PetHIS: A New Histopathological Image Dataset for Detecting Breast Cancer in Domestic Canines

Vinicius Barbosa da Silva*, Fabio Henrique Evangelista de Andrade†, and Omar Andres Carmona Cortes‡
* Bacharelado em Sistemas de Informação
Instituto Federal do Maranhão (IFMA), São Luis – MA Brazil
Email: viniciusbarbosa@acad.ifma.edu.br
†Departamento de Patologia
Universidade Estadual do Maranhão (UEMA), São Luis – MA Brazil
Email: fabioandrade@uema.br
‡Omar Andres Carmona Cortes
Departamento de Computação (DComp)
Instituto Federal do Maranhão (IFMA), São Luis – MA Brazil
Email: omar@ifma.edu.br

***Abstract*—Cancer is a severe disease that demands early discovery to increase the chances of survival and the probability of a cure. Its discovery is usually made through biopsy, a slow process subject to fatigue and psychological effects. Building intelligent computational visual tools to accelerate the diagnostic process is crucial in this context. The problem is that tools based on machine learning or deep learning, such as convolutional neural networks, depend on datasets to train the algorithm to create a proper model. Thus, this work proposes a dataset using histopathological images and augmentation of data from examinations of pets by the Laboratory of Veterinary Pathology. Tumor samples were obtained at the Veterinary Hospital School of the State University of Maranhão. Then, the dataset is used in a case study using VGG16 along with double transfer learning to demonstrate the applicability of the dataset. Results show that double transfer learning improves VGG16 efficiency, increasing metrics by around 5% using k-fold with k=5.**

***Index Terms*—cancer detection, dataset, pets, convolutional neural networks, VGG16.**

I. Introduction

According to a report by Radar Pet, a survey carried out in 2021 by the Companion Animal Commission (Comac) of the National Union for the Industry of Products for Animal Health (Sindan) found a substantial increase, reaching 30% more adoptions of pets during the pandemic [1]. The survey results also showed a growth in the percentage of tutors who started to see their pets as family members and even as children between 2019 and 2020. In contrast, the proportion of individuals considering them only as companion animals decreased [2]. These trends naturally contribute to improving the life expectancy of pets because their tutors care more than before.

Moreover, dogs and cats live longer than ever, with an average lifespan almost doubled compared to previous years. However, a side effect of this increase in life expectancy is that these animals face an increased risk of developing cancer, a chronic disease involving uncontrolled cell growth and tissue invasion. This can lead to uncomfortable symptoms, disability, and even death in humans and animals [3].

Despite the distinct biological categories, there are striking parallels in cancer development between humans and animals. Histopathologically, several forms of cancer exhibit substantial similarities in dogs and humans. This convergence is visible, for example, in cases of osteosarcoma, melanoma, lymphoma, and mammary carcinoma, all types of cancer that can manifest in both species [4]. Similarly, felines also show significant correspondences with human conditions. Let us take as an example the oral squamous cell carcinoma found in cats, which bears analogies with head and neck cancers in humans and cats. Furthermore, mammary tumors in cats show traits that resemble the "triple negative" phenotype identified in human breast cancers.

Thus, we can realize that the study of cancer in pets is as crucial as the study in humans, considering that both species share similarities in developing the disease. Particularly in dogs, they are more predisposed to cancer, often leading to poor prognosis and high mortality [5]. In the case of pet animals, mammary tumors have the highest incidences (16.8%) in females [6]. In unspayed female dogs, canine mammary tumor is the most common malignancy with a significantly higher mortality rate as compared to human breast cancer [7]. Therefore, diagnosing cancer as early as possible is crucial to increase the chances of cure or patient survival, and computational tools can be particularly useful for this purpose. Mainly because the diagnosis is made by incision biopsy with subsequent histopathological analysis, a highly time-consuming task that depends on the pathologist's background and is influenced by external factors such as fatigue and decreased attention [8].

The problem is that datasets containing histopathological images of pets are scarce. Therefore, this paper proposes the creation of a database comprising histopathological images of neoplasms in pets to serve as valuable tools for preventing and controlling cancer in domestic animals. Additionally, we present an experiment using the VGG16 [9] architecture, both with and without double transfer learning, to establish

a baseline for pattern recognition systems.

In this context, this investigation is divided into the following sections: Section II presents some related works; Section III shows how the dataset was created and enhanced by using data augmentation; Section IV shows some experiments proving the applicability of the new dataset; finally, Section V presents the conclusions of this work and future work.

II. Related Works

Histopathological datasets for human cancer detection are readily available and can be easily found on platforms like Kaggle. Many of these datasets are derived from well-known sources like BreakHIS [8], a renowned dataset for breast cancer detection in humans. In contrast, histopathological datasets for pets are scarce, primarily because most research on pets utilizes imaging techniques such as Magnetic Resonance (MR) or Computer Tomographies (CT), as seen in studies like Romeo et al. and Groheux, among others.

Additionally, research conducted by Michalski et al. [12] utilizes histopathological images as auxiliary tools to assess responses to treatments like chemotherapy. Similarly, Gamba et al. [13] perform histopathological exams and immunohistochemical staining assessments of invasive micro-papillary mammary carcinoma in dogs. While both works utilize histopathology, they lack comprehensive datasets dedicated to this purpose.

To the best of our knowledge, Kumar et al. [14] are the unique researchers to introduce a database specifically for canine mammary tumors, known as CMTHis. This dataset comprises 352 images from 44 clinical cases of canine mammary tumors at the Referral Veterinary Polyclinics at ICAR–Indian Veterinary Research Institute (IVRI), Izatnagar, India. In this context, our work aims to complement this existing resource by providing an additional tool for testing cancer detection algorithms.

III. The Dataset

Histology involves the detailed analysis of biopsy samples through a microscopic examination [15], in which the samples are subjected to a processing and fixation process in buffered formalin, diaphanization, embedded in paraffin and sectioning in a rotating microtome resulting in thin layers of tissue, stained with specific dyes and mounted on glass slides. These slides, in turn, are prepared from macroscopic sections of the biopsied tissue, which are treated with wax and specific dyes. This staining is crucial since it enables pathologists to differentiate the various cellular components, allowing a detailed analysis of tissue structure and architecture for diagnostic purposes. Through this procedure, it is possible to explore the microscopic characteristics of tissues, revealing views relevant to identifying diseases and anomalies.

In short, histopathological images are a fundamental visual representation of biological tissues at the microscopic level [15]. They are crucial in pathology, medical research, and disease diagnosis. These images offer detailed insights into tissues' structure and cellular characteristics, allowing clinicians, researchers, and scientists to analyze patterns and anomalies that may be imperceptible to the naked eye. Figure 1 illustrates a carcinoma with two magnitudes.

In this sense, the Laboratory of Veterinary Medicine of the State University of Maranhão made the histopathological images available, totaling 82 colored images with a 1944×2592 resolution and three RGB channels with a depth of 8 bits each. The images were divided into three classes: carcinomas, carcinosarcomas, and normal breast, resulting in 62 malignant and 20 non-cancer images comprising a small dataset. Table I shows the number of images according to their zooming magnitude.

TABLE I
Number of images per magnitude

Magnitude	Negative	Positive	Subtotal
4x	10	22	32
10x	10	21	31
40x	0	19	19
Total	20	62	82

Due to the low number of images, we built a binary dataset composed of benign and malignant classes called PetHIS.

A. Data Augmentation

As previously shown in Table I, the number of images is low; thus, we applied data augmentation in the images to increase the number. We used some transformations, as suggested by Hao et al. [16], such as rotations, mirroring, cropping, and zooming, to expose the model to a broader range of case possibilities. Consequently, increasing the model's ability to learn patterns reduces the risk of overfitting and improves generalization.

In our dataset, the data augmentation process allowed to multiply the number of neoplasm images by five and the number of breast cancer by ten, totaling 372 malignant and 220 normal breast images, *i.e.*, 592 images. To perform the transformation, we used a software so-called Imgaug [17] performing horizontal flip, vertical flip, and random cuts from 0 to 50 pixels as shown in Table II.

TABLE II
Augmentation parameters

Transformation	Parameter Value
Horizontal Flip	0.5
Vertical Flip	0.5
Random Crop	px(0, 50)

Some transformations happen at runtime, creating variations of original images without storing all possibilities in advance, enriching the training set, and improving the model's ability to generalize unknown possibilities. These transformations were carried out according to Table III.

IV. Experiments

To assess the suitability of the dataset for training deep learning algorithms, we evaluated using the VGG16 model.

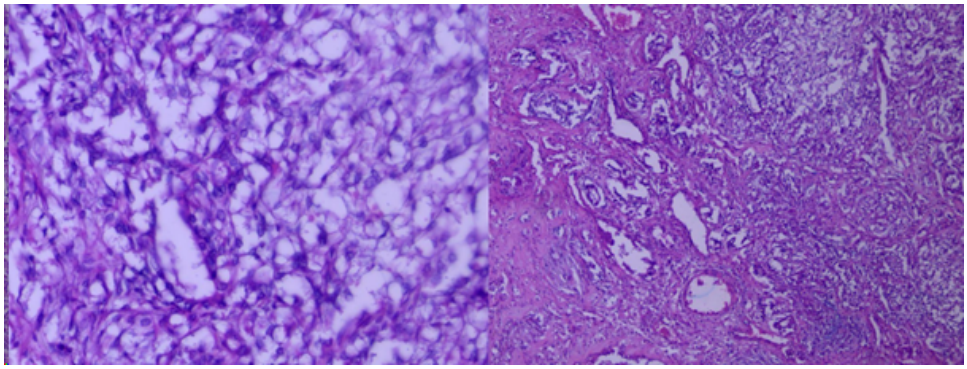

Fig. 1. Carcinoma with magnitude 40X on the left and magnitude 10X on the right

TABLE III
AUGMENTATION PARAMETERS

Transformation	Parameter Value
Rotation Range	20
Width Shift Range	0.2
High Shift Range	0.2
Horizontal Flip	TRUE

This well-established model has demonstrated good performance in biomedical image-based algorithms. Next, we provide details on the software and hardware configurations, metrics employed, and the obtained results.

A. Double Transfer Learning

Transfer Learning is a technique in machine learning where a pre-trained model in one task is used as a starting point for a new related task. The idea is to take advantage of the knowledge learned by the pre-trained model to improve performance in the new task [19]. Generally, the initial layers of the pre-trained model capture general, valuable features such as edges and textures, while the final layers capture more specific characteristics of the original task.

Double Transfer Learning is an advanced data processing technique combining two transfer learning stages to improve the performance of a machine learning or deep learning model. In the first stage, a pre-trained model is tuned for a related task, allowing it to capture general domain characteristics. The model is tuned for the specific interest task in the second stage. Matos et al. [20] proposed a dual transfer learning approach for classifying breast cancer histopathological images, improving classification accuracy from this perspective. In this context, we applied double transfer learning as presented in Figure 2, in which the VGG16 comes pre-trained with ImageNet, and we performed a second pre-trained test using BreakHIS.

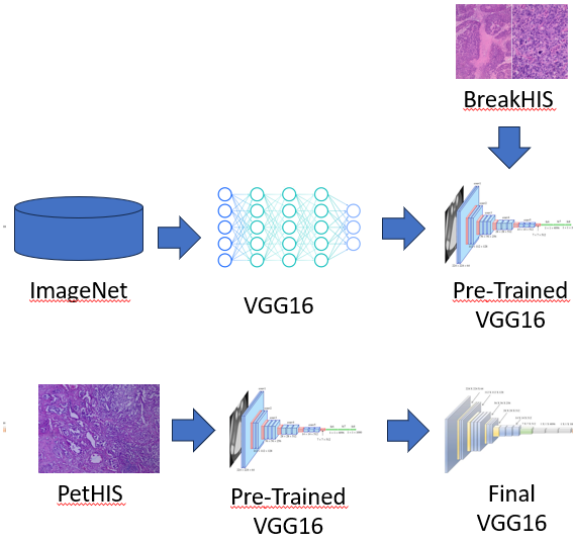

Fig. 2. VGG16 and its double transfer learning

B. Setup

For the data augmentation of the initial dataset of histopathological images of Pets, we used the Jupyter Lab, which is an interactive development interface that extends the functionality of the Jupyter Notebook, in which we use Python 3.11.2 as programming language, OpenCV 4.7.0 for image manipulation and visualization, Imgaug 0.4.0 for data augmentation, Numpy 1.23.5 for data manipulation and data transformation and Matplotlib 3.7.1 for image viewing.

For evaluation, we used cross-validation with the K-Fold technique. Cross-validation balances the assessment of model performance by splitting the data into multiple parts, ensuring that each fold contains mutually exclusive test sets. This technique improves experimental design, allowing a more comprehensive understanding of model performance and ensuring that it works well across all classes, regardless of significant imbalance in the data set. Furthermore, cross-validation can maximize the efficiency of using limited data to obtain more realistic estimations of the model's generalizability.

We performed two tests using k-fold cross-validation, with k=5 and k = 10, training a VGG16 model. Firstly, we used regular transfer learning with the ImageNet [18] dataset. Next, we utilized double transfer learning, employing a training stage with the BreakHIS [8], a histopathological dataset for detecting breast cancer in humans, before training the VGG16 with PetHIS images. The model is compiled with Adam Optimizer

with a parameter equal to 0.001, a binary cross entropy loss function (binary_crossentropy), and a batch size equal to 32 with ten epochs.

C. Metrics

To evaluate the classification models, we used the traditional metric for machine learning algorithms: Sensitivity (recall), Specificity, Accuracy, Precision, f1_score, and AUC. Thus, considering TP as the number of true positives, FP the number of false positives, TN the number of true negatives, and FN the number of false negatives, The five first metrics are obtained by Equations 1 to 5.

$$recall = \frac{TP}{TP + FN} \tag{1}$$

$$sensitivity = \frac{TN}{FP + TN} \tag{2}$$

$$acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$f1\_score = \frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

In this particular case, $recall$ plays an essential role in the model's evaluation because it uses the false negatives in its calculation, representing the worst-case scenario if the biopsy results in a false negative, leading the beloved pet to a certain death, a quick one depending on the cancer aggressiveness. Consequently, $f1\_score$ can be used as a general performance metric.

Finally, the area under the curve (AUC) is obtained by the Receiver Operating Characteristic Curve (ROC) by considering the TP rate (recall) and the false positive rate, which is the inverse of specificity ($1 - especificity$). The closer the AUC from 1, the better the model. In other words, the ROC curve makes it possible to compare several classifiers and determine which is superior based on different cutoff points. In practice, when the curve approaches the top of the Y-axis (AUC = 1), the classifier is more efficient.

D. Results

Table IV presents the mean metrics obtained through k-fold cross-validation with $k = 5$ and their respective standard deviations. It is evident from the table that the metrics exhibit an improvement of about 5% when employing double transfer learning. Moreover, the standard deviations remain relatively low, underscoring the beneficial impact of training the model with BreakHIS on overall performance.

In Figure 3, we notice the ROC curves for two distinct approaches: regular transfer learning (utilizing ImageNet) on the top and double transfer learning incorporating BreakHIS on the bottom. The central curve highlighted in the graph represents the mean, while the surrounding curves depict individual ROC curves per fold.

TABLE IV
METRICS WITH REGULAR TRANSFER LEARNING AND DOUBLE TRANSFER LEARNING WITH K = 5

Metric	Regular $mean \pm std$	Double TL $mean \pm std$
Accuracy	0.898 ± 0.042	0.946 ± 0.036
Precision	0.920 ± 0.022	0.947 ± 0.035
Recall	0.899 ± 0.041	0.892 ± 0.018
f1_score	0.900 ± 0.045	0.946 ± 0.036

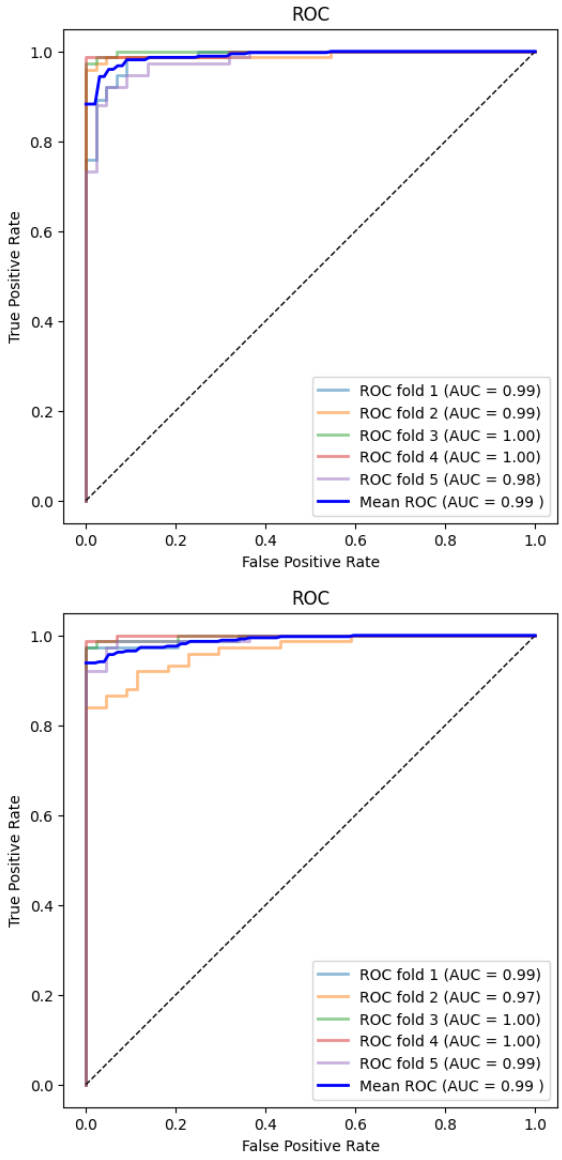

Fig. 3. ROC curve with regular transfer learning and double transfer learning with k = 5

A visual inspection immediately reveals that the mean curve for double transfer learning tends towards a near-perfect performance, even though when AUC are similar, the blue curve with double transfer learning is closer to one than the other one.

Table V shows the mean metrics reached through k-fold

cross-validation with $k = 10$, along with their respective standard deviations. The improvements are lower than the VGG with k = 5, probably because the test dataset is minor. Nonetheless, the improvements are still toward double transfer learning.

TABLE V
METRICS WITH REGULAR TRANSFER LEARNING AND DOUBLE TRANSFER LEARNING WITH K = 10

Metric	Regular $mean \pm std$	Double TL $mean \pm std$
Accuracy	0.914 ± 0.029	0.929 ± 0.021
Precision	0.927 ± 0.020	0.934 ± 0.016
Recall	0.910 ± 0.028	0.924 ± 0.021
f1_score	0.910 ± 0.031	0.924 ± 0.023

In Figure 4, we still notice the ROC curves for regular transfer learning (utilizing ImageNet) on the top and double transfer learning incorporating BreakHIS on the bottom. The central curve highlighted in the graph represents the mean, while the surrounding curves depict individual ROC curves per fold. Visually, the double transfer learning approach is still closer to the ideal ROC curve.

V. CONCLUSIONS

This study introduced the development of a dataset comprising histopathological images aiming to serve as a tool for training machine learning models to detect breast cancer in canines. Data augmentation techniques were employed to enhance the dataset's diversity, resulting in a fivefold increase in malignant cases and a tenfold increase in non-cancer images. Next, we investigated the dataset's applicability using the VGG16 model with regular and double transfer learning. The results revealed that double transfer learning can deliver improvements in metrics around 5% in k-fold with k = 5.

In terms of future work, several options are being considered:

(i) Expanding the dataset's size by acquiring additional histopathological cases and applying data augmentation techniques once again.
(ii) Assessing the dataset's performance using other well-known CNN architectures, employing regular and double transfer learning methodologies.
(iii) Exploring the possibility of merging this dataset with the one introduced in Kumar et al.'s study [14], followed by testing various VGGs and ensemble CNN architectures to enhance the detection capabilities further.

These proposed forthcoming steps aim to improve the dataset's robustness and applicability in canine breast cancer detection, contributing to more effective diagnostic tools and research in this domain. The dataset can be downloaded at https://zenodo.org/records/8322517.

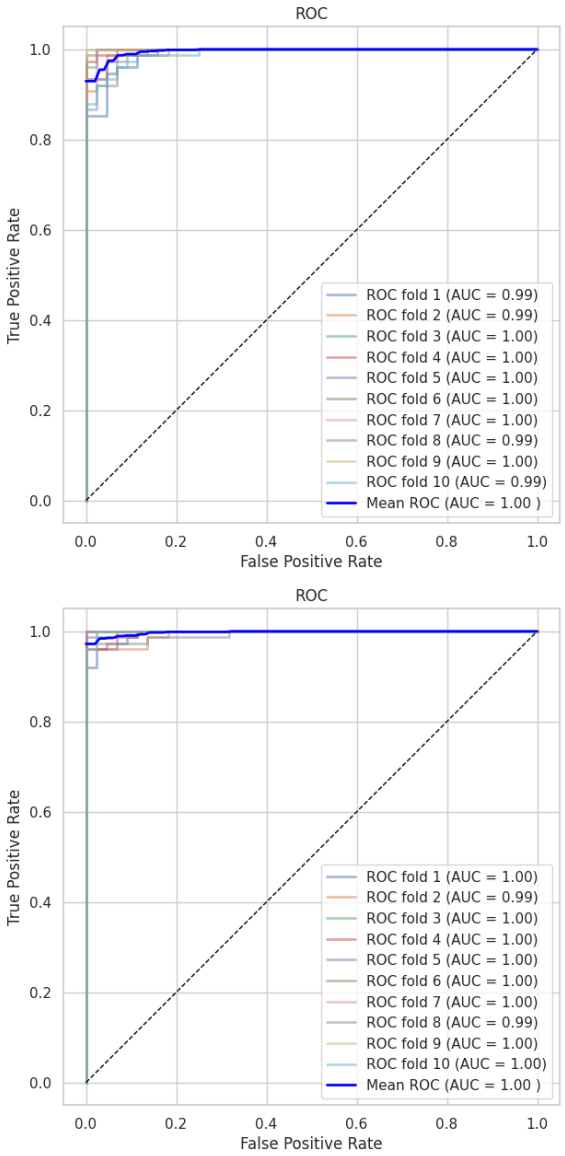

Fig. 4. ROC curve with regular transfer learning and double transfer learning with k = 10

REFERENCES

[1] Comissão de Animais de Companhia (COMAC), Coletiva de imprensa radar 2021 mercado de pet na pandemia. Brasil: Comac, 2021.

[2] Comissão de Animais de Companhia (COMAC), Pesquisa Radar Pet: Brasil conta com a segunda maior população pet do mundo. Available at: https://sindan.org.br/release/pesquisa-radar-pet-brasil-conta-com-a-segunda-maior-populacao-pet-do-mundo/. Access date: Ago-17th-2023

[3] Vessoni, A., Pesquisadores da Unesp participam de primeiro consenso mundial sobre hemangiossarcoma em cães. Available at: https://jornal.unesp.br/2023/05/03/pesquisadores-da-unesp-participam-de-primeiro-consenso-mundial-sobre-hemangiossarcoma-em-caes/. Access date: Ago-17th-2023

[4] Cannon, C.M., Cats, Cancer and Comparative Oncology. Vet. Sci. 2015, 2, 111-–126.

[5] Grüntzig, K.; Graf, R.; Hässig, M.; Welle, M.; Meier, D.; Lott, G.; Erni, D.; Schenker, N.; Guscetti, F.; Boo, G. et al. The swiss canine cancer registry: a retrospective study on the occurrence of tumours in dogs in switzerland from 1955 to 2008, J. Comp. Pathol., 152 (2–3), 2015, pp. 161–171.

[6] Salas, Y.; Márquez, A.; Diaz, D.; Romero, L., Epidemiological study of mammary tumors in female dogs diagnosed during the period 2002–2012: a growing animal health problem, PLoS One, 10 (5) (2015),

p. e0127381.
[7] Egenvall, A.; Bonnett, B. N.; Öhagen, P.; Olson, P.; Hedhammar, V.; Euler, von H., Incidence of and survival after mammary tumors in a population of over 80,000 insured female dogs in sweden from 1995 to 2002, Prev. Vet. Med., 69 (1-2), 2005, pp. 109–127
[8] Spanhol, F. A.; Oliveira, L. S.; Petitjean, C.; Heutte, L., A Dataset for Breast Cancer Histopathological Image Classification. Ieee Transactions On Biomedical Engineering, [S.L.], v. 63, n. 7, p. 1455-1462, Jul. 2016. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tbme.2015.2496264.
[9] Simonyan, K.; Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv, 2015.
[10] Romeo, V,; Moy, L.; Pinker, K., AI-Enhanced PET and MR Imaging for Patients with Breast Cancer, PET Clinics, Volume 18, Issue 4, 2023, pp. 567–575.
[11] Groheux, D., FDG-PET/CT for Primary Staging and Detection of Recurrence of Breast Cancer, Seminars in Nuclear Medicine, v. 52, Issue 5, 2022, pp. 508–519.
[12] Michalski, K.; Stoykow, C.; Bronsert, P.; Juhasz-Böss, I.; Meyer, P. T.; Ruf, J.; Erbes, T.; Asberger, J., Association between gastrin-releasing peptide receptor expression as assessed with [68Ga]Ga-RM2 PET/CT and histopathological tumor regression after neoadjuvant chemotherapy in primary breast cancer, Nuclear Medicine and Biology, v. 86-87, 2020, pp. 37–43.
[13] Gamba, C. O.; Dias, E. J.; Ribeiro, L. G. R.; Campos, L. G. R.; Estrela-Lima, A.; Ferreira, E.; Cassali, G. D., Histopathological and immunohistochemical assessment of invasive micropapillary mammary carcinoma in dogs: A retrospective study, The Veterinary Journal, v. 196, Issue 2, 2013, pp. 241–246.
[14] Kumar, A.; Singh, S. K.; Saxena, S.; Lakshmanan, K.; Sangaiah, A. K.; Chauhan, H.; Shrivastava, S.; Singh, R. K., Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer, Information Sciences, v. 508, 2020, pp. 405–421, https://doi.org/10.1016/j.ins.2019.08.072.
[15] Belsare, A. D., Histopathological Image Analysis Using Image Processing Techniques: an overview. Signal & Image Processing: An International Journal, [S.L.], v. 3, n. 4, 2012, pp. 23–36.
[16] Hao, R.; Namdar, K.; Liu, L.; Haider, M. A.; Khalvati, F., A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using Convolutional Neural Networks. Journal Of Digital Imaging, [S.L.], v. 34, n. 4, 2021, pp. 862–876, http://dx.doi.org/10.1007/s10278-021-00478-7.
[17] IMGAUG. Imgaug. Available at: https://imgaug.readthedocs.io/en/latest/. Access date: Mar-4th-2023.
[18] Krizhevsky, A.; , Alex; Sutskever, I.; Hinton, G. E., ImageNet classification with deep convolutional neural networks. Communications Of The Acm, [S.L.], v. 60, n. 6, p. 84–90, 24 maio 2017. Association for Computing Machinery (ACM), http://dx.doi.org2017/10.1145/3065386.
[19] Iman, M.; Arabnia, H. R.; Rasheed, K., A Review of Deep Transfer Learning and Recent Advancements. Technologies, [S.L.], v. 11, n. 2, 2023, MDPI, http://dx.doi.org/10.3390/technologies11020040
[20] Matos, J. de; Britto, A. de S.; Oliveira, L. E. S.; Koerich, A. L., Double Transfer Learning for Breast Cancer Histopathologic Image Classification, 2019 International Joint Conference On Neural Networks (Ijcnn), [S.L.], jul. 2019. IEEE. http://dx.doi.org/10.1109/ijcnn.2019.8852092.