS CRM chow diet (Special Diets Services, Witham, UK) and tap water *ad libitum* and housed in specific pathogen-free conditions at Central Animal Laboratory, University of Turku. All animal handling in the study was approved by the Finnish Animal Ethics Committee and the Institutional animal care policies of the University of Turku (Turku, Finland), which fully meet the requirements as defined in the Guide for the Care and Use of Laboratory Animals (Committee for the Update of the Guide for the Care and Use of Laboratory Animals, 2011).

Determination of the estrous cycle stage from vaginal smears

To determine the estrous cycle stage, vaginal smears from female mice were collected in PBS. The sampling was always performed at the same time, i.e., 8:00-9:00 am. The vaginal smears were air-dried on microscopic slides, fixed with ethanol, and stained with Mayer's hematoxylin, and the stage of the cycle was determined based on the cytology. The relative amount of different cell types defines the specific estrous stage as follows: In proestrus (P), most of the cells are nucleated immature vaginal epithelial cells. In estrus (E), the epithelial cells are prominently cornified and without a nucleus. In metestrus (M), cornified epithelial cells are still present, while a few leucocytes and isolated nucleated epithelial cells appear. In diestrus (D),

the leucocytes are the most common cell type, and also, a few nucleated epithelial cells are present (Figure 1) (Byers et al., 2012).

Digitalization of the slides

Slides were digitized using a Pannoramic P1000 Flash digital slide scanner (3DHISTECH Ltd., Budapest, Hungary) with 20X magnification, 0.24 µm/pixel resolution, and 20X/0.8 NA objective to create digital whole-slide images (WSIs). The WSIs were compressed using a 50% jpeg compression and uploaded to Aiforia software (Aiforia Technologies).

Training, verification, and validation of the convolutional neural network

The CNN training was based on fully supervised pixel-level multiclass semantic segmentation of image features. Thus, each pixel belonged to one of the semantic classes: proestrus, estrus, metestrus or diestrus. Training of the CNN was performed using cloud-based Aiforia software. The training set included 171 hematoxylin-stained vaginal smear WSIs (link to the slides below). To increase the robustness of the CNN, diverse training data, i.e. slides with different cell densities and color intensities, prepared of mice with two different genetic background, were included in the training data. The ground truth was initially established by an experienced researcher who trained the CNN following the criteria described in two previous articles (Byers et al., 2012; Cora et al., 2015). However, in unclear cases, all four researchers who tested the CNN discussed, and the consensus between researchers established the ground truth. In Aiforia, the ground truth was established by freehand manual annotations of representative features and image areas (2-3 areas/slide) (Figure 2). Image artefacts (including blurry areas, folds, dust, etc.) were annotated as background to train the CNN to automatically classify and exclude the artefacts in the analysis. After each training, a small set of slides were analyzed using the model, and the trainer validated the results. Based on the validation, training annotations were corrected and added and parameters for digital augmentation were adjusted. Altogether, 126 slides were used for validation and annotated for further training after each validation round. Thus, 171 training slides included the validation slides. After eight

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

rounds of annotation and validation, the final version of the CNN was trained with 36 347 iterations, 400 µm field of view and ultra-complex neural networks. Aiforia training framework automatically stops the training if the training loss does not decrease over a period of iterations set by the user. To increase the generalizability of the CNN, the training data were digitally augmented for size (from -1% to 1%), aspect ratio (1%), shear distortion (1%), luminance (from -1% to 1%) and contrast (from -1% to 1%). Furthermore, all training data were rotated (360 degrees) and flipped (see Table S1). To test the performance of the CNN, verification was done by comparing the inference of the CNN against the annotated ground truth.

Testing the CNN

Finally, the CNN was tested against external data produced by independent human experts. For testing, 148 WSIs (different from those included in the training data set) were analyzed by CNN (link to the slides and the results of the analyses below) and by four independent researchers from different laboratories experienced with estrous cycle determination. One of the researchers had produced annotations for the CNN training. In addition, we tested 22 extra slides prepared and stained according to our protocol in another laboratory. All testing samples were collected from the female mice of C57BL/6N genetic background. Images out of focus were excluded from the validation. The researchers visually evaluated the samples under a microscope and determined the predominant stage of the cycle according to a previously published method (Byers et al., 2012). When the preparation did not represent any single stage, but the transition from one stage to another was obvious, as judged by the researcher, the stage was described as transitional (e.g. P/E). In contrast, the stage in the digital WSIs with features of several stages was classified to be the one with the highest value as quantified by CNN. However, if there were two predominant stages on WSI (representing >30% of the area of all stages), the stage was determined as transitional. The results from each researcher and the CNN were compared using confusion matrices. For clarity, WSIs quantified as

transitional stages by CNN were excluded from the validation data. However, the slides classified by a researcher to belong to a transitional stage were included in the data, and the more prominent stage (estimated by the researcher) was selected as the actual stage.

Statistical analyses

Agreement among several raters (Fleiss Kappa) was calculated with MAGREE macro, including all researchers and algorithms, and also for the researchers only. The methodology implemented by MAGREE macro is presented by Fleiss (Fleiss, 1981) and Kendall (Kendall, 1995). The data analysis was generated using SAS software, Version 9.4 of the SAS System for Windows (SAS Institute Inc., Cary, NC, USA). The principal component analysis was done using the GraphPad Prism 10 software (GraphPad Software, La Jolla, CA, USA) on all analyzed slides, with the percentages of areas assigned by the model as each stage as variables, principal components selected based on parallel analysis, and results presented as a score plot.

Results

Verification of the CNN

An example of an output of the AI model is presented in Figure 3. The verification was performed by comparing the inference of the CNN against the annotated ground truth and by measuring false positive and false negative rates. According to the verification results, CNN has learned the ground truth well (Table 1). The method showed high precision with a false positive rate of 0.19 and a false negative rate of 0.92. An example of the visual presentation of the verification by the AI model is presented in Figure 4.

Principal component analysis (Figure 5) clearly indicates the dynamic feature of the estrous cycle. As percentages of areas per stage determined by the CNN for each sample were plotted together, the PCA shows how estrous cycle is moving constantly from one stage to another.

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

Testing the CNN against independent researchers

The data were then collected from three independent experiments using the same mouse strain (C57BL/6N). Initially, 148 WSIs were selected and analyzed using the AI model and by four independent researchers from three laboratories. In 147/148 of the samples, at least one of the researchers agreed with the AI model, and in 103/148 of the samples, the results by all the researchers and the AI model were identical.

When transitional stages (= two predominant stages on the slide, representing >30% of the area of all stages) determined by CNN were excluded from the testing data, the final number of slides was 132. Statistical analyses (see later) were performed with these 132 slides. The number of the different stages determined by CNN and all the researchers is shown in Table 2.

The agreements and disagreements between the CNN model and the four researchers are presented in a confusion matrix (Table 3). The results of the CNN are plotted on the *x*-axis, and the researchers' results are on the *y*-axis. The numbers in the diagonal section of the table (bold) present the values in which both CNN and the four researchers determined the same estrous cycle stage. The confusion matrix indicates that the CNN training was completed successfully, as the total agreement between CNN and R1, who trained the model, was 93 %. The corresponding percentages between CNN and the other three researchers were also very high (80-89%) but also reflect the variation typical also between experienced researchers.

Fleiss kappa value representing agreement between CNN and researchers was 0.75 (Table 4), confirming that overall agreement between CNN and all researchers was excellent. When the different stages were analyzed separately, D, E and P showed very good or excellent agreement between researchers and CNN (Fleiss kappa values 0.89, 0.79 and 0.74, respectively). In contrast, the agreement was poor for the M stage, and the kappa value was only 0.27. However, the Fleiss kappa value for M was poor even between the researchers

(0.26), although the overall agreement (for all stages) between them was good (kappa 0.73). This indicates that the problem with analyzing M stage is in defining its ground truth.

To evaluate the robustness of our CNN model, we tested a set of samples that were prepared and stained according to our protocol, but in a different laboratory. Two independent researchers agreed with the AI model in 20 out of 22 samples analyzed, and only in one sample both of the researchers disagreed with the AI model. Therefore, we suggest that the CNN can be used confidently to analyze estrous cycle samples from different laboratories, provided the same staining protocol.

Discussion

The estrous cycle stage is an important indicator of the functional status of the hypothalamus-pituitary-ovarian axis, and consequently, the hormonal status of a vast number of different tissues in preclinical mouse models at the time of sampling. In the present study, we generated a CNN model to define the estrous stage of the female mice from hematoxylin-stained vaginal smears placed on microscopic slides. We utilized the Aiforia platform, which provided an AI architecture for developing an algorithm to analyze WSIs, and the method applies high-performance cloud computing and deep neural networks, specifically CNN. The aim was to facilitate analyses of large cohort studies and maximize the results' reliability and reproducibility.

Our model was trained to detect all the estrous cycle stages: proestrus (P), estrus (E), metestrus (M) and diestrus (D). The CNN was tested using a different set of WSIs, which were not used for the training. The results indicated that the total agreement between the researchers and the CNN was excellent. The internal verification tool further confirmed that the CNN had learned the annotated ground truth. Both CNN and researchers classified most cases as D stage, which aligns with the published data showing that D is the longest stage of the estrous cycle (Cora et al., 2015). In accordance with previous results (Ajayi and Akhigbe,

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

2020), our data showed that M is likely the shortest stage in the estrous cycle, as only 3/132 slides were classified by the CNN to belong predominantly to M. In addition, the fact that the vaginal smears were always taken at the same time of day probably contributed to the low chance of hitting the cycle at M stage.

Although the model recognizes the P, E and D stages well, the M stage could not be recognized with high confidence. Similar observations have been made by Wolcott and coworkers (Wolcott et al., 2022), who developed a DL-based method to analyze rodent estrous cycle with a large set of samples from different species and stains. However, in their model, the M stage caused confusion mainly with P, while in our model, there was confusion with all other stages. In the present study, the human validators classified a few more cases to M stage (5-23/132) than CNN (3/132), but the variation between the researchers was evident, and the discrepancies between researchers in defining the ground truth regarding M stage could not be solved by a computational method. Taken together, the above results indicate the importance of to better defining the cytological appearance of the vaginal smear at metestrus and, thus, the transition from estrus to diestrus.

The CNN developed has a comparable accuracy to the other two deep learning-based methods for analyzing the estrous cycle stage. In the method combining several species and stainings (Wolcott et al., 2022), the test accuracy varied from 69% (metestrus) to 96% (estrus), whereas in our study, the overall accuracy between the researchers and the CNN varied from 80% to 93%. In the other previously published work, the overall accuracy was reported to vary from 79% to 91% (Sano et al., 2020), while their model did not include the metestrus stage.

As far as we know, our model is the first WSI-based model to classify the mouse's estrous cycle by a deep learning-based algorithm. The other recently published algorithms to analyse estrous cycle are based on data obtained from micrographs (Sano et al., 2020, Wollcott et al., 2022). The benefits of Whole Slide Images (WSIs) include the availability of information at multiple scales and at multiple z-stack levels. The model, thus, analyzes and quantifies the

entire slide, eliminating human biases in area selection. Most importantly, our model has a crucial capability of classifying not only the primary stage of the sample but also determining the relative abundance of each stage in the sample, thereby demonstrating the transitional stage of the specimen. Thus, the model provides more precise results than manual 4-stage classification, including stage transitions, for better estimation of estrogen stimulus at the time of sample collection. We conclude that the developed model is a reliable and effective method to classify the estrous cycle stages in preclinical studies utilizing female mice.

Declaration of interest

There is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

Sigrid Juselius Foundation, Drug Research Doctoral Programme, University of Turku

Author contribution statement

Strauss L: Design of the study, creating the CNN, validating the results, writing the manuscript

Junnila A: Preparing estrous cycle samples for the study, validating the results

Wärri A: Validating the results, writing the manuscript

Manti M: Validating the results

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

Jiang Y: Preparing estrous samples for the study at the University of Gothenburg

Löyttyniemi E: Performing statistical analyses

Stener-Victorin E: Validating the results, reviewing manuscript

Lagerquist M: Organizing the samples for the study at the University of Gothenburg

Kukoricza K: Arranging/preparing samples for the study

Heinosalo T: Arranging/preparing samples for the study

Blom S: Consulting in CNN construction

Poutanen M: Study design, writing the manuscript

Data availability

Training slides

Test slides and analysis results

Acknowledgements

The authors thank Jenni Airaksinen, Katri Hovirinta, Leena Koskinen, Nina Messner, and Heli Niittymäki for their technical support and advice. All the animal work was performed at the Turku Center for Disease Modeling, part of Biocenter Finland network.

Figure legends

Figure 1. Vaginal smear stained with hematoxylin. Based on cytology, the estrous cycle can be divided into four stages: A) Proestrus with mainly immature vaginal squamous epithelial

cells with visible nuclei (arrows), B) Estrus with mainly mature vaginal squamous epithelial cells that are cornified, without having nuclei (arrow), C) Metestrus with mature epithelial cells (arrow) and few leukocytes (oval) and D) Diestrus with many leukocytes (oval) and few immature epithelial cells. Bar = 200 µm.

Figure 2. Defining the ground truth. Only areas clearly representing one estrous cycle stage were selected for training to define the ground truth. A) Proestrus, B) Estrus, C) Metestrus, D) Diestrus. Bar = 200 µm.

Figure 3. An example of an output of the algorithm. A) The representative areas are stained with pseudo colors. Blue indicates metestrus and red diestrus. B) The algorithm provides numerical values as the percentage area of each stage in the sample. Bar = 2.5 mm.

Figure 4. Internal verification of the AI model. The area of interest was determined and consequently annotated as estrus by an experienced researcher. The green color in the picture indicates the area, which the AI model also classified as estrus. The AI model classified the orange area (star) as proestrus, and thus, indicated as a" false positive". The colorless area (arrow) was not recognized to belong to any stage by the model and thus indicated as a" false negative." Bar = 500 µm.

Figure 5. Principal component analysis on all slides analysed by the CNN. The percentages of the areas indicating different estrous cycle stages were plotted together. The figure shows the dynamic feature of the estrous cycle.

References:

Ajayi, A.F., Akhigbe, R.E., 2020. Staging of the estrous cycle and induction of estrus in experimental rodents: an update. Fertil. Res. Pract. 6, 5. https://doi.org/10.1186/s40738-020-00074-3

Beery, A.K., Zucker, I., 2011. Sex bias in neuroscience and biomedical research. Neurosci. Biobehav. Rev. 35, 565–572.

Byers, S.L., Wiles, M.V., Dunn, S.L., Taft, R.A., 2012. Mouse Estrous Cycle Identification Tool and Images. PLoS ONE 7, e35538.

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

Calanni-Pileri, M., Weitzel, J.M., Langhammer, M., Wytrwat, E., Michaelis, M., 2022. Altered insulin, leptin, and ghrelin hormone levels and atypical estrous cycle lengths in two highly fertile mouse lines. Reprod. Domest. Anim. 57, 577–586.

Committee for the Update of the Guide for the Care and Use of Laboratory Animals; National Research Council: Guide for the Care and Use of Laboratory Animals: Eighth Edition. Washington, DC, National Academies Press, 2011

Cora, M.C., Kooistra, L., Travlos, G., 2015. Vaginal Cytology of the Laboratory Rat and Mouse: Review and Criteria for the Staging of the Estrous Cycle Using Stained Vaginal Smears. Toxicol. Pathol. 43, 776–793.

Elderman, M., de Vos, P., Faas, M., 2018. Role of Microbiota in Sexually Dimorphic Immunity. Front. Immunol. 9, 1018.

Fleiss, J.L., 1981. Statistical Methods for Rates and Proportions, second edition. John Wiley & Sons Inc., New York.

Fuentes, N., Roy, A., Mishra, V., Cabello, N., Silveyra, P., 2018. Sex-specific microRNA expression networks in an acute mouse model of ozone-induced lung inflammation. Biol. Sex Differ. 9, 18.

Hokenson, R.E., Short, A.K., Chen, Y., Pham, A.L., Adams, E.T., Bolton, J.L., Swarup, V., Gall, C.M., Baram, T.Z., 2021. Unexpected Role of Physiological Estrogen in Acute Stress-Induced Memory Deficits. J. Neurosci. 41, 648–662.

Hua, L., Zhuo, Y., Jiang, D., Li, Jing, Huang, X., Zhu, Y., Li, Z., Yan, L., Jin, C., Jiang, X., Che, L., Fang, Z., Lin, Y., Xu, S., Li, Jian, Feng, B., Wu, D., 2018. Identification of hepatic fibroblast growth factor 21 as a mediator in 17β-estradiol-induced white adipose tissue browning. FASEB J. 32, 5602–5611.

Jaric, I., Rocks, D., Greally, J.M., Suzuki, M., Kundakovic, M., 2019. Chromatin organization in the female mouse brain fluctuates across the oestrous cycle. Nat. Commun. 10, 2851.

Kemiläinen, H., Adam, M., Mäki-Jouppila, J., Damdimopoulou, P., Damdimopoulos, A.E., Kere, J., Hovatta, O., Laajala, T.D., Aittokallio, T., Adamski, J., Ryberg, H., Ohlsson, C., Strauss, L., Poutanen, M., 2016. The Hydroxysteroid (17β) Dehydrogenase Family Gene HSD17B12 Is Involved in the Prostaglandin Synthesis Pathway, the Ovarian Function, and Regulation of Fertility. Endocrinology 157, 3719–3730.

Kendall, M.G., 1995. Rank Correlation Methods, Second Edition. Charles Griffin & Co. Ltd., London.

Kim, I., Field, T.S., Wan, D., Humphries, K., Sedlak, T., 2022. Sex and Gender Bias as a Mechanistic Determinant of Cardiovascular Disease Outcomes. Can. J. Cardiol. 38, 1865–1880.

Konstandi, M., Andriopoulou, C.E., Cheng, J., Gonzalez, F.J., 2020. Sex steroid hormones differentially regulate CYP2D in female wild-type and CYP2D6-humanized mice. J. Endocrinol. 245, 301–314.

Mayneris-Perxachs, J., Arnoriaga-Rodríguez, M., Luque-Córdoba, D., Priego-Capote, F., Pérez-Brocal, V., Moya, A., Burokas, A., Maldonado, R., Fernández-Real, J.-M., 2020. Gut microbiota steroid sexual dimorphism and its impact on gonadal steroids: influences of obesity and menopausal status. Microbiome 8, 136.

Mohamed, M.S., Moulin, T.C., Schiöth, H.B., 2021. Sex differences in COVID-19: the role of androgens in disease severity and progression. Endocrine 71, 3–8.

Nilsson ME, Vandenput L, Tivesten Å, Norlén AK, Lagerquist MK, Windahl SH, Börjesson AE, Farman HH, Poutanen M, Benrick A, Maliqueo M, Stener-Victorin E, Ryberg H, Ohlsson C. Measurement of a Comprehensive Sex Steroid Profile in Rodent Serum by High-Sensitive Gas Chromatography-Tandem Mass Spectrometry. Endocrinology. 2015 Jul;156(7):2492-502.

Roa, J., Ruiz-Cruz, M., Ruiz-Pino, F., Onieva, R., Vazquez, M.J., Sanchez-Tapia, M.J., Ruiz-Rodriguez, J.M., Sobrino, V., Barroso, A., Heras, V., Velasco, I., Perdices-Lopez, C., Ohlsson, C., Avendaño, M.S., Prevot, V., Poutanen, M., Pinilla, L., Gaytan, F., Tena-Sempere, M., 2022. Dicer ablation in Kiss1 neurons impairs puberty and fertility preferentially in female mice. Nat. Commun. 13, 4663.

Sano, K., Matsuda, S., Tohyama, S., Komura, D., Shimizu, E., Sutoh, C., 2020. Deep learning-based classification of the mouse estrous cycle stages. Sci. Rep. 10, 11714.

Vanderschueren, D., Laurent, M.R., Claessens, F., Gielen, E., Lagerquist, M.K., Vandenput, L., Börjesson, A.E., Ohlsson, C., 2014. Sex Steroid Actions in Male Bone. Endocr. Rev. 35, 906–960.

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

Wood, P.A., Hrushesky, W.J.M., 2005. Sex cycle modulates cancer growth. Breast Cancer Res. Treat. 91, 95–102.

Xing, E., Billi, A.C., Gudjonsson, J.E., 2022. Sex Bias and Autoimmune Diseases. J. Invest. Dermatol. 142, 857–866.

Yang, M., Liu, Q., Huang, T., Tan, W., Qu, L., Chen, T., Pan, H., Chen, L., Liu, J., Wong, C.-W., Lu, W.W., Guan, M., 2020. Dysfunction of estrogen-related receptor alpha-dependent hepatic VLDL secretion contributes to sex disparity in NAFLD/NASH development. Theranostics 10, 10874–10891.

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

Table 1. Verification the inference of the CNN against the annotated ground truth

ESTROUS CYCLE STAGE	***TOTAL ERROR (%)**	**†FALSE POSITIVE (%)**	**‡FALSE NEGATIVE (%)**
P	1.13	0.35	0.78
E	1.50	0.15	1.35
M	0.45	0.07	0.39
D	1.37	0.20	1.17
TOTAL	**1.11 (±0.47)**	**0.19 (±0.12)**	**0.92 (±0.43)**

The *Total Error, †False Positive, and ‡ False Negative percentage values indicate how well the trained CNN performs in the training data set and low values together with low training loss indicate that the labelled features have been learned well by the CNN. False positive pixel is defined as a pixel where the label and CNN prediction disagree so that CNN defines a semantic class incorrectly. False negative pixel is defined as a pixel where CNN calls a background pixel incorrectly.

Equations:

***Total error = (False Positive pixels + False negative pixels) / All annotated pixels per class

†False Positive = False Positive pixels / All annotated pixels for the given semantic class

‡False Negative = False Negative pixels / All annotated pixels for the given semantic class

Table 2. Number of estrous cycle stages as predicted by CNN and researchers

GROUND TRUTH	CNN	R1 LS	R2 AW	R3 MM	R4 AJ
P	30	26	22	16	22
E	41	40	43	31	27
M	3	5	12	23	16
D	59	62	56	63	68

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

Table 3. Confusion matrix: CNN versus different researchers

R1 (LS)					**R2 (AW)**				**R3 (MM)**				**R4 (AJ)**			
Ground truth	P	E	M	D	P	E	M	D	P	E	M	D	P	E	M	D
P	**26**	1	0	3	**22**	5	3	0	**16**	2	9	3	**22**	0	4	4
E	0	**39**	1	1	0	**38**	2	1	0	**29**	11	**1**	0	**27**	9	5
M	0	0	**2**	1	0	0	**3**	0	0	0	**3**	0	0	0	**3**	0
D	0	0	2	**57**	0	0	4	**55**	**0**	**0**	0	**59**	0	0	0	**59**
Total Agreement	**93 %**				**89 %**				**80 %**				**83 %**			

Table 4. Fleiss Kappa values

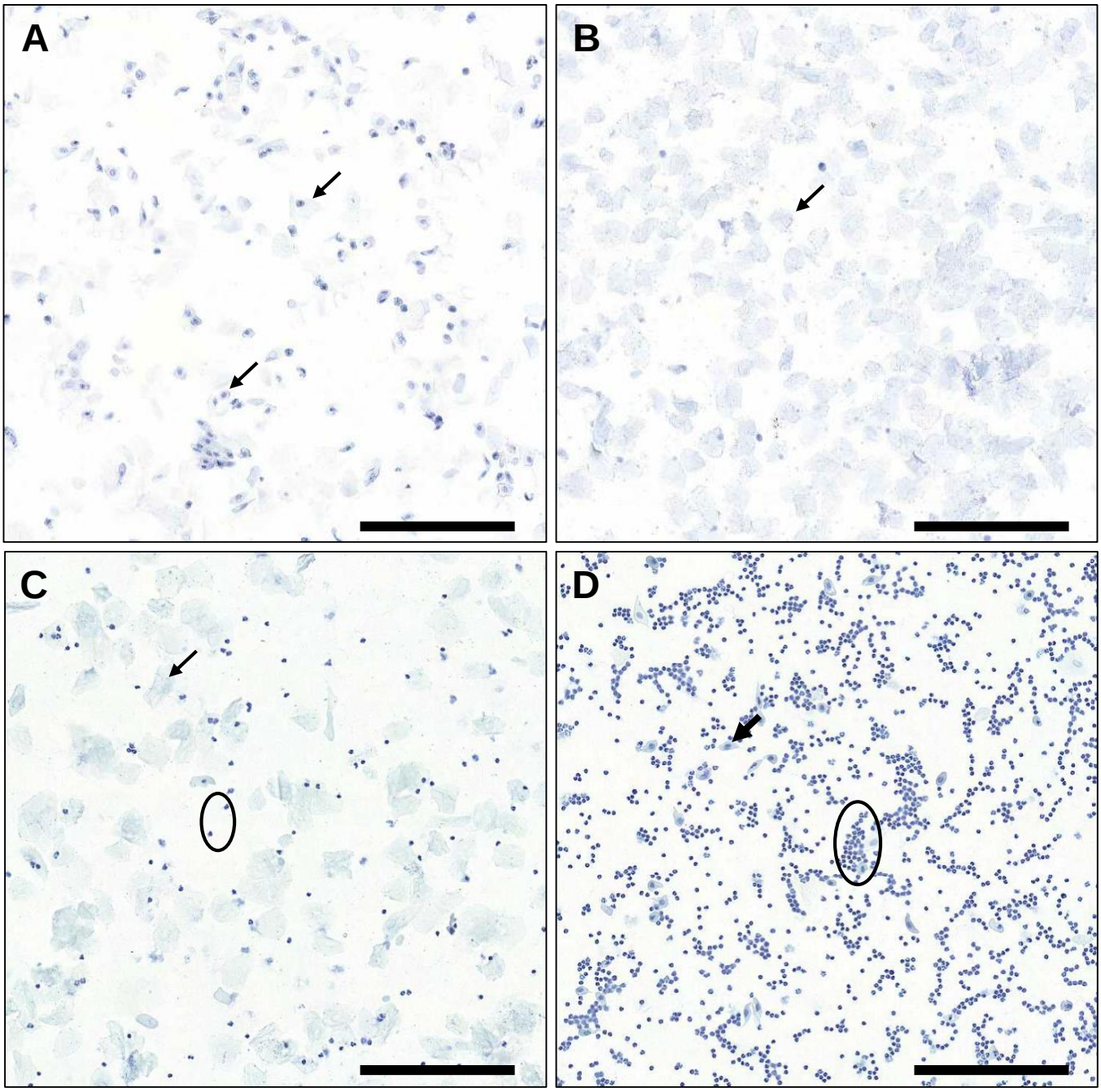
Ground truth	Kappa (CNN vs R1-R4)	Standard error (CNN vs. R1-R4)	Kappa (researchers only)	Standard error (researchers only)
P	0.74	0.0275	0.71	0.0355
E	0.79	0.0275	0.76	0.0355
M	0.27	0.0275	0.26	0.0355
D	0.89	0.0275	0.88	0.0355
Total	**0.75**	0.0176	**0.73**	0.0224

 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

Supplementary Table S1. CNN details (Aiforia model hyperparameters)

		CNN
Classes		Proestrus, estrus, metestrus, diestrus
Type (semantic segmentation)		Region
Complexity		Ultra Complex
Field of view		400 µm
Training parameters	Weight decay	0.0001
	Mini-batch size	5
	Mini-batches per iteration	20
	Iteration without progress	1500
	Initial learning rate	0.1
Image augmentation	Scale (min/max)	-1/1
	Aspect ratio	1
	Maximum shear	1
	Luminance (min/max)	-1/1
	Contrast (min/max)	-1/1
	Max with balance change	1
	Noise	0

Figure 1.



 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

Figure 2.

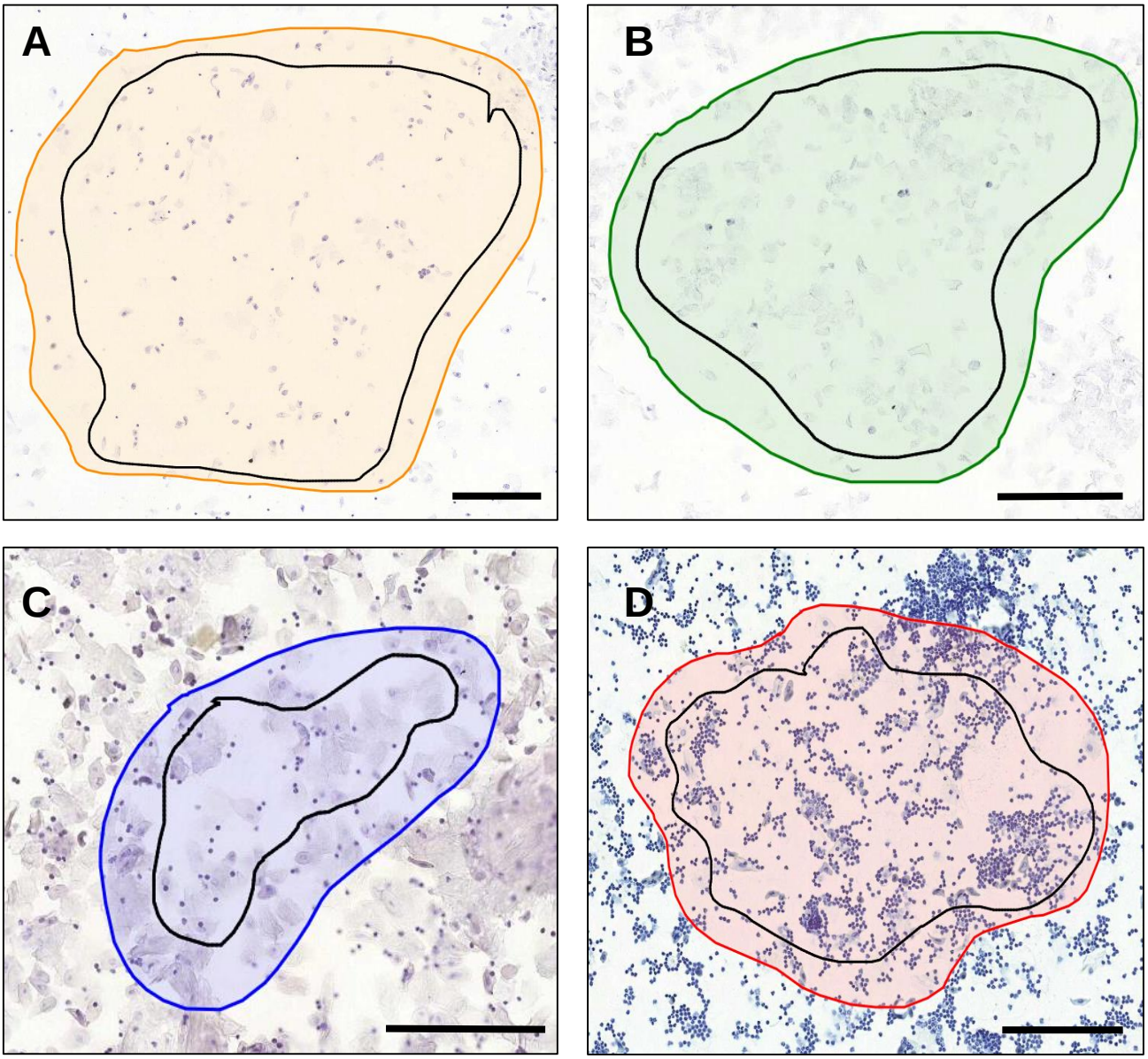
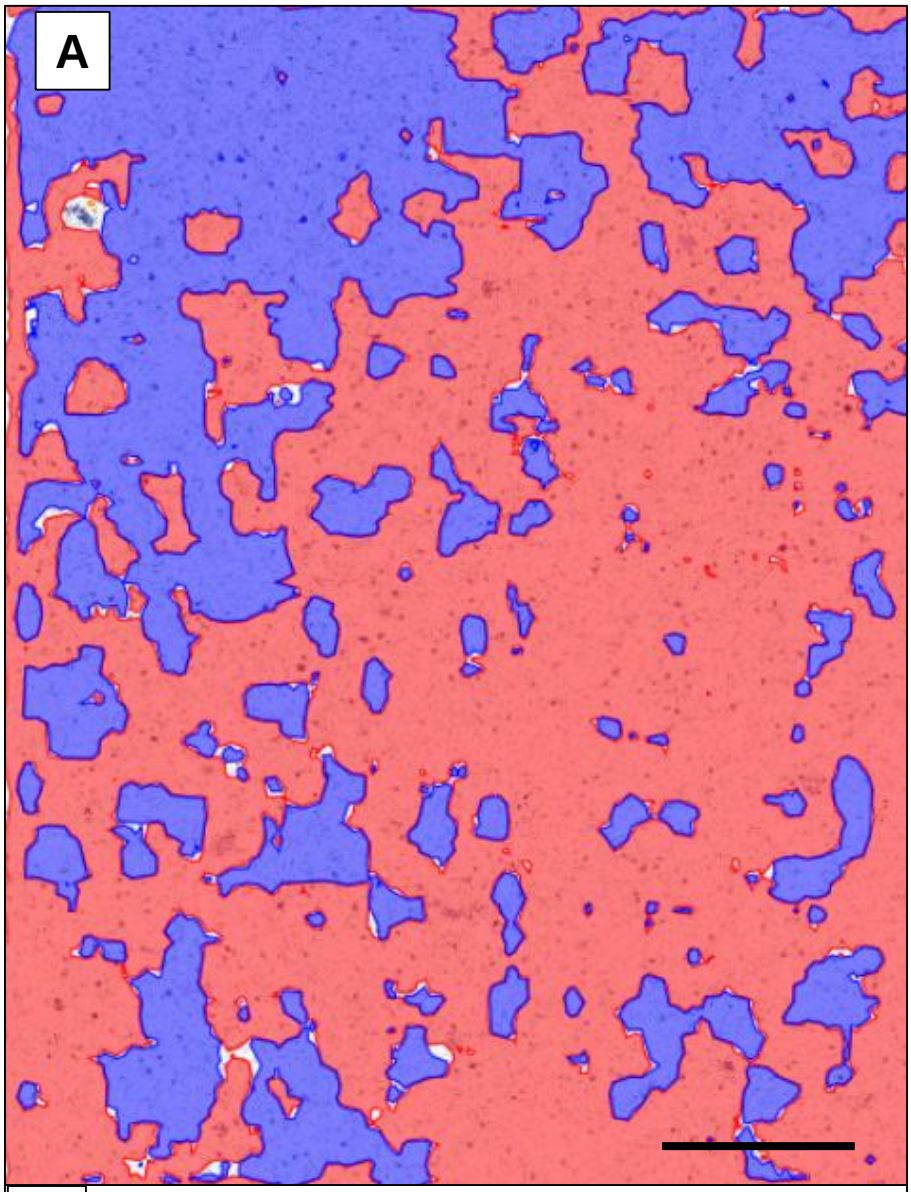


Figure 3.



 Accepted Manuscript published as JOE-23-0204.R2. Accepted for publication: 08-Apr-2024

Figure 4.

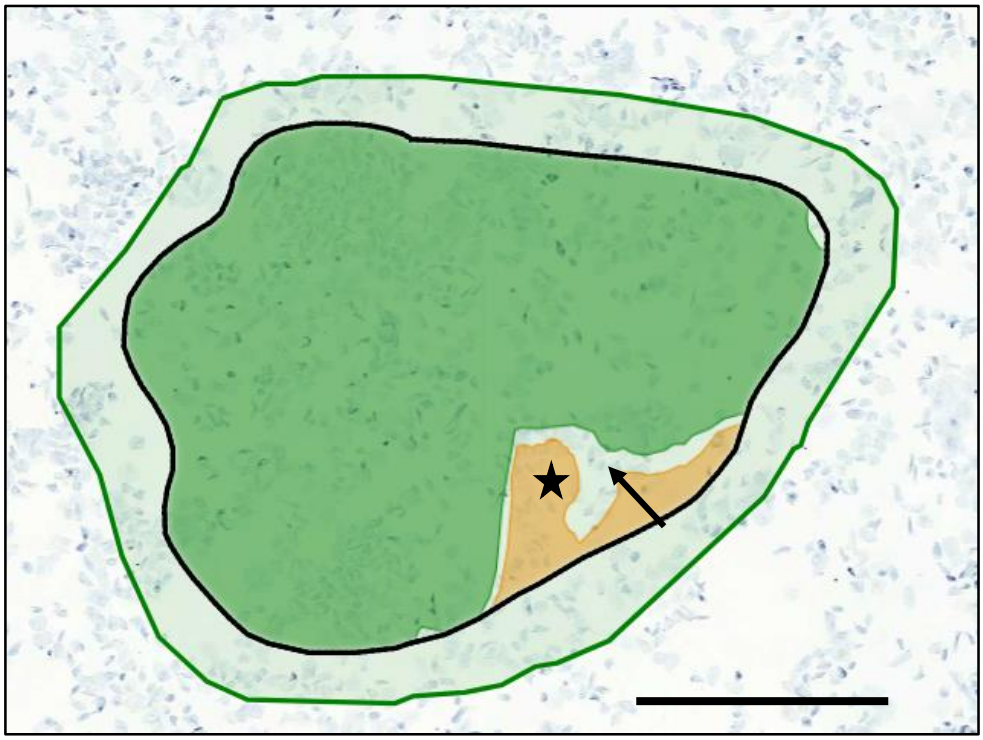


Figure 5