# COMBINATION OF DEEP LEARNING AND MACHINE LEARNING FOR AI-ASSISTED DIAGNOSTICS

https://doi.org/10.56238/sevened2025.008-005

Fábio Lofredo Cesar[1] and Hygor Santiago Lara[2].

## ABSTRACT

This work proposes a hybrid system that combines ResNet (Residual Network - widely recognized for its impact on deep learning, being a milestone in the area of computer vision) with Extremely Random Trees (Extra Trees) to classify chest X-ray images and assist in the detection of diseases. The approach uses the Transfer Learning technique, where ResNet, previously trained, is used to extract relevant characteristics from the images. Then, the Extra Trees algorithm performs the classification based on these characteristics. In the initial stage, using only ResNet combined with a small neural network, we obtained an accuracy of 95.40% in validation and 79.33% in tests. With the implementation of the hybrid system, the results were significantly improved, reaching 96.90% accuracy in validation and 89.98% in tests, representing a significant improvement of approximately 10 percentage points in tests. These results highlight the potential of the hybrid system in applications, demonstrating how the combination of advanced deep learning and machine learning techniques can contribute significantly to improving accuracy.

**Keywords:** X-ray. CNN. Transfer learning. Resnet.

[1] Degree in Physics
State University of Campinas (UNICAMP)
E-mail:fabiolofredo@gmail.com
[2] Mechanical Engineer, Dr. in Solid Mechanics and Mechanical Projects
State University of Campinas (UNICAMP)
ORCID: https://orcid.org/0000-0002-4835-5498
CV Lattes: http://lattes.cnpq.br/8858059011756333
E-mail:hsantiagolara@gmail.com

## INTRODUCTION

Computer vision is a field that involves several areas designing some interpretation of human vision computationally. The field seeks to solve challenges in areas such as object or face recognition and detection. Motion detection, three-dimensional analysis from two-dimensional images, scene reconstruction and image restoration. And this brings the interest of researchers and companies (Kovaleski, 2018).

Convolutional neural networks use layered convolution operations. The filters, in these networks, contribute to extracting characteristics from the image, which makes them useful for recognizing objects and faces (Kovaleski, 2018).

One of the motivations of this work is to use x-ray images for the detection of diseases, something that can be very useful in the medical field, acting as a diagnostic tool for doctors.

## HISTORY

In the historical context of Convolutional Neural Networks (CNN) and Fully Convolutional Networks (FCN) can be divided into:

Between 1989 and 1999, the origin of Convolutional Neural Networks (CNNs) occurred. During this time, networks were able to automatically learn filter patterns and recognize rotational variations in images, marking the beginning of a new approach to image processing. In the early 2000s, however, there was a stagnation in the development of CNNs, with few relevant advances in the area. Between 2006 and 2011, CNNs underwent a renaissance thanks to the introduction of techniques such as greedy layer-wise unsupervised training and max pooling, as well as improvements in processing provided by the evolution of available hardware (Cunha, 2020).

The period between 2012 and 2014 was marked by the rise of CNNs, with significant advances in their performance, boosting their popularity and applicability in several areas. In 2014, the discovery of Fully Connected Convolutional Neural Networks (FCN) occurred. This innovation consisted of replacing the fully connected layer with a new convolution layer, allowing CNNs to perform semantic image segmentation more efficiently and accurately (Cunha, 2020).

## OBJECTIVES

There is a search for more efficient architectures, and this article aims to help show the efficiencies of two techniques for image prediction.

The objective of this paper is to create a hybrid CNN(Resnet) system with extremely random trees, using transfer learning and to measure its accuracy.

## METHODOLOGY: WORDING AND CODE

Certain parts of this article and in the code have been written with the aid of ChatGPT artificial intelligence, with the aim of improving textual clarity. In this process, we incorporate both excerpts of our own The authors. and fragments of original articles, always ensuring the proper citation of the original authors in the passages used. We subsequently submit these sections for revisions to ensure the accuracy and completeness of the content.

## DISCUSSION AND BIBLIOGRAPHIC ANALYSIS

According to the analysis of (Rodrigues, 2018), which uses a convolutional neural network to show the feasibility of the technique for reading characters from vehicle license plates, inference results in the order of 89.24% are obtained. Among the possibilities of applications are controlling road traffic, identifying cars in parking lots, or checking traffic violators.

According to (Cruz, 2019), according to the study of emotion recognition by facial expression using convolutional neural networks, there are several possible applications such as human-computer interaction, psychiatry and medical care, visual impairment, human-robot interaction and virtual characters and animation.

Second (Alvear-Sandoval et al., 2019) in which improvement techniques are applied in the application of CNN and in Stacked Denoising Auto-Encoder classifiers. And then Stacked Denoising Auto-Encoder classifiers are applied to the output of a CNN for better results. It is concluded that combining techniques of different natures can acquire better results.

In Ahlawat & Choudhary, (2020) and in Niu & Suen (2012) they carry out an approach in which a hybrid system of CNN with Support Vector Machine (SVM) is used, the CNN works by extracting features and the SVM as a binary classifier, thus obtaining an accuracy of 99.28% in the MNIST digit database in Ahlawat & Choudhary (2020) and 99.81% without rejection and 94.40% with a rejection of 5.60% in Niu & Suen (2012).

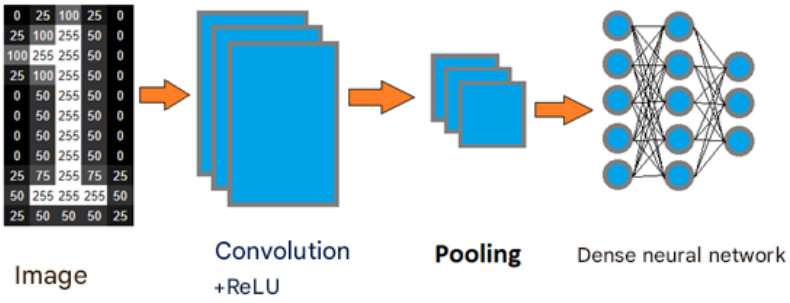
## THEORY

### CNN PROCESSES

A Convolutional Neural Network (CNN) is made up of three fundamental elements: the convolution layer, the pooling layer, and the dense neural network. The convolution layer extracts features from the input using small-size filters, which convolute the data in width, height, and depth. During training, filters adjust to identify common characteristics in the data, such as edges and colors, evolving into more complex structures. The pooling layer reduces the size of the data after convolution, allowing the network to learn different representations of the data and avoid overfitting. The dense neural network, typically at the end of the architecture, uses the extracted features to classify the output. The balance between resources and performance is essential in the definition of architecture (Rodrigues, 2018).

Figure 1 illustrates a complete process of a Convolutional Neural Network (CNN). The pixels, represented by quantitative values, initially pass through convolution layers activated by the ReLU function. Then, a pooling process is applied for dimensional reduction and extraction of more relevant characteristics. Finally, the processed information is forwarded to a dense neural network, which performs the final stage of classification or regression.

**Figure 1.** Example of a CNN process



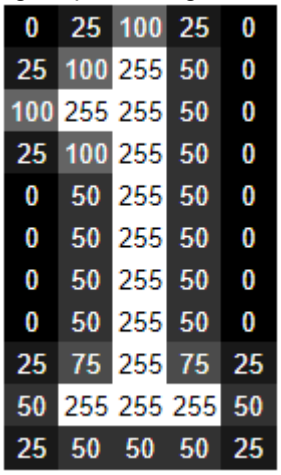

**Source:** The authors..

### IMAGE INPUT

The ability of humans to interpret images is an intrinsic ability, but computers rely on a numerical representation to process images. In the machines, the images are translated into pixel matrices, where each pixel is represented by a number ranging from 0 to 255, reflecting the color intensity. These arrays organize pixels, allowing computers to process and analyze images algorithmically. This conversion process is critical for machines to be able to understand and make decisions based on visual information, playing a crucial role in

fields such as computer vision and machine learning. Monochrome images have 1 channel, while colored with RGB have 3 channels (Cunha, 2020).

**Figure 2.** Monochrome image representing the intensity range from 0 to 255.

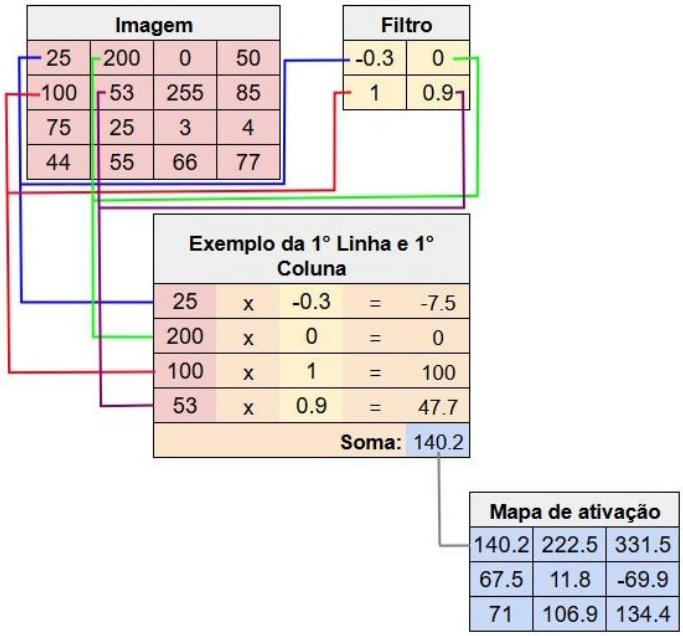



**Source:** The authors.

## CONVOLUTION

The convolution layer is the central component of a CNN, where convolution is a mathematical operation that involves sliding one function over another to process specific sub-regions of the image. This differs from the traditional approach, where neurons are connected to all the input data. The advantage of CNN lies in the efficiency of processing local features and reducing parameters, making it ideal for computer vision tasks, such as pattern recognition and image classification (Cruz, 2019), that is, a filter is slid in the image mathematically producing one or more activation maps (feature map)(Cunha, 2020).

**Figure 3.** An image being convoluted by a filter generating an activation map.
Also showing examples of the calculations in the generation of the matrix.



| Imagem | | | |
|---|---|---|---|
| 25 | 200 | 0 | 50 |
| 100 | 53 | 255 | 85 |
| 75 | 25 | 3 | 4 |
| 44 | 55 | 66 | 77 |

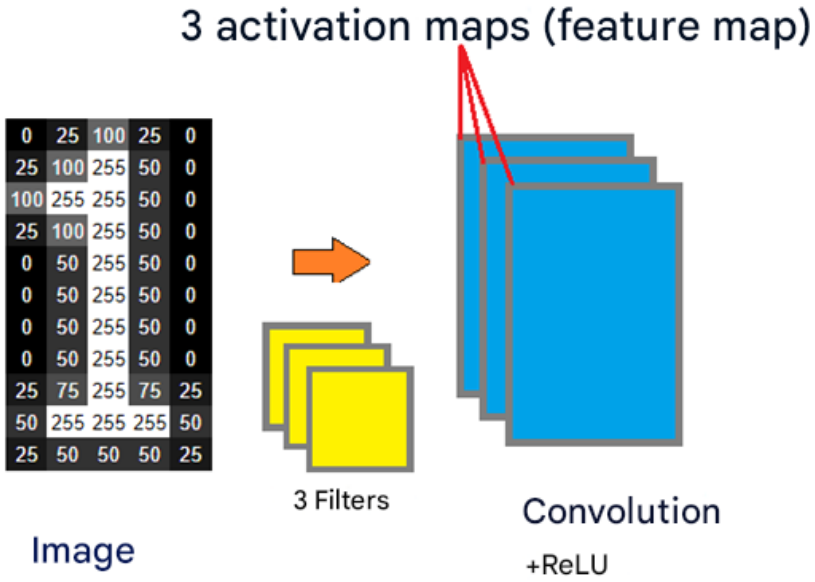
| Filtro | |
|---|---|
| -0.3 | 0 |
| 1 | 0.9 |

| Exemplo da 1° Linha e 1° Coluna | | | | |
|---|---|---|---|---|
| 25 | x | -0.3 | = | -7.5 |
| 200 | x | 0 | = | 0 |
| 100 | x | 1 | = | 100 |
| 53 | x | 0.9 | = | 47.7 |
| | | | Soma: | 140.2 |

| Mapa de ativação | | |
|---|---|---|
| 140.2 | 222.5 | 331.5 |
| 67.5 | 11.8 | -69.9 |
| 71 | 106.9 | 134.4 |

**Source:** The authors.

**Figure 4.** Example of generating 3 activation maps using 3 filters.




**Source:** The authors.

## ACTIVATION FUNCTION

After the convolution layers in a conventional neural network, it is common to insert a nonlinear layer, known as the activation layer. This layer is crucial for introducing nonlinearity into a system that, up to that point, has performed predominantly linear operations, such as multiplication and addition (Cunha, 2020).

Recent research has demonstrated that the ReLU (Rectified Linear Units) function offers superior benefits, allowing for faster and more efficient network training without

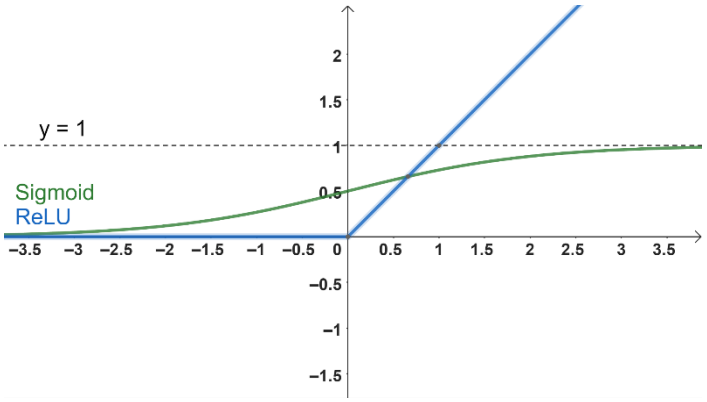
compromising accuracy. The ReLU layer applies the function f(x) = max(0, x) to all input volume values, replacing negative values with zero, increasing the nonlinear modeling capability without negatively affecting the convolution layer activations. The ReLU function is (Cunha, 2020):

$$f(x) = \max(0, x) \text{ , } (1)$$

The Sigmoid function, similar to biological neurons, restricts its values to a range between 0 (non-activation) and 1 (activation), reflecting the activation or inactivation behavior of neurons in the face of incoming inputs. The Sigmoid function is given by (Rodrigues, 2018):

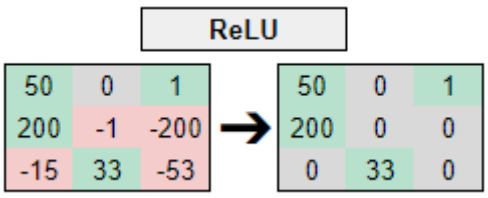
$$\sigma(x) = \frac{1}{1+e^{-x}} \text{ , } (2)$$

**Figure 5.**Graphs of the ReLU and Sigmoid activation functions.




**Source:** The authors.

**Figure 6.** Example of the ReLU function in a matrix.




**Source:** The authors.

They usually use the Softmax function at the end of a CNN to classify, it is a generalized model of a logistic regression and can estimate probabilities for multiple classes (Cruz, 2019). The probability for a class i, given an input x is obtained in the formula:

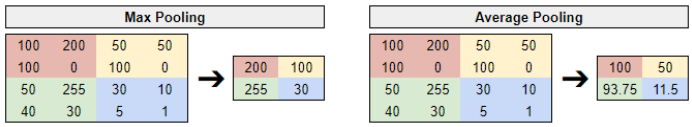
$$p(classe = i|x) = \frac{e^{y_i}}{\sum_{j=1}^{K} e^{y_j}}, (3)$$

Where K is the total number of classes (Kovaleski, 2018).

## POOLING

The pooling layer plays a crucial role in Convolutional Neural Networks (CNNs), focusing on subsampling to reduce image size during processing. This helps to reduce the computational load, the number of parameters and, consequently, to prevent overfitting and save memory. The pooling operation is inspired by the workings of the visual cortex, where local reception fields represent sub-regions of the visual field of specific neurons. Reducing the image size makes the CNN more robust to variations in the position of objects and eliminates minor noise. There are two main types of pooling: max pooling, which selects the maximum value in a receive field, and average pooling, which averages the values close together. Max pooling is the most commonly used technique in CNNs (Cruz, 2019).

**Figure 7.** Examples of Pooling to reduce image size. In Max Pooling you keep the highest number and in Average Pooling you average the values.

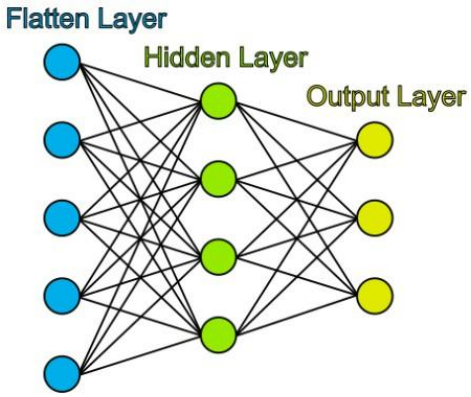



**Source:** The authors..

## DENSE NEURAL NETWORK

The fully connected layer in a Convolutional Neural Network (CNN) plays an essential role in the classification decision of the image as a whole. Its function is to unite all the characteristics and attributes extracted from the image by the previous layers, i.e., the convolution and pooling layers. This layer interconnects all neurons in a traditional way, with the output of the previous layer being flattened into a single vector before entering the fully connected layer (Cunha, 2020).

**Figure 8.** Dense neural network.




**Source:** The authors.

## TRAINING

The training involved dividing the database into batches for processing and updating CNN weights after each batch. After running all batches, a new epoch (epoch) is calculated. The ideal model was defined based on the lowest loss function rate (Loss Function or loss) on the basis of validation. The Learning Rate is dynamically adjusted by the optimizer, starting high and decreasing throughout the training, working as an accelerator, thus obtaining a satisfactory accuracy (Cruz, 2019).

## TRANSFER LEARNING

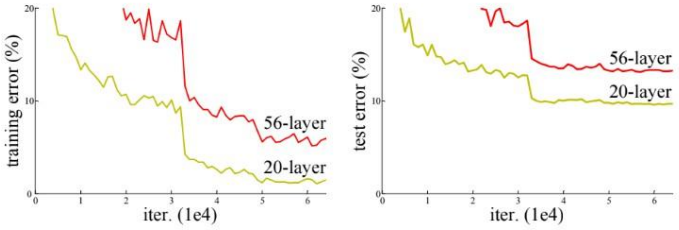
Transfer learning is based on transferring learning, already learned, to learning something in some different, but similar domain. In Zhuang et al. (2021) cite the intuitive examples of transfer learning, such as those who learn the violin can learn the piano faster, just as those who learn to ride a bicycle can learn to ride a motorcycle more quickly. But care must be taken, because if the domains have little in common, learning will not be efficient, because those who learn to ride a bicycle may not help learn to play the piano.

Thus, in the practical context of artificial intelligence and this work, something already trained is used to learn in another similar domain, thus taking less training time and needing less data. The already trained model used in this work will be Resnet, layers will be added that will be trained with X-ray images, optimizing time and requiring fewer images.

## RESNET

Resnet is a residual neural network trained, which won first place in the 2015 ILSVRC competition. The network has an output for 1000 classes, having been trained with 1.28 million images (He et al., 2015).

**Figure 9.** Training error on the left and test error on the right in CIFAR-10 with 20 and 56 layers. The deeper network has higher training and testing error.




**Source:** He et al., 2015.

According to the article in figure 9, neural networks with many layers will not necessarily give a better result in accuracy.

**Figure 10.** An example of a residual learning block in Resnet.

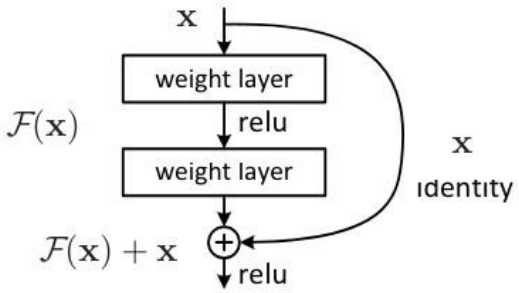



Figure 2. Residual learning: a building block.

**Source:** He et al., 2015.

In figure 10 above, the example of the operation of a part of the Resnet network, in which it has "shortcut connections", they skip one or more layers. Shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked layers.

3.9 EXTREMELY RANDOM TREES

Extremely random trees, or Extra Tree, is a Machine Learning algorithm for prediction. With it, it is possible to train the algorithm with data following a logic similar to decision trees. According to Geurts et al. (2006), the main difference in relation to other tree algorithms, translating into Portuguese, is:

> The Extra-Trees algorithm constructs a set of unpruned decision or regression trees according to the classic top-down procedure. Its two main differences with other tree-based set methods are that it splits nodes by choosing entirely random cutoff points, and that it uses the entire learning sample (rather than a bootstrap replica) to grow the trees.

## METHODOLOGY

### FIRST STAGE

First, the database of chest x-ray images from the site was used, in which we took the training and test images with and without pneumonia. We also adopted the architecture

of resnet_v2. We added a layer of 64 neurons with Relu activation function and 20% dropout, another layer of 16 neurons also with Relu and 20% dropout and at the end we added 2 neurons with Softmax activation function for the classification of healthy or sick. As in Table 1:

**Table 1.** Architecture being the keras_layer the resnet_v2. Additionally, a layer of 64, another of 16 and the last of 2 neurons.

| Layer | Name | Neuron outputs |
|---|---|---|
| 1 | Resnet | 1001 |
| 2 | Densa64 + Dropout | 64 |
| 3 | Densa16 + Dropout | 16 |
| 4 | Dense2 | 2 |

**Source:** The authors.

The architecture was trained using transfer learning, i.e., freezing the resnet and training the following neurons.

## SECOND STAGE

In the second stage, the layers of 2 neurons and the dropout of the layer of 16 neurons were removed, as shown in Table 2. After that, the architecture of extremely random trees was trained on the result of the layer of 16 neurons in a supervised manner. The result of healthy or sick was obtained in the output of the model of extremely random trees.

**Table 2.** Architecture being the keras_layer the resnet_v2. Additionally a layer of 64, another of 16.

| Layer | Name | Neuron outputs |
|---|---|---|
| 1 | Resnet | 1001 |
| 2 | Densa64 + Dropout | 64 |
| 3 | Dense16 | 16 |

**Source:** The authors.

## OUTCOME AND DISCUSSIONS

In the first stage, the resnet with 2 neuron outputs, an accuracy of 95.4% was obtained for validation and for the test of 79.33%, clearly showing overfitting.

In the second stage, the resnet, with the neural network plus the extremely random tree, an accuracy of 96.90% for validation and 89.98% for the test was obtained, improving the test results.

This shows that there was an increase of approximately 10 percentage points for the test, significantly improving the accuracy of the model.

The efficiency of the result obtained was similar to that of the technique of reading characters on vehicle license plates (Rodrigues, 2018), which reached 89.24%. However, it

is important to highlight the difference between the realities of these two cases: while one is about digits on license plates, the other involves chest x-ray analysis.

Leão et al. (2020), who investigated the application of convolutional networks for the detection of Covid-19 in X-ray images, in their analysis of with two and three classes and with and without transfer learning, obtained accuracies between 82.11% and 87.91%. Showing that the second stage, carried out in this work, obtained superior results.

In Castro et al. (2023), who explored the use of Principal Curves as a technique to optimize the screening of tuberculosis patients, also using X-rays, they obtained accuracies between 84% and 89%. Showing similar results obtained in this work.

## CONCLUSION

The transfer learning process in the resnet with the extremely random trees performed better than with transfer learning alone for the chest x-ray images, with approximately 10 percentage points difference.

The final accuracy results of this study are comparable to those in the literature, but the improvement in accuracy from the first to the second models shows a potential tool to be used in conjunction with other models.

For future work it is possible to think about applying this system in other data and contexts, it is also possible to try new architectures based on the hybrid system, with the potential to increase the final accuracy. It is also possible to use "Finning tuning", which consists of also training the Resnet network, to try to obtain better results.

## ACKNOWLEDGMENTS

I thank Professor Hygor S. Lara for all his support and dedication to the project, without which it would not be possible to carry out this work. I also thank the staff of the GEIA (Study Group on Artificial Intelligence).

## REFERENCES


1. Ahlawat, S., & Choudhary, A. (2020). Hybrid CNN-SVM classifier for handwritten digit recognition. Procedia Computer Science, 167, 2554-2560. https://doi.org/10.1016/j.procs.2020.03.387

2. Alvear-Sandoval, R. F., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2019). On improving CNNs performance: The case of MNIST. Information Fusion, 52, 106-109. https://doi.org/10.1016/j.inffus.2019.04.009

3. Castro, D. H. H. de, et al. (2023). Utilização de curvas principais na triagem de pacientes com tuberculose. In XVI Brazilian Conference on Computational Intelligence (CBIC 2023), Salvador, BA, 8-11 de outubro, 2023.

4. Cruz, A. A. (2019). Uma abordagem para reconhecimento de emoção por expressão facial baseada em redes neurais de convolução (Dissertação de Mestrado em Engenharia Elétrica). Universidade Federal do Amazonas, Manaus, Brasil.

5. Cunha, L. C. (2020). Redes neurais convolucionais e segmentação de imagens - Uma revisão bibliográfica (Trabalho de Conclusão de Curso em Engenharia de Controle e Automação). Universidade Federal de Ouro Preto, Ouro Preto, Brasil.

6. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63, 3-42. https://doi.org/10.1007/s10994-006-6226-1

7. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385. https://arxiv.org/abs/1512.03385

8. Kovaleski, P. A. (2018). Implementação de redes neurais profundas para reconhecimento de ações em vídeo (Trabalho de Conclusão de Curso em Engenharia de Computação). Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

9. Leão, P. P. de S., et al. (2020). Detecção de Covid-19 em imagens de raio-X utilizando redes convolucionais. J. Health Inform., Número Especial SBIS, 393-398.

10. Niu, X.-X., & Suen, C. Y. (2012). A novel hybrid CNN–SVM classifier for recognizing handwritten digits. Pattern Recognition, 45(4), 1318-1325. https://doi.org/10.1016/j.patcog.2011.11.019

11. Rodrigues, D. A. (2018). Deep learning e redes neurais convolucionais: Reconhecimento automático de caracteres em placas de licenciamento automotivo (Trabalho de Conclusão de Curso em Ciência da Computação). Universidade Federal da Paraíba, João Pessoa, Brasil.

12. Zhuang, F., et al. (2021). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555