

Performance of convolutional neural networks in the histopathological grading of canine cutaneous mast cell tumors: a pilot study

Dyessica Ferreira Maia¹, Dyonnatán Ferreira Maia², Alice Assunção dos Santos¹, Ana Maria de Souza Almeida², Estela Vieira de Souza Silva¹, Veridiana Maria Brianezi Dignani de Moura², Eduardo de Paula Nascente¹ +

¹ Pontifícia Universidade Católica de Goiás

² Universidade Federal de Goiás

Abstract. Canine cutaneous mast cell tumor is one of the most prevalent skin neoplasms in dogs, and histological grading is essential for prognosis and therapeutic decision-making. However, histological classification according to the Patnaik and Kiupel systems presents limitations related to subjectivity and low reproducibility among pathologists. This study evaluated the performance of convolutional neural networks (CNNs) in the histopathological classification of canine cutaneous mast cell tumors. Eighteen histological slides labeled by specialists were used and processed into image patches. The AlexNet, InceptionV3, ResNet50, and YOLO11 architectures were trained using cross-validation and evaluated based on accuracy, sensitivity, precision, F1-score, and area under the curve (AUC). Grad-CAM techniques and feature maps were also applied for interpretative analysis of the models' decisions. The YOLO11 and ResNet50 models showed the best performance, achieving AUC values of up to 93.5%, but did not reach the arbitrarily defined minimum clinical threshold of 95%. Visual analysis demonstrated that the models recognized relevant morphological patterns, such as granules and pleomorphism, but failed to differentiate between grades II and III. It is concluded that, although promising as auxiliary tools, CNNs still require methodological improvements, such as refined curation, semantic segmentation, and cellular structure identification, for clinical application.

Keywords: Dermatopathology; Artificial intelligence; Veterinary pathology.

Introduction

Mast cell tumor is one of the most common malignant skin neoplasms in dogs, accounting for approximately 11% of cutaneous tumors diagnosed in this species (De Nardi et al., 2022). It belongs to the group of round cell neoplasms and derives from mast cells—immune system cells that participate in inflammatory processes and tissue repair (Furlani et al., 2008; Ríos, 2008). When neoplastic, mast cells can release mediators such as histamine and heparin, producing significant systemic effects including gastrointestinal ulcers, vomiting, diarrhea, pruritus, hemodynamic disorders, and even circulatory collapse (Prado et al., 2012).

To define the prognosis and select the optimal treatment, classifying the tumor based on histopathological evaluation is essential (Blackwood et al., 2020). The system proposed by Patnaik et al. (1984) categorizes mast cell tumors into three grades (I, II, and III), considering criteria such as cellular morphology and tissue organization. However, it is considered subjective and can lead to inter-observer variability among pathologists

(Simões et al., 1994). As an alternative, Kiupel et al. (2011) proposed a two-tier system, low-grade and high-grade tumors, with more objective criteria. Despite improving consistency, this system still faces limitations, particularly in cases where the neoplasm does not clearly fit into either category.

Given these challenges, artificial intelligence (AI) emerges as a promising alternative to assist and enhance diagnostic accuracy. However, its application in veterinary histopathology remains limited by a scarcity of validated studies (Yan & Hill, 2024). Therefore, this study aimed to investigate the performance, limitations, and potential of AI specifically convolutional neural networks (Norvig et al., 2022), for the histological grading of canine cutaneous mast cell tumors.

Previous research, such as that by Salvi et al. (2021), demonstrated that models like AlexNet, InceptionV3, and ResNet can achieve high accuracy in classifying these neoplasms. However, most of these studies rely on small datasets with low diversity and lack external validation, thereby limiting the applicability of their findings in real-world

veterinary clinical settings. Furthermore, some authors emphasize that the subjectivity of histopathological diagnosis and inter-pathologist discordance pose significant challenges to AI model training (Christof et al., 2020).

To ensure clinical acceptability in pathology, McGenity et al. (2024) established performance targets of sensitivity $\geq 95\%$ and AUC $\geq 98\%$. Algorithm validation studies in digital pathology demonstrated a mean sensitivity of 96.3% (95% CI: 94.1–97.7%) across various applications, reflecting the ability to identify true positives without excessively increasing false positives (McGenity et al., 2024). Finally, on the computational side, state-of-the-art models achieve, on average, over 91% and 99% accuracy on the ImageNet and CIFAR-10 benchmarks, respectively (Chen et al., 2021). Since these values are both feasible and desirable, and given the absence of specific metric reference standards for clinical use, we deemed it reasonable to establish a sensitivity target of 95% and, for simplicity, adopt this same threshold as a baseline for the other metrics (accuracy, precision, F1-score, and AUC). Consequently, the following hypothesis was formulated: Convolutional neural network models, even when trained on a limited dataset of histopathological images, will correctly classify Patnaik grades (I, II, and III) with high diagnostic performance (accuracy, sensitivity, and AUC exceeding 95%), demonstrating high reproducibility compared to diagnoses performed by pathologists.

In this context, this study aims to evaluate the performance of different Convolutional Neural Network (CNN) models in the histopathological classification of canine cutaneous mast cell tumors according to the grading system proposed by Patnaik (1984).

Material and Methods

The study was conducted in four stages. The first stage involved the collection, selection, and labeling of histological images according to histopathological examination results. In the second stage, image preprocessing was performed to prepare data for input into the computational models, including cropping, resizing, and partitioning the dataset into training, validation, and test sets. In the third stage, training and testing of the main Convolutional Neural Network (CNN) architectures were executed. Finally, the fourth stage comprised the analysis of performance metrics, complemented by visual evaluation and interpretation of the results. In the first stage, histopathological examination results from the Veterinary Pathology Department of the School of Veterinary Medicine and Animal Science at the Federal University of Goiás (SPA-EVZ-UFG), located in Goiânia, Goiás, were analyzed. Cases diagnosed with canine cutaneous mast cell tumor were selected according to the histopathological criteria defined by the World Health Organization (Misdorp et al., 1999) and Meuten (2017).

Image selection was based on representative morphological criteria for each mast cell tumor grade, derived from diagnoses established by two experienced veterinary pathologists using the criteria proposed by Patnaik et al. (1984). For Grade I, regions exhibiting well-differentiated mast cells in the superficial or deep dermis arranged in cords or small nests were selected; cells were round to oval or uniform with abundant, well-demarcated cytoplasm, round nuclei, and absence of mitotic figures. For Grade II, tumoral regions of moderate to marked cellularity invading the deep dermis and subcutis were captured; cells were round to oval and moderately pleomorphic, with occasional binucleated cells; distinct to indistinct cytoplasm; round nuclei with one or more visible nucleoli; rare mitotic figures (0 to 2 per high-power field – 400x); and, when present, areas of edema and necrosis. For Grade III, criteria included extensive extension into the dermis and subcutis; pleomorphic cells arranged in sheets or narrow cords; anisocytosis and anisokaryosis; indistinct cytoplasm; round nuclei with one or more prominent nucleoli; 3 to 6 mitotic figures per high-power field; multinucleated and giant cells; and frequent edema, hemorrhage, and necrosis.

Histological samples from six cases for each grading category (Figure 1) were selected, totaling 18 analyzed slides. Images were captured at the SPA-EVZ-UFG using light microscopy coupled to a digital image acquisition system. An average of ten distinct histological fields were recorded per slide, resulting in a total of 167 images (40 Grade I, 58 Grade II, and 69 Grade III). Captures were performed using a 40x objective lens, a magnification frequently recommended for offering an ideal balance between magnification and spatial resolution, thereby enabling the visualization of relevant morphological details (Montanari, 2016).

In the second stage, the image dataset was divided into 90 images for training (30 per grade) and 28 images for testing (ten Grade I, eight Grade II, and ten Grade III), ensuring that the test data were completely distinct from those used during training. Subsequently, the images were segmented into patches, generating square tiles with a resolution of 540×540 pixels (Figure 2). The use of this format promotes standardization and preserves critical morphological details, as widely adopted in computational histopathology studies (Petříková et al., 2023). Patch generation also increased the overall dataset size, yielding a total of 1,989 images. This updated dataset was partitioned into five folds for cross-validation (Mahlich et al., 2024), with each iteration comprising approximately 80% for training/validation and 20% for testing.

In the third stage, model training was conducted using pre-trained weights from the ImageNet database (Ridnik et al., 2021). Various CNN architectures were evaluated, including AlexNet, Inception-v3, and ResNet as tested in the pilot study by Salvi et al. (2021) along with a modern CNN architecture, YOLOv11 (Khanam & Hussain,

2024). Individual training procedures were performed for each classifier using 5-fold cross-validation ($k=5$) to classify images into one of the three categories (Grades I, II, or III). Computational implementation was carried out using the

Transformers (Wolf et al., 2020), scikit-learn (Pedregosa et al., 2011), and PyTorch (Paszke et al., 2019) libraries in the Python programming language.

Figure 1. Photomicrographs of canine cutaneous mast cell tumors. A.1 and A.2: Grade I; B.1 and B.2: Grade II; C.1 and C.2: Grade III. HE, 40x.

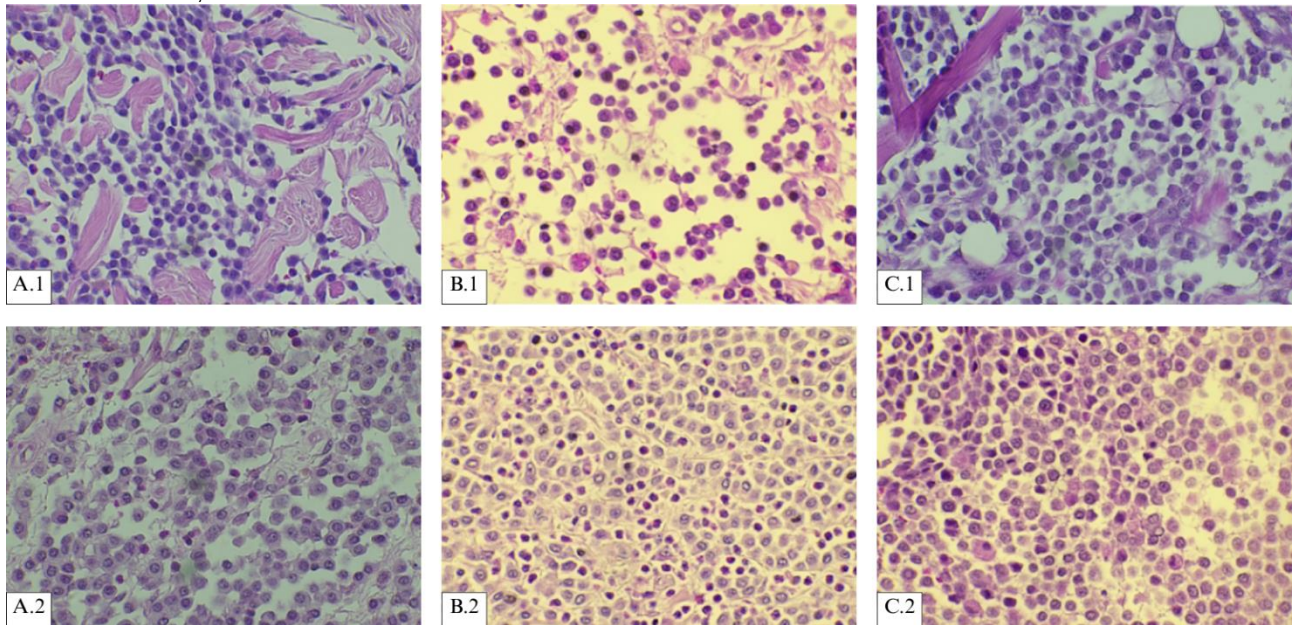
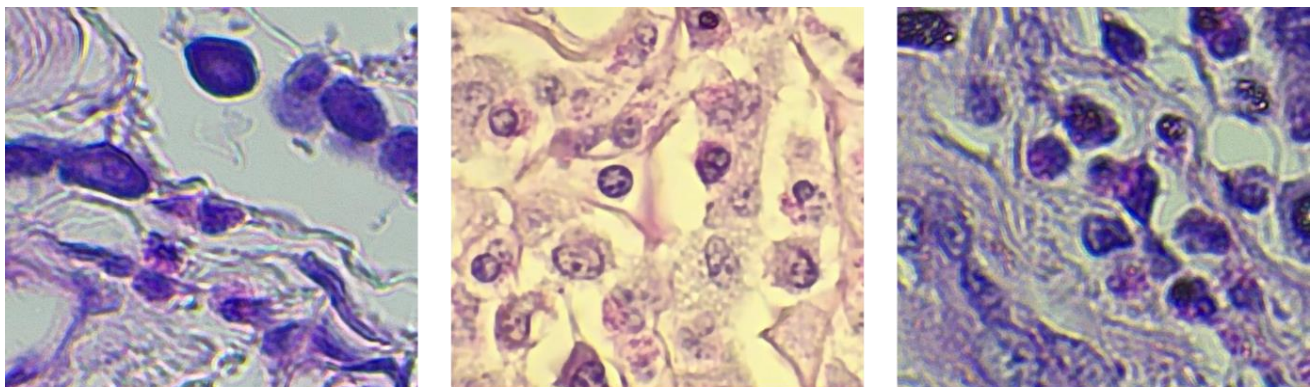


Figure 2. Photomicrographs of canine cutaneous mast cell tumors. Example of image patches cropped from the original images of Grade I, Grade II, and Grade III mast cell tumors.



In the final stage, the models were evaluated using accuracy, precision, sensitivity (recall), and F1-score metrics, considering the grading system proposed by Patnaik et al. (1984). The top-performing models in this classification were also evaluated on the grading system proposed by Kiupel et al. (2011) to verify their applicability in systems with more objective criteria. Additionally, a qualitative analysis was conducted focusing on the visual interpretation of model decisions using two approaches: feature map visualization, as described by Zeiler and Fergus (2014), and the Grad-CAM technique (Selvaraju et al., 2017), which highlights

the image regions that most influenced the model's final prediction.

Results and Discussion

For the dataset used in the experiments, Table 1 outlines the final number of images for each grade across the training, validation, and test sets. Table 2 presents the test set results for the AlexNet, InceptionV3, ResNet50, and YOLOv11 models following the fine-tuning process of the architectures. Evaluated metrics included accuracy, precision, sensitivity, and F1-score for the histopathological classification of canine cutaneous mast cell tumors according to the Patnaik et al. (1984) grading system.

Table 1. Number of examples per grade across the training, validation, and test dataset splits, with absolute values and their respective percentages for each set.

Grades	Training	Validation	Test	Total
Grade 1	408	102	170	680
Grade 2	408	102	136	646
Grade 3	421	106	136	663
Total	1237 (62%)	310 (15%)	442 (22%)	1989

Table 2. Performance of CNN models applied to the test set - Mean and standard deviation of metrics for accuracy, sensitivity, specificity, and F1-score.

Models	Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)	AUC (%)
Alexnet	53,27 ± 17,9	47,45 ± 30,6	54,41 ± 17,39	44,75 ± 23,6	69,24 ± 15,94
InceptionV3	75,4 ± 25,0	78,6 ± 19,5	75,2 ± 23,1	72,2 ± 30,9	81,03 ± 9,1
ResNet50	83,36 ± 9,77	86,11 ± 9,36	81,35 ± 11,19	79,13 ± 14,72	93,53 ± 6,42
Yolo11	87,77 ± 13,1	84,39 ± 20,9	87,55 ± 14,07	85,33 ± 18,72	91,96 ± 15,03

The ResNet50 and YOLOv11 architectures stood out among the evaluated models, each exhibiting distinct strengths. YOLOv11 demonstrated greater overall robustness, maintaining high and balanced performance across accuracy, sensitivity, and F1-score, which suggests strong generalization capability across different tumor grades. Its confusion matrix reveals consistent accuracy across all grades, particularly for Grade II (72.4%) and Grade III (75.8%), demonstrating that the model maintains effective class separation even in the most challenging cases.

Conversely, although ResNet50 presented lower overall performance, it achieved the highest precision and AUC among the models, indicating that when it assigns a prediction, it rarely misclassifies, and possesses excellent ranking capability between positive and negative classes. That is, it is particularly effective when well-defined morphological contrast is present. This interpretation is corroborated by its confusion matrix, which shows high reliability for Grade I (95.4%) and maximum performance for Grade II (100%), but a marked decline for Grade III, where correct predictions were diluted across classes (48.6% correct and up to 32% misclassified as Grade I) (Figure 3). This highlights a limitation in forming a clear separating hyperplane for cases with greater morphological complexity. On the other hand, the AlexNet and InceptionV3 models displayed more unstable performances and insufficient metrics for clinical application, reinforcing the necessity of more modern architectures for complex diagnostic tasks such as mast cell tumor classification.

Visual inspection of misclassified cases (Figure 4) reveals intermediate morphological patterns that may justify model predictions. In certain images, such as the first and third in the top row, cells exhibited relatively uniform nuclei without evidence of severe nuclear disorganization or disintegrating nuclei, which may have rendered them visually similar to the Grade II pattern. Thus, although labeled as Grade III, these samples may not fully express the most aggressive histological features, suggesting that the network may have reacted correctly to the visual field provided.

Conversely, in the last image of the bottom row, high cellularity, mitotic figures, and at least two binucleated cells are observable alterations consistent with greater biological aggressiveness. In this case, the misclassification and high standard deviations in the results suggest significant interference from the selection of the test dataset subset, indicating low representativeness of certain data that were not robustly learned, such as subtle morphological patterns including binucleation and frequent mitotic figures, criteria considered critical for differentiating Grade II from Grade III mast cell tumors.

Since the ResNet50 and YOLOv11 models achieved the best metrics, both architectures proved to be strong candidates for tumor grading according to the Kiupel et al. (2011) system. Thus, using the same training scheme with cross-validation, Table 3 presents the performance metrics obtained by the model, highlighting minimal differences compared to the Patnaik et al. (1984) system and remaining within the margin of variation. This was expected, as the thresholds between 'low grade' and 'high grade' coincide with the characteristics that define Grades II and III, which represent the primary challenges for the models. Because the ResNet50 model was the most precise regarding the AUC metric which indicates the confidence level of the model's classification and possesses lower architectural complexity, thereby facilitating human interpretability, it proved to be a strong candidate for generating visual analyses to interpret the regions and morphological patterns that most influenced the neural network's decision-making process. Furthermore, ResNet50 requires no structural architectural modifications for Grad-CAM implementation, whereas YOLOv11 lacks average pooling and fully connected layers, which are essential for heatmap generation (Selvaraju et al., 2017; Band et al., 2023). The first applied approach involved extracting feature maps from intermediate convolutional layers, allowing us to observe how ResNet50 processed and transformed the original image information across its layers.

Figure 3. Confusion matrices displaying mean $\pm$ standard deviation: ResNet50 (left) and YOLOv11 (right).

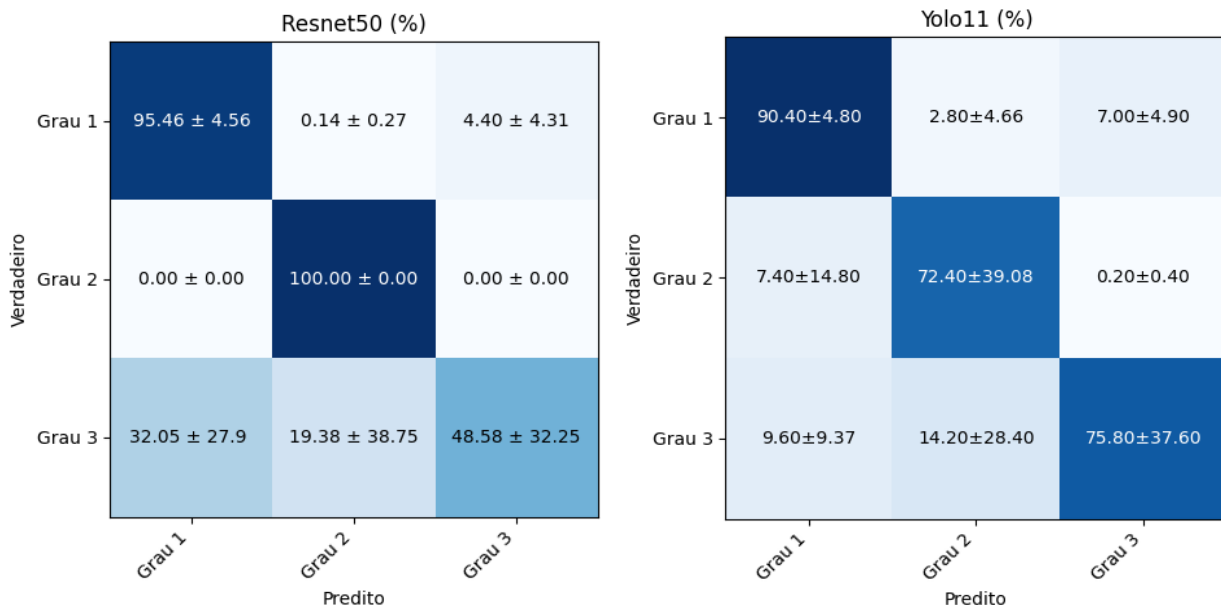


Figura 4. Examples of images misclassified by ResNet50.

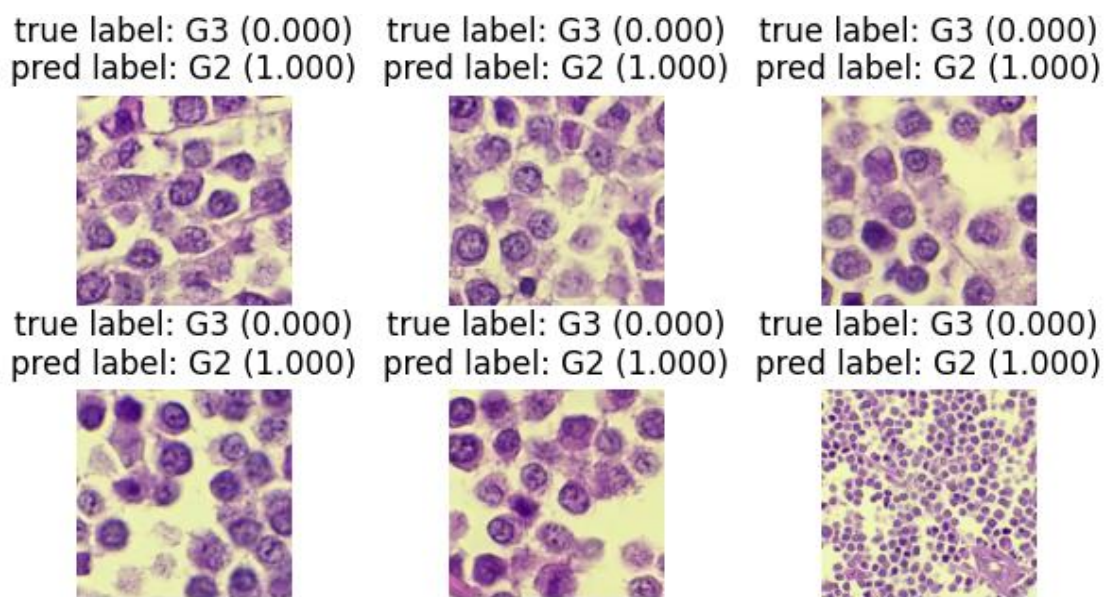


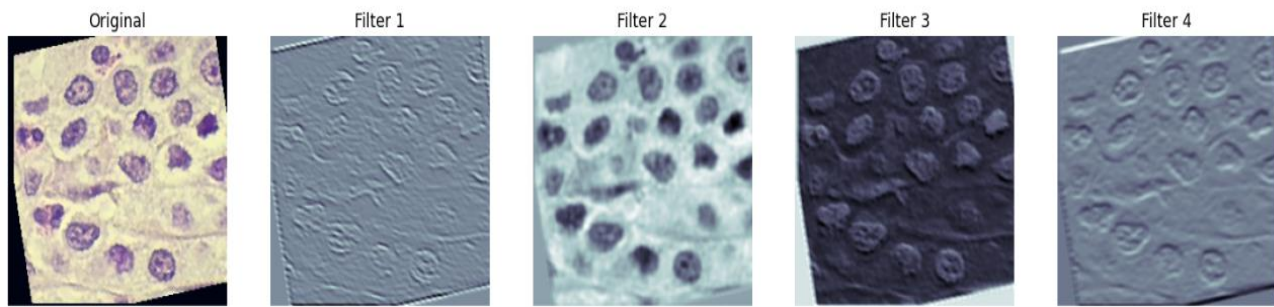
Table 3. Performance of the ResNet50 and YOLOv11 models, expressed as percentages, applied to test sets labeled according to the Kiupel grading system.

Models	Accuracy	Precision	Sensitivity	F1	AUC
ResNet50	81,05 $\pm$ 8,30	76,49 $\pm$ 21,06	76,52 $\pm$ 14,67	74,57 $\pm$ 18,24	86,2 $\pm$ 16,07
Yolo11	86,8 $\pm$ 16	82,5 $\pm$ 22	86,6 $\pm$ 16	84,1 $\pm$ 19	83,6 $\pm$ 22

In Figure 5, different filters responded to specific morphological patterns. Filter 1 and Filter 4 highlighted areas with high granular density, suggesting that the model recognized cytoplasmic granules a characteristic feature of mast cells. Filter 2 emphasized rounded structures of varying sizes, likely associated with neoplastic cells and their nuclei. Furthermore, Filters 2 and 3 demonstrated sensitivity to histological staining, highlighting purple

and bluish tones corresponding to the affinity of hematoxylin for nuclear components, as well as occasional yellowish hues that may be related to digital artifact variations or specific cytoplasmic characteristics. These findings indicate that the network incorporated relevant morphological patterns and chromatic features consistent with the histological criteria used in the classification of canine cutaneous mast cell tumors.

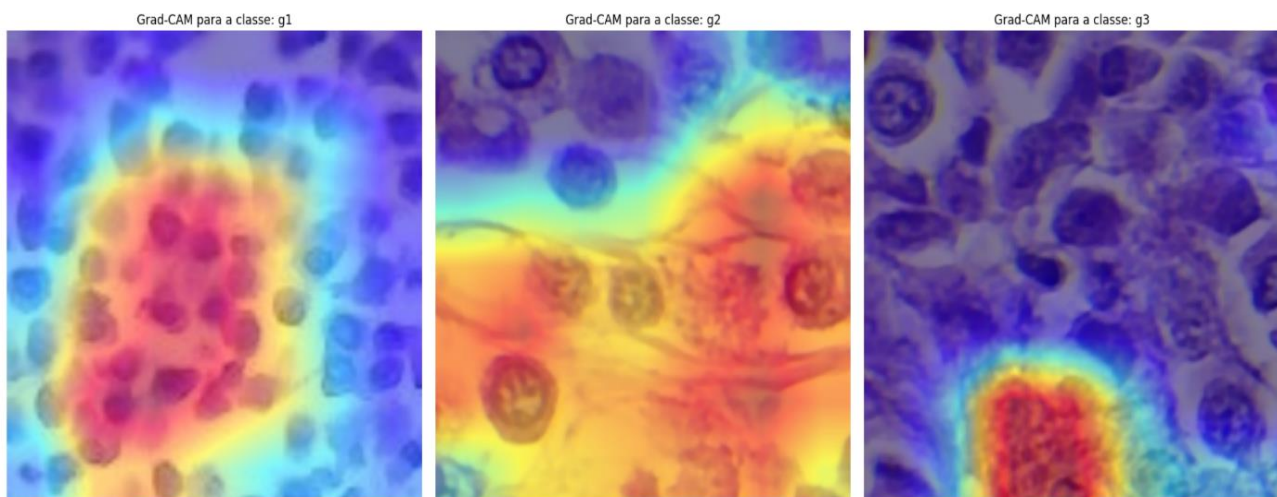
Figure 5. Visual representation of the output from ResNet50 convolutional layers.



The applied Grad-CAM technique (Figure 6) revealed that the ResNet50 architecture focused its attention predominantly on regions containing mast cell aggregates, areas with pronounced nuclear variations (pleomorphism), mitotic figures, and alterations in tissue architecture. This activation hotspots aligned with the primary histopathological criteria used in the classification of canine cutaneous

mast cell tumors, as described by Patnaik et al. (1984) and Kiupel et al. (2011). However, the activation generated by Grad-CAM was diffusely distributed, encompassing broad regions of the image rather than concentrating on specific cellular structures such as nuclei, nucleoli, or cytoplasmic granules.

Figure 6. Visual representation of ResNet50 attention generated by Grad-CAM for Grades I, II, and III, as proposed by Patnaik et al. (1984).



Visualization using guided backpropagation (Springenberg et al., 2014) (Figure 8, center) highlighted fine morphological details with greater clarity, such as cellular contours and nuclear outlines, emphasizing the importance of these structures in the model's decision-making process. The combined application with the Grad-CAM technique (Figure 7, right) enabled a more detailed visualization of regions deemed relevant by ResNet50, uniting the broad spatial localization of Grad-CAM with the structural refinement provided by guided backpropagation.

However, this approach also revealed a recurring challenge: the model's difficulty in precisely distinguishing between Grade II and, particularly, Grade III, especially in regions with significant morphological overlap. This behavior may be attributed to the histological similarity between the two grades, often characterized by moderate pleomorphism and the presence of mitotic figures in both, which compromises classification accuracy

even among expert pathologists. Such a limitation underscores the inherent subjectivity of histopathological diagnosis, which can influence the labeling of training data and the model's capacity to accurately replicate classification criteria.

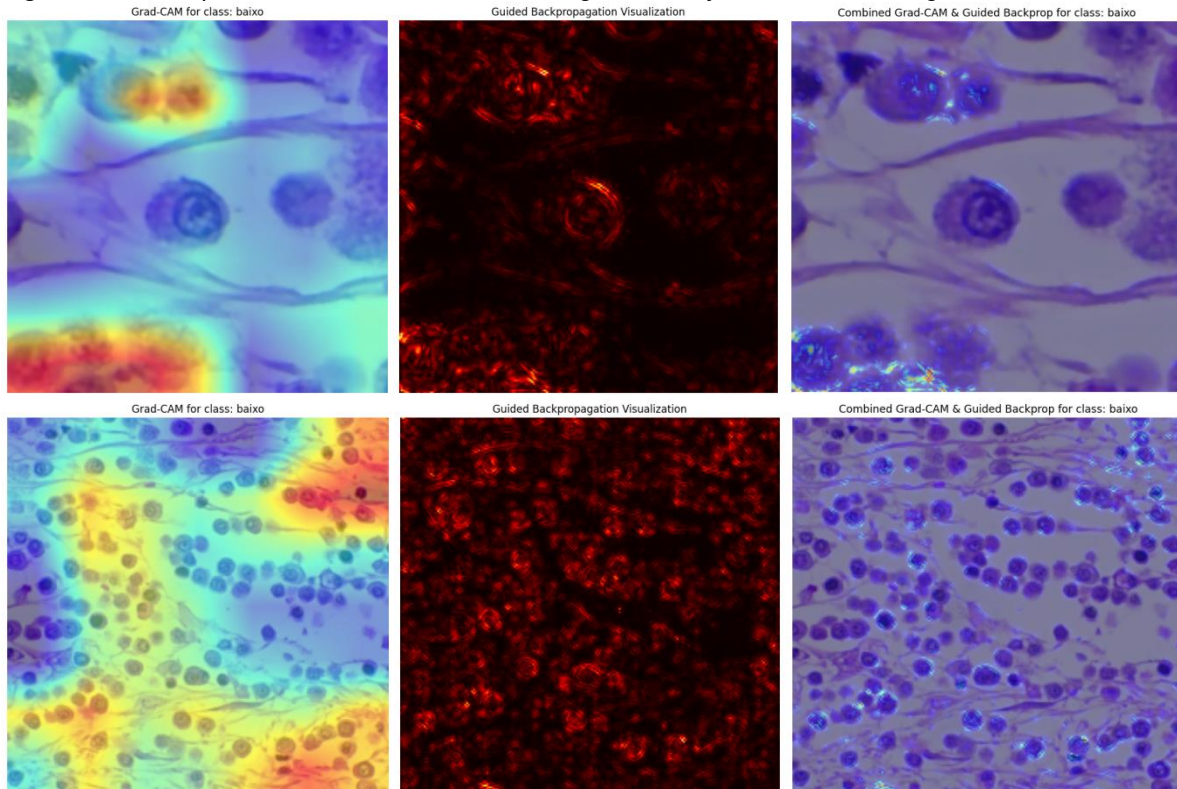
The results which fell below the pre-established threshold and were accompanied by high standard deviation variance indicate that the complexity involved in the histopathological classification of canine cutaneous mast cell tumors was not fully overcome by the convolutional neural network architectures employed, thereby compromising the degree of reliability required for practical applications. Nevertheless, the ResNet50 and particularly the YOLOv11 models, owing to its consistency, stood out with noteworthy results under the evaluated conditions, indicating that the field continues to evolve steadily.

Due to the scarcity of scanned and publicly available slides, relying on a small sample size of cases remains a common practice across studies.

For instance, the ARCTA dataset by Salvi et al. (2021) utilized 20 cases, resulting in 216 images for training and 15 for testing, though incorporating other round cell tumors. Other related computational histopathology studies did not focus on tumor grading but followed a similar trend of limited sample sizes, such as the 'HTT dataset' by Elfer et al. (2022) with 64 cases, and the 'MITOS_WSI_CCMCT' dataset by Bertram et al. (2019) with 32 cases.

To mitigate the effects of data scarcity in CNN training, several techniques have been developed, such as semi-supervised learning (Jiao et al., 2024) and few-shot learning (Wang et al., 2020), with Transfer Learning via fine-tuning standing out as the primary approach adopted in this domain. However, while CNNs inherently excel at extracting visual features from images, they lack more advanced logical reasoning capabilities, such as those found in Large Language Models (LLMs) (Bommasani et al., 2021).

Figure 7. Visual representation of ResNet50 attention generated by Grad-CAM for low grade.



Without detailed annotations, dataset training becomes even more challenging for models, which may end up learning only a fraction of the tissue's morphological diversity rendering them more susceptible to overfitting and less capable of coping with natural sample variations. However, Salvi et al. (2021) reported 96% training accuracy and 100% test accuracy using five images per grade, suggesting either a potential underrepresentation of the test set or indicating a problem that is easily solvable by CNNs.

Thus, in accordance with our initial hypothesis, we assumed that a small dataset would be sufficient for the model to learn and correctly classify Patnaik grades (Patnaik et al., 1984), thereby implying that the model could replicate the grading established by pathologists and demonstrate high diagnostic reproducibility for grading canine cutaneous mast cell tumors. However, as demonstrated in these experiments, standard CNN models did not approach 100%

accuracy in the most challenging scenarios. This corroborates other studies that highlight low inter-pathologist agreement in grading, reflecting the inherent difficulty of computationally replicating subjective human annotations (Bertram et al., 2020).

YOLOv11 and ResNet50 stood out with promising performance yet fell short of the established 95% threshold across all selected metrics. Consequently, the results of this study rejected the working hypothesis. Thus, the null hypothesis applies to the scenario evaluated here, indicating that the developed classifiers do not yet achieve sufficient and reliable diagnostic reproducibility for routine clinical use in veterinary pathology.

This limitation is primarily attributed to the small dataset size and the multifactorial nature of tumor classification under the Patnaik et al. (1984) and Kiupel et al. (2011) grading systems. The integrated analysis of quantitative metrics and visual representations suggests that the models

successfully identified relevant histological patterns, such as cellular pleomorphism and tissue architectural disorganization. However, persistent difficulties remained in differentiating Grade II from Grade III, underscoring the morphological overlap between these categories and reaffirming the diagnostic challenges inherent to this neoplasm. Furthermore, the high standard deviation highlights the difficulty of consistently establishing a training set that is sufficiently representative for model generalization alongside equally representative test sets.

The experimental results, coupled with visual analysis, emphasize the importance of strategies that reduce ambiguity in input data. Direct classification of complex images governed by multiple diagnostic factors revealed significant limitations of this approach for practical clinical application in pathology. This finding underscores the complexity of the task and highlights the need for more advanced methodological approaches and robust datasets to advance AI models in veterinary pathology.

Accordingly, the evolution of the dataset used in this study could be achieved through the incorporation of pixel-level localized morphological annotations, combined with instance detection and semantic segmentation techniques (Lu et al., 2021; Jiao et al., 2024). This would enable the identification of key structures, such as mitotic figures, preventing models from learning spurious correlations such as associating necrosis, which is not a formal criterion, with high tumor grade (Gross et al., 2005). Furthermore, expanding the dataset with whole-slide images (WSIs), including all cropped regions, special stains such as toluidine blue, and immunohistochemical techniques (Ramos-Vara & Miller, 2014), could refine ground-truth labels and validate histological patterns through annotations tailored for detecting nuclei, cytoplasmic granularity, and mitotic activity.

Consequently, the evaluation conducted in this study addressed the robustness of CNN classifiers, assessing whether they could accomplish the task with the necessary level of abstraction for the problem, even without fully detailed fine-grained annotations. Furthermore, it evaluated their capacity to extract such information in the presence of noisy data and datasets that reflect regional realities in histopathological diagnosis. The non-resolution of the problem indicates that it is not, in fact, trivially solvable and requires greater data preprocessing as well as refined criteria during case grading before revisiting image classification techniques.

The results obtained allow for an evaluation of this initial approach's performance, aiming for the gradual refinement of analytical techniques to achieve the requisite diagnostic accuracy and clinical utility. Beyond the quantitative validation of algorithms, identifying the limitations of applying AI to the histological grading of mast cell tumors is essential to proposing feasible pathways for technical and technological advancements in the

field. Thus, this research contributes not only experimental results but also provides insight into model inference through visual analysis techniques.

Conclusion

Despite the promising performance of the YOLOv11 and ResNet50 architectures, challenges associated with subjective labeling, small dataset size, and the morphological complexity of neoplasms prevented robust clinical accuracy, thereby rejecting the hypothesis that the model would achieve values equal to or exceeding 95% across quantitative metrics. Nevertheless, the findings highlight viable avenues for future work, such as employing semantic segmentation and object detection for relevant structures, including nuclei and mitotic figures. Additionally, the judicious integration of artificial intelligence into routine diagnostics may, in the near future, serve as a valuable decision-support tool for veterinary pathologists.

References

- BAND, Shahab S. et al. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, v.40, p.101286, 2023.
- BERTRAM, C. A. et al. MITOS_WSI_CCMCT: a dataset for mitotic figure assessment in canine cutaneous mast cell tumors. *Veterinary Pathology*, v.56, n.3, p.393–402, 2019. DOI: <https://doi.org/10.1177/0300985818815317>.
- BERTRAM, Christof A. et al. Are pathologist-defined labels reproducible? Comparison of the TUPAC16 mitotic figure dataset with an alternative set of labels. In: *Interpretable and Annotation-Efficient Learning for Medical Image Computing: MICCAI 2020*. Cham: Springer, 2020. p.204–213.
- BLACKWOOD, L. et al. European consensus document for mast cell tumours in dogs and cats. *Veterinary and Comparative Oncology*, v.18, n.3, p.439–452, 2020. DOI: <https://doi.org/10.1111/vco.12665>.
- BOMMASANI, R. et al. On the opportunities and risks of foundation models. *arXiv preprint*, arXiv:2108.07258, 2021.
- CHEN, Wuyang; GONG, Xinyu; WANG, Zhangyang. Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective. *arXiv preprint arXiv:2102.11535*, 2021.
- CHRISTOF, A. et al. Challenges in applying deep learning to histopathological images in veterinary medicine. *Veterinary Pathology Insights*, v. 13, p. 1–10, 2020.
- DE NARDI, A. B. et al. Diagnosis, Prognosis and Treatment of Canine Cutaneous and Subcutaneous

- Mast Cell Tumors. *Cells*, v. 11, n. 4, p. 618, jan. 2022.
- DE NARDI, A. B. et al. Tumores cutâneos em cães: uma revisão de 10 anos. *Veterinária e Zootecnia*, v. 29, n. 1, p. 10–20, 2022.
- ELFER, K. et al. Pilot study to evaluate tools to collect pathologist annotations for validating machine learning algorithms. *Journal of Medical Imaging (Bellingham)*, v.9, n.4, p.047501, jul. 2022. DOI: [10.1117/1.JMI.9.4.047501](https://doi.org/10.1117/1.JMI.9.4.047501).
- FURLANI, J. M. et al. Mastocitoma canino: estudo retrospectivo. *Ciência Animal Brasileira*, v. 9, n. 1, p. 242–250, 2008.
- FURLANI, R. P. Z. et al. Mastocitoma cutâneo canino: aspectos clínicos, diagnósticos e terapêuticos. *Ciência Rural*, v.38, n.4, p.1132–1138, 2008. DOI: <https://doi.org/10.1590/S0103-84782008000400038>.
- GROSS, T. L. et al. *Skin diseases of the dog and cat: clinical and histopathologic diagnosis*. 2. ed. Oxford: Blackwell Publishing, 2005.
- JIAO, Y. et al. A semi-supervised deep learning framework for histopathology image analysis in low-data regimes. *Medical Image Analysis*, v. 85, 102747, 2024. DOI: <https://doi.org/10.1016/j.media.2024.102747>.
- KHANAM, R.; HUSSAIN, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv preprint*, arXiv:2410.17725, 2024.
- KIUEP, M. et al. Proposal of a 2-tier histologic grading system for canine cutaneous mast cell tumors to replace the Patnaik system. *Veterinary Pathology*, v. 48, n. 1, p. 147–155, 2011. DOI: <https://doi.org/10.1177/0300985810386469>.
- LU, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, v. 5, p. 555–570, 2021. DOI: <https://doi.org/10.1038/s41551-020-00682-w>.
- MAHLICH, L.; WILHELM, F.; HOPPE, T. Efficient e-Fold Cross-Validation for Deep Neural Networks. *arXiv preprint*, arXiv:2401.13416, 2024.
- MCGENITY, Clare et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. *npj Digital Medicine*, v. 7, n. 1, p. 114, 2024.
- MEUTEN, D. J. *Tumors in Domestic Animals*. 4. ed. Iowa: Iowa State Press, 2017.
- MISDORP, W. et al. *Histological classification of mammary tumors of the dog and cat*. 2. ed. Washington, DC: Armed Forces Institute of Pathology, 1999.
- MONTANARI, M. F. Técnicas de microscopia óptica aplicadas à histopatologia veterinária. In: MEUTEN, D. J. (Org.). *Patologia veterinária especial*. 5. ed. São Paulo: Roca, 2016. p. 31–45.
- NORVIG, P. *Inteligência Artificial*. 4. ed. Rio de Janeiro: Grupo Gen, 2022.
- PASZKE, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv preprint*, arXiv:1912.01703, 2019.
- PATNAIK, A. K. et al. Canine cutaneous mast cell tumor: morphologic grading and survival time in 83 dogs. *Veterinary Pathology*, v. 21, n. 5, p. 469–474, 1984.
- PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.
- PETRÍKOVÁ, I. et al. A deep learning approach for histopathological image classification in veterinary oncology. *Computers in Biology and Medicine*, v. 158, 106957, 2023. DOI: <https://doi.org/10.1016/j.compbiomed.2023.106957>.
- PRADO, Aline Ambrogi Franco et al. Mastocitoma em cães: Aspectos clínicos, histopatológicos e tratamento. *Enciclopédia Biosfera, Centro Científico Conhecer*, v. 8, n. 14, p. 2151, 2012.
- RAMOS-VARA, J. A.; MILLER, M. A. When tissue antigens and antibodies get along: revisiting the technical aspects of immunohistochemistry—the red, brown, and blue technique. *Veterinary pathology*, v. 51, n. 1, p. 42–87, 2014.
- RIDNIK, T. et al. Imagenet pretraining: should we train from scratch? In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. p. 9224–9233.
- RÍOS, R. M. Mastócitos: funções biológicas e implicações patológicas. *Revista de Imunologia Veterinária*, v. 13, n. 2, p. 45–52, 2008.
- SALVI, M. et al. Histopathological classification of canine cutaneous round cell tumors using deep learning: a multi-center study. *Frontiers in Veterinary Science*, v.8, p.640944, 26 mar. 2021. DOI: [10.3389/fvets.2021.640944](https://doi.org/10.3389/fvets.2021.640944)
- SELVARAJU, R.R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, p.618–626, 2017. DOI: [10.1109/ICCV.2017.74](https://doi.org/10.1109/ICCV.2017.74)

SIMÕES, D. M. et al. Variabilidade interobservador na classificação histológica de mastocitomas caninos. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, v. 46, n. 3, p. 327–334, 1994.

SPRINGENBERG, Jost Tobias et al. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. DOI: <https://doi.org/10.48550/arXiv.1412.6806>

WANG, Yaqing et al. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, v. 53, n. 3, art. 63, p. 1–34, 2020. DOI: <https://doi.org/10.1145/3386252>.

WOLF, Thomas et al. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 2020. p. 38-45.

YAN, R.; HILL, S. Harnessing artificial intelligence for enhanced veterinary diagnostics. *Veterinary Clinical Pathology*, v. 53, n. 1, p. 45–56, 2024. DOI: <https://doi.org/10.1111/vcp.1340>.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: *European Conference on Computer Vision (ECCV)*. Springer, 2014. p. 818–833. DOI: https://doi.org/10.1007/978-3-319-10590-1_53.