COBEM 2021 Florianópolis - Brasil

ABCM Associação Brasileira de Engenharia e Ciências Mecânicas

26th ABCM International Congress of Mechanical Engineering
November 22-26, 2021. Florianópolis, SC, Brazil

COB-2021-1052 - PILOT: ARTIFICIAL INTELLIGENCE APPLIED TO THE IDENTIFICATION OF BONE CHANGES IN CANINE PELVIC RADIOGRAPHIES

Bárbara Emmanuelle Sanches Silva
"Graduate Program in Mechanical Engineering" – Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 – Pampulha, Belo Horizonte - MG
barbarasilvadvm@gmail.com;

Rudolf Huebner
Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 – Pampulha , Belo Horizonte - MG
rudolf@ufmg.br

Anelise Carvalho Nepomuceno
Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 – Pampulha, Belo Horizonte - MG
anelise-imagem@vet.ufmg.br;

Carolina Costa Cardoso
Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 – Pampulha, Belo Horizonte - MG
carolinaccardoso@vetufmg.edu.br

Jonathan Cristovão Ferreira da Silva
Universidade Federal de Ouro Preto, Campus Morro do Cruzeiro – Bauxita, Ouro Preto - MG
jonathancristovao13@gmail.com

Abstract. *Artificial intelligence has been applied to the health sciences as a new method of metrology and instrumentation as it uses a database to create statistical reference standards related to this database. In veterinary medicine, the low structure and the great diversity of data results in a still incipient use of this technology. The application of artificial intelligence as an auxiliary diagnostic method in images is already used in human medicine with high precision and accuracy to aid decision making, especially, in the identification of changes in physiological patterns in radiographies. This work proposes a comparison of application of a convolutional neural network (CCN) and the neural network Multi-layer Perceptron for identification of bone changes of different causes in ventrodorsal radiographies of dogs of different breeds and sizes. Images with normal radiological patterns and with different radiological alterations were collected from different radiology clinics to create the database, dividing the data between normal and not normal patterns. Database expansion techniques were used to avoid overfitting, resized and the images were divided into training (90% and 83%) and testing (10% and 17%). The performances of the neural networks are compared to the radiological reports issued by human specialists. The results of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and F1-Score are obtained. The results aim to show that, even for data with high variability, the application of neural networks as an auxiliary method of diagnosis is a valuable resource in veterinary medicine and enlargement of the dataset is recommended.*

Keywords: *Convolutional neural network, machine learning, radiology, veterinary medicine.*

1. INTRODUCTION

The development of stochastic models for understanding and predicting the behavior of biological systems is a study in constant focus, whether aimming the development of classification systems for the physiological states of individuals, to evaluate population dynamics or applied to epidemiology. Predictions related to the probability of occurrence of a disease help in decision-making regarding conducts in relation to the diagnosis and its treatment, which can significantly influence the success obtained (Santos, 2018). These models commonly adopt simplification of scenarios to verify specific hypotheses surrounding the problem.

Given the complexity of biological systems, machine learning (ML) algorithms have been widely used in the creation of these models, considering their focus on predicting new observations from a set of input examples rather than interpreting a small number of specific parameters (Breiman, 2001; Tarca, 2007). The increase in data availability allows a more comprehensive assessment of the influences on a given response and the processing of this factors multiplicity and their interactions is optimized and made possible with ML techniques.

Imaging diagnosis is one of the areas within the health sciences with roominess for neural network analysis applications as its interpretation depends on various factors and, particularly, on the experience of the professional who produces the reports in relation to the pathology to be measured. The application of neural networks for this purpose can expand the diagnostic capacity and quantify this diagnosis in relation to a statistical pattern of a model developed on a given database, favoring the obtaining of an increasingly accurate final diagnosis by professionals (de Bruijne, 2016). This approach has been used extensively by human medicine to help detect different diseases based on models of different neural networks, whether machine learning or deep learning, including convolutional neural networks (Wang, 2021).

Park et al. (2020) evaluated the performance of a convolutional neural network (CNN) for classification of child hip dysplasia, obtaining an accuracy greater than 95%, comparable to results of experienced professionals in the field, based on a 5,076 images database. Cheng et. al (2019) verified the accuracy and sensitivity of deep learning algorithms for identifying and locating fractures in human pelvic radiographs modeled with a 25,505 radiographic images (Xs) database, obtaining an accuracy in identifying the lesion of 95.9%. Saraiva et. al (2019), using 5,840 images, compared the performance of two neural networks – the Multi-layer Perceptron (MP) and the CNN – to identify pneumonia in human chest Xs, obtaining results of accuracy of 92.2% and 94.4%, respectively.

Regardless of the network used, the image processing method, the adequate previous classification and the data volume for training the network are fundamental for the models performance. The volume of data is especially important in veterinary medicine, considering the high variability of species, breeds and size of patients. Expressive external pressures such as health insurers, public health policies and pharmaceutical industries, in human medicine, favor the massive databases structuring shared between different clinics and entities, which does not easily occur in veterinary medicine (Lustgarten, 2020). Even so, it is possible to observe an increase in the interest in the development of applications using artificial intelligence in this area, given its potential to raise the accuracy of diagnoses and improve decision-making.

Boissady et al. (2020), with a 22,000 Xs images database, compared the performance of CNN from the open-source Pytorch framework in identifying 15 classes of diagnosable disorders in chest Xs of dogs and cats among with the results of a group composed of veterinarians with different experience levels, obtaining an error rate of 10.7% for CNN, 17.2% for veterinarians group and 16.8% veterinarians assisted by CNN. Li et al. (2020), in turn, evaluated the application of the Visual Geometry Group 16 neural network to identify left atrial enlargement in dogs using 792 chest Xs, obtaining an accuracy of 82.71% for the network, compared to 82 .71% by certified professionals.

The increase in complexity of the evaluated region and the number of possible associated pathologies demands greater attention in relation to the number of images used. Zhou et al. (2021) studied the possibility of circumventing the low images availability in veterinary medicine using 40.005 images of human chest Xs from the Musculoskeletal Radiograph (MURA) dataset in a pre-training of the neural network to verify the performance of the network in identifying different pathologies in 15 distinct anatomical regions and an overall 500 images database of dogs and cats. They observed that the pretraining with human dataset increased by three times the network's ability to locate pathologies and reduce the misclassification of soft tissue as a lesion, not being enough to improve the prediction of false injuries in joints, a fact attributed to the absence of limbs joints in the MURA examples.

Thus, the main objective of this study is to stimulate and demonstrate the importance of structuring an aggregated database in Veterinary Medicine to develop accurate diagnostic aid tools using machine learning, especially when applied to regions of high radiographic complexity, being a pilot project to develop a model for the analysis and classification of pelvis radiographs of dogs as an auxiliary diagnostic method.

2. METHODOLOGY

The design proposed for the present work is a retrospective diagnostic study as it evaluates the accuracy of a new diagnostic auxiliary method on radiographic images previously classified by a veterinary radiologist with more than 10 years of experience. Two supervised ML algorithms were used comparatively to verify the performance of each in the presented training scenario. The CNN MobileNet-V2 (MNV2) algorithm and the ML Multi-layer Perceptron (MP) algorithm were used to classify the images and the accuracy, precision and sensitivity of each were compared based on the previous classification of the images made by the professionals – the study's gold standard. The same database was used in the training of both networks.

2.1 DATABASE

The database was assembled using Xs of the pelvic region of dogs, generated between 2019 and 2021, and their respective reports provided by the Veterinary Hospital of the Federal University of Minas Gerais. Overall, 155 images were obtained, 81.9% in the ventrodorsal position and 18.1% in the laterolateral position. In the present study, only ventrodorsal images with extended hind limbs animals position were used considering their higher prevalence and the need for a large image bank for training and testing of neural networks. The images were separated into two groups:

normal or altered radiographic patterns, based on the results expressed in the respective reports. The alterations were not discriminated, and may include fractures, bone deformities, tumors, joint degenerations, among others. In addition, a selection was not performed based on the size, breed, sex or age of the animals, considering that one of the goals of the design of the algorithm is to be able to act with precision regardless of the particularities of the area in which it is applied, being necessary to be trained as per your requirements.

A total of 126 images were obtained, 71 with an altered pattern and 55 with a normal radiological pattern, manually separated based on the report. The images were extracted in their original format – DICOM (Digital Imaging and Communications in Medicine). The size of the images ranged between 1722x1430 and 2446x2010 pixels, depending on the plate used to generate the image on the device and the output format selected by the radiologist. An algorithm for extracting only the pixel matrix from the DICOM image was generated.

2.2 IMAGE PROCESSING

Initially, the images were converted into JPEG (Joint Photographics Experts Group) format using an adaptation of the Python algorithm (version 3.7.10) proposed by Shafique (2020) to minimize the loss of image quality, standardize and automate the conversion and to avoid using external software. The algorithm optimized the application of saturation and contrast by reducing the DICOM image format from 12 bytes to 8 bytes in JPG, using for this the OpenCV (version 4.5.2) and Pydicom (version 1.2.4) packages for reading and writing.

After conversion, each type of network used received a distinct image processing, as recommended in the frameworks documentation and literature. The MP neural network training is based on the image's histograms, not its pixel matrix, as is the case of the deep learning network. Thus, it was necessary to pre-process the JPG images format for this model, initially converting the BGR color standard to the HSV standard to obtain the saturation channel, using the OpenCV package, considering that the images were in grayscale only. Binary pattern classification was sequentially added for each histogram and the data saved as a matrix.

At the same time, considering the prerogatives described in the literature for applying deep learning algorithms (Sandler), all images were resized to the standard size of 224x224 pixels for MNV2 training. The database was then randomly divided between training and test images in two proportions for later comparison of the obtained accuracy, being 90% and 10%, and 83% and 17%, respectively. A two-level factorial study was set up, the first factor being the neural network applied and the second the database split ratio between training and testing. The random images split in each training was performed using the Scikitlearn framework (version 0.24.2), with the target being the binary classification of the categories "Altered" for radiological patterns with alterations and "Normal" for both network types.

The increase in CNN robustness by expanding the training database was performed using the Tensor Flow Keras package (version 2.3.0) for random application of rotation (±10°), expansion (±0.15), horizontal inversion and vertical and horizontal displacement (±0.2 each).

2.3 MODEL CONSTRUCTION AND EVALUATION

The first model based on MP was trained using the solver "adam", alpha reduction factor of 1.0E-04, function and activation "ReLU6", maximum number of iterations of 3000 and initial learning rate of 1.0E-04. After training, the test was conducted and the confusion matrix generated as shown in Table 1.

Table 1 - Confusion Matrix

		Predicted label	
		Altered	**Normal**
True Label	**Altered**	True Positive	False Negative
	Normal	False Positive	True Negative

The results in which the network correctly classified the test images are classified as True (positive - TP or negative - TN) and the images erroneously classified by the network were added to the results classified as False (positive - FP or negative - FN), being positive the results for images with radiological alterations and negative without alteration.

MNV2, developed by Mark Sandler et al (Sandler, 2018), was used for possible later mobile applications and for its optimized image processing capacity. The model's architecture is based on the use of 3 convolution layers, the first being data expansion, the second depth wise and the third projection.

The layer density chosen for training was 128 with a dropout of 0.5. A batch size of 24 and 40 epochs were used as input configuration, in addition to pre-training the model with the ImageNet database. The initial learning rate was first set to 0.01, being reduced by 20% with each training whenever observed an overall increase in the cost function between epochs. The confusion matrix of this model was also generated based on the test image classification results.

The models performance was measured using the parameters of Accuracy, Specificity, Sensitivity and the F1¬Score, calculated from the confusion matrix obtained for each model, according to equations (1) to (4):

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Sensibility\ ou\ Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = 2 * \frac{Precision * Sensibility}{Precision + Sensibility} = 2 * \frac{\left(\frac{TP}{TP+FP}\right)*\left(\frac{TP}{TP+FN}\right)}{\left(\frac{TP}{TP+FP}\right)+\left(\frac{TP}{TP+FN}\right)} \quad (4)$$

Accuracy corresponds to the model's ability to correctly predict any class, whereas sensitivity and specificity represent its ability to correctly predict each classification, positive or negative, respectively. The F1-Score, on the other hand, refers to a harmonic average between precision metrics, which is the ability to correctly classify positive cases, and sensitivity, being more sensitive to low values for any of these metrics (Geron, 2019; Raschka, 2019).

Finally, the algorithm was designed to present to the final user the image converted and already classified by the models, including the probability of correct classification according to the training.

3. RESULTS AND DISCUSSION

The test database was structured as shown in Table 2.

Table 2 - Test database structuring based on training and testing rate.

Train-Test Ratio	Altered	Normal	Total
0,9:0,1	8	6	14
0,83:0,17	12	10	22

These ratios aimed to define an even number of images for each group to facilitate the identification of a performance close to the random probability of 50% for each of the two classes, an expected result in an unstructured and disconnected training. Higher values for the test group were initially disregarded as their increase would impact the already reduced training database.

Figure 1 exemplifies the output of the algorithms with the classification performed by neural networks.

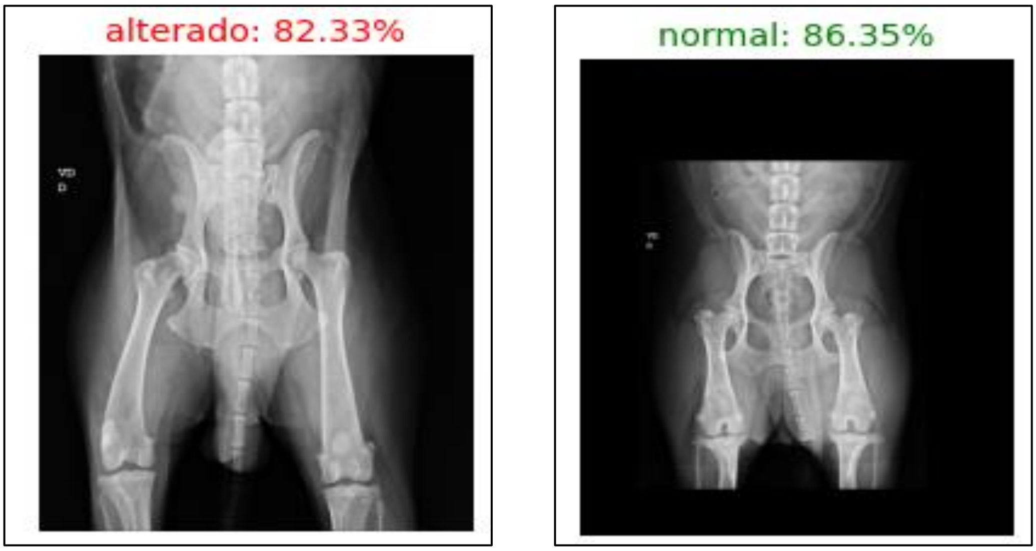

Figure 1 - Output of the Xs classification algorithm indicating the classification and probability of success of the model in real life examples.

Due to the variability of limb positioning, image size, size and individual characteristics of the animals from which the images were obtained, in addition to the number of images available for training the networks, 15 training repetitions were performed for each model, seeking to minimize the effects of unwanted random selection of more characteristic images in the test groups for one or another model in each training. Confusion matrices with normalized prediction mean values and standard deviation for each factor studied were then generated as shown in Tables 3 to 6.

Table 3 - Confusion matrix for the model using the MP for the 10% test image ratio.

		Predicted label	
		Altered	**Normal**
True Label	**Altered**	0.7±0.2	0.3±0.2
	Normal	0.4±0.2	0.6±0.2

Table 4 - Confusion matrix for the model using the MNV2 network for the 10% test image ratio.

		Predicted label	
		Altered	**Normal**
True Label	**Altered**	0.8±0.1	0.2±0.1
	Normal	0.5±0.2	0.5±0.2

Table 5 - Confusion matrix for model using MP for test image ratio of 17%.

		Predicted label	
		Altered	**Normal**
True Label	**Altered**	0.7±0.1	0.3±0.1
	Normal	0.4±0.2	0.6±0.2

Table 6 - Confusion matrix for the model using the MNV2 network for the test image ratio of 17%.

		Predicted label	
		Altered	**Normal**
True Label	**Altered**	0.93±0.08	0.07±0.08
	Normal	0.5±0.2	0.5±0.2

The data presented in Tables 3 to 6 suggest, for any proportion of database division, a possible improvement in performance using the MNV2 network to predict altered radiological patterns. Setting the test ratio at 10% or 17%, an increase of 14.3% or 32.9%, respectively, was observed in the TP prediction, changing the network from MP to MNV2. For TN prediction, a reduction of 16.7% was observed in both configurations. It is also possible to verify a satisfactory prediction rate for the identification of altered patterns for the MNV2 network, with the range with lower values between 70 and 90% of hit chance.

In none of the observed scenarios it was possible to satisfactorily predict the occurrence of normal patterns, since the range of uncertainty with the standard deviation encompassed the probability of correctness of 50%, expected in a normal distribution of a random binary classification. However, these values these values indicate that the algorithm does not indiscriminately classify all radiographs as altered.

It is also possible to verify a possible improvement in the performance of networks using a larger test database, since a smaller deviation was obtained in both models and the MNV2 model achieved better performance in identifying the altered class. Confirmation of these inferences was performed through the statistical analysis of the Design of the Experiment (DOE), conducted in the Minitab ® software, over this study previously determined metrics, namely: accuracy, sensitivity, specificity and F1-Score. The results for each factor are shown in Table 7 and the DOE analysis (α = 5%) summarized in Table 8.

Table 7 - Accuracy, sensitivity, specificity and F1-Score results of the models for each test-ratio

Model	**Test Ratio**	**Accuracy**	**Sensibility**	**Specificity**	**F1-Score**
MP	10%	0,6±0,2	0,7±0,2	0,6±0,2	0,7±0,2
MP	17%	0,6±0,1	0,7±0,1	0,6±0,2	0,7±0,1
MNV2	10%	0,68±0,08	0,8±0,1	0,5±0,2	0,74±0,06
MNV2	17%	0,72±0,05	0,93±0,07	0,5±0,2	0,79±0,03

Table 8 - P-Value obtained by DOE analysis.

	P-Value			
Factor	**Accuracy**	**Sensibility**	**Specificity**	**F1-Score**
Test-Ratio	0,026	0,026	0,441	0,026
Model	<0,001	<0,001	0,021	<0,001
Model*Test-Ratio	0,363	0,363	0,961	0,363

Binary image classification models containing two classes have the characteristic that sensitivity corresponds to the normalized TP value, and specificity corresponds to the normalized TN fraction. Accuracy, on the other hand, represents the total correctness capacity of the models. As in the previous results, the data in Table 7 suggest an improvement in the final performance and dispersion with the MNV2 network and test ratio of 17%, a result also observed in the F1-Score. An increase of 13.3% and 20.0% was observed for the accuracy in the test ratios of 10% and 17%, respectively, and in the F1-Score of 5.7% and 12.9% using the MNV2 network instead of MP. The variation of the test ratio for each neural network resulted in a 5.9% increase in accuracy and 6.8% in F1-Score increasing the test rate from 10% to 17% for the MNV2 net, not being observed change in the average value of these factors for the MP network.

The reliability of the influence of the model and test ratio factors on the responses is confirmed through the P-Value shown in Table 8. As expected, it is not possible to state that the increase in the test database influences the specificity and although the choice the network has a significant influence (P<0.05), it was not possible to observe an improvement in the performance in the predictive capacity of normal images with the association of these two factors for this experiment. On the other hand, both accuracy and sensitivity and F1-Score, which has a non-linear dependence on sensitivity, were significantly influenced by both the model choice (P<0.001) and the test ratio (P<0.05) separately, showing no significant influence with the association of variation in these two factors. Finally, the training boxplots for each F1-Score factor are shown in Figure 2.

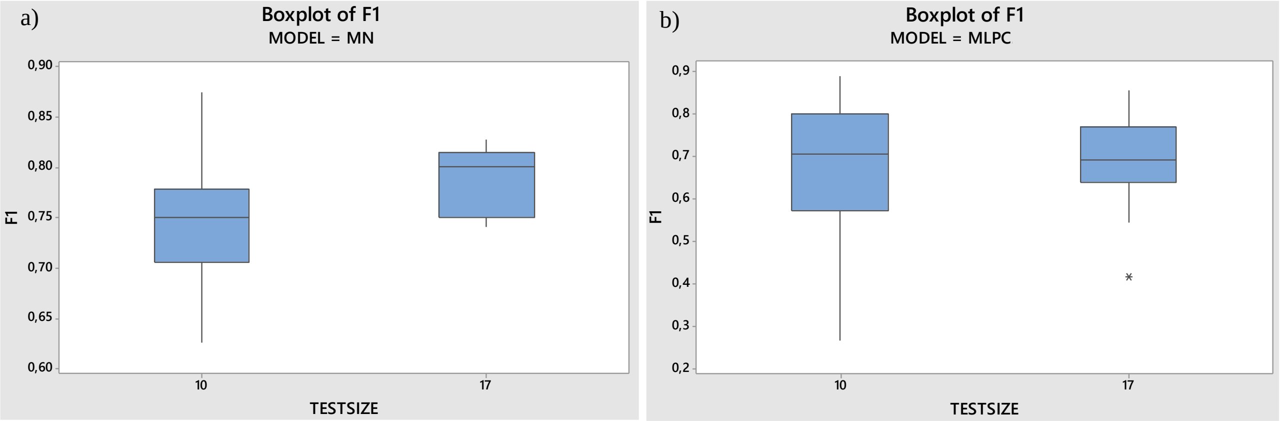

Figure 2 - Boxplot of the F1-Score obtained as a function of the neural network and the test fraction a) for the MNV2 network and b) MP.

It is possible to observe, through Figure 2 a and b, the importance of the statistical analysis of the model for conditions similar to those adopted in this experiment, that is, high data variability and low availability. The isolated training and testing of networks under these conditions may wrongly discourage their development and the continuity of their application. The MP model, for a test-ratio and 10%, showed F1-Score values close to 90%, but also values lower than 30%, a result that is not applicable when the objective is to increase the diagnostic accuracy. The increase in the volume of images for testing, although reducing the number of images to train the network, for both neural networks, favored the reduction of the F1-Score dispersion, as mentioned above, confirming the importance of expanding the image bank.

4. CONCLUSION

The results obtained indicate that it was possible to train a network based on the MNV2 structure with an accuracy between 60.0 and 77.0%, with its performance being better for the prediction of altered radiological patterns compared to those without alteration. The sensitivity of this model ranged between 70.0% and 100.0%, but the specificity ranged between 30.0% and 70.0%. Even so, the F1-Score obtained was between 68.0% and 82.0%.

The MP-based network obtained accuracy and specificity values ranging from 40.0% to 80.0%; and sensitivity and F1-Score between 50.0% and 90.0%, demonstrating an average performance lower than MNV2 and a higher standard deviation of training. It is believed that the values obtained for the standard deviations of each metric were a consequence of the volume of images used for training and testing the models. Through the statistical analysis of the results, it was possible to infer that the choice of the model has a significant influence on them ($P<0.05$), but it is not possible to confirm the significance of the increase in the test ratio from 10% to 17% on the results.

The results obtained for an original database of 125 pelvic x-ray images of dogs, for this pilot work, were considered promising, particularly, for the MNV2 network, as they are close to the results obtained in the literature for Xs analyses in animals and humans, comparatively, using a massive database. However, yet the accuracy falls short of the final goals for the development of an application that can be used as a diagnostic aid tool for precision medicine, so that the continuity of the study with database augment may be justifiable. In addition, the localization of lesions using heat maps is a fundamental step to verify the points considered in the classification of images by the neural network. Such a study can further elucidate the limits of structuring a statistical model based on its input data.

5. REFERENCES

BOISSADY E., DE LA COMBLE A., ZHU X., HESPEL A.M.. Artificial intelligence evaluating primary thoracic lesions has an overall lower error rate compared to veterinarians or veterinarians in conjunction with the artificial intelligence. Vet Radiol Ultrasound. 2020;61:619–627.

BREIMAN, L. Random forest. Machine Learning. 2001,45, 5–32.

DE BRUIJNE M.. Machine learning approaches in medical image analysis: From detection to diagnosis. Medical Image Analysis. Volume 33, 2016. Pages 94-97.

CHENG, C. T., HO, T. Y., LEE, T. Y., CHANG, C. C., CHOU, C. C., CHEN, C. C., CHUNG, I. F., & LIAO, C. H. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. European radiology, 29(10), 5469–5477. 2019.

GERON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd Edition. 2019.

LI S., WANG Z., VISSER L.C., WISNER E.R., CHENG H. Pilot study: Application of artificial intelligence for detecting left atrial enlargement on canine thoracic radiographs. Vet Radiol Ultrasound. 2020;61:611–618.

LUSTGARTEN J. L., ZEHNDER A., SHIPMAN W., GANCHER E., WEBB. T. L. Veterinary informatics: forging the future between veterinary medicine, human medicine, and One Health initiatives—a joint paper by the Association for Veterinary Informatics (AVI) and the CTSA One Health Alliance (COHA), JAMIA Open, Volume 3, Issue 2, July 2020, Pages 306–317.

PARK H.S., JEON K., CHO Y.J., KIM S.W., LEE S.B., CHOI G., LEE S., CHOI Y.H., CHEON J.E., KIM W.S., RYU Y.J., HWANG J.Y. Diagnostic Performance of a New Convolutional Neural Network Algorithm for Detecting Developmental Dysplasia of the Hip on Anteroposterior Radiographs. Korean J. Radiol. 2021 Apr;22(4):612-623.

RASCHKA S., MIRJALILI, V. Learning Best Practices for Model Evaluation and Hyperparameter Tuning. Python Machine Learning — 3rd Edition. 2019.

SANDLER M., HOWARD A., ZHU M., ZHMOGINOV A., CHEN L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.

SANTOS, H. G. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. 206f. Tese (Doutorado) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2018.

SARAIVA, A.; SANTOS, D.; COSTA, N.; SOUSA, J.; FERREIRA, N.; VALENTE, A. AND SOARES, S. Models of Learning to Classify X-ray Images for the Detection of Pneumonia using Neural Networks. In Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOIMAGING, ISBN 978-989-758-353-7; ISSN 2184-4305, pages 76-83. 2019.

TARCA A.L., CAREY V.J., CHEN X-w., ROMERO R., DRĂGHICI S. Machine Learning and Its Applications to Biology. PLoS Comput Biol 3(6): e116. 2007.

WANG J., ZHU H., WANG S., ZHANG Y. A Review of Deep Learning on Medical Image Analysis. Mobile Networks and Applications. 26. 2021.

ZHOU S., AHN E., FULHAM M., KIM J. Intelligent Interpretation of Veterinary Musculoskeletal X-rays Trained with Human Thoracic Limb X-rays, 10 June 2021.

6. RESPONSIBILITY NOTICE

The author(s) is (are) the only responsible for the printed material included in this paper.