Classification of chest X-ray images using Machine Learning and Histogram of Oriented Gradients

Fellipe M. C. Barbosa[1], **Anne Magaly de P. Canuto**[1]

[1]Departamento de Informática e Matemática Aplicada
Universidade Federal do Rio Grande do Norte (UFRN)
Natal – Rio Grande do Norte – Brasil

fellipecosta@ppgsc.ufrn.br, anne@dimap.ufrn.br

Abstract. *This work proposes a machine learning model trained from scratch to classify and detect the presence of pneumonia from a collection of chest X-ray images. Unlike most works that use deep learning to classify whether the image is of a lung with pneumonia or not, that is, two classes to achieve a remarkable classification performance, this model uses Oriented Gradient Histogram for extra features from a provided chest X-ray image and classify it into three classes, determining whether or not a person is infected with viral or bacterial pneumonia. Despite greater complexity and use of traditional machine learning techniques, the highest accuracy achieved was 91.32% more higher than works that use deep learning approaches and seeking to solve the same degree of complexity.*

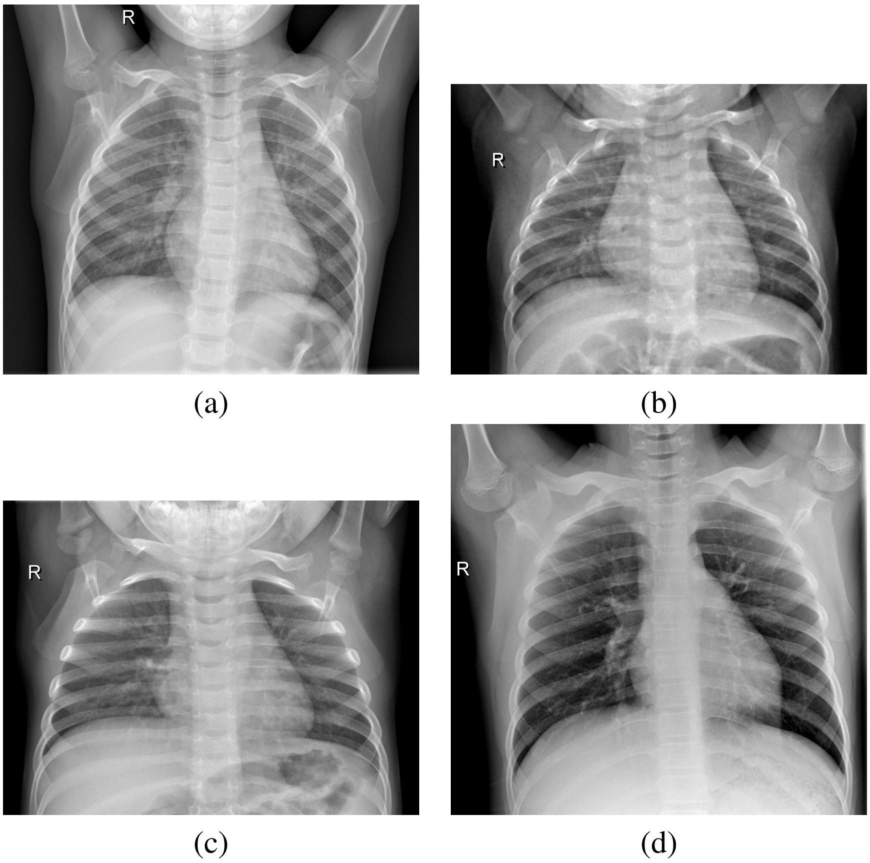
Resumo. *Este trabalho propõe um modelo de aprendizado de maquina para classificar e detectar a presença de pneumonia a partir de uma coleção de amostras de radiografias do tórax. Ao contrário da maioria dos trabalhos que utilizam abordagens de aprendizado profundo para classificar se a imagem é de um pulmão com pneumonia ou não, ou seja, duas classes para assim alcançar um desempenho de classificação notável, este modelo utiliza Histograma de Gradientes Orientados para extrair características de uma determinada imagem de raio-X de tórax e classificá-la em três classes, determinando se uma pessoa está ou não infectada com pneumonia viral ou bacteriana. Apesar de uma maior complexidade e utilização de modelos tradicionais de aprendizado de maquina, a maior acurácia alcançada foi de 91.32% superior a de trabalhos que utilizam redes profundas e buscam resolver o mesmo grau de complexidade.*

1. Introduction

Pneumonia is a common disease that remains the leading cause of death for children in developing countries and elderly people in developed countries. WHO estimates that more than 4 million premature deaths occur annually from diseases related to air pollution, including pneumonia [World Health Organization 2018]. More than 150 million people are infected with pneumonia annually, especially children under the age of 5 [Rudan et al. 2004]. In poor regions, the problem may be further aggravated by the scarcity of medical resources and specialists. For example, in the 57 nations of Africa, there is a gap of 2.3 million doctors and nurses [Narasimhan et al. 2004], [Naicker et al. 2009]. For these countries, accurate and rapid diagnosis it is very important to save time and money.

Neural network models and experiments are conventionally designed and performed by experts in a continuous trial and error method. This process requires a lot of time, knowledge and resources as most of them use deep learning techniques and digital image processing. To overcome this problem, a series of new and simpler models capable of automatically classifying X-ray images were analyzed.

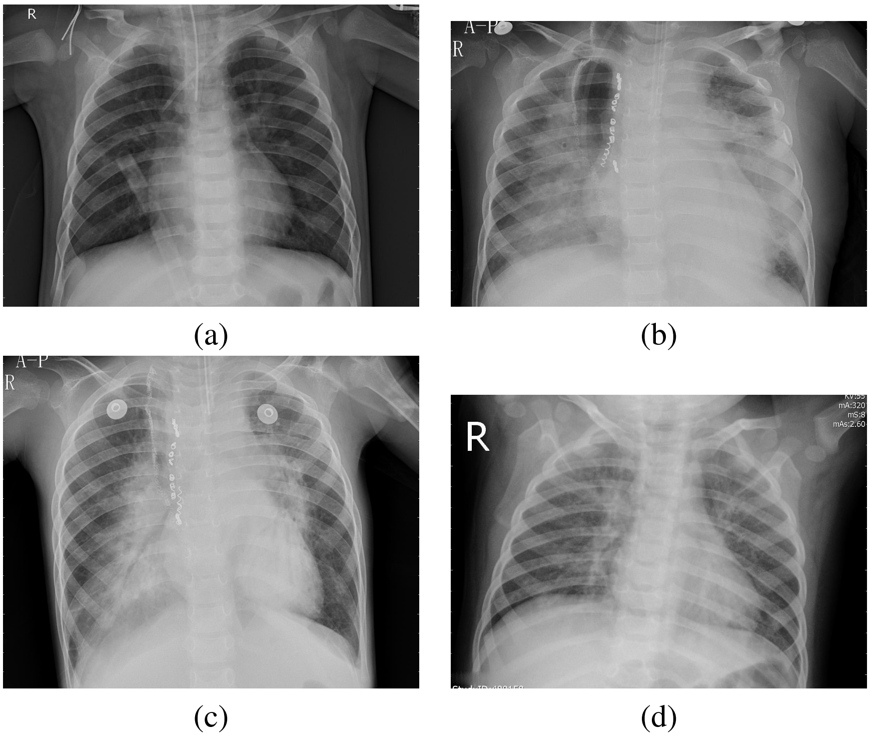
The proposed technique is based on machine learning algorithms, using Oriented Gradient Histogram to extract relevant features from X-ray images. In order to demonstrate the effectiveness of the proposed method with the minimization of computational cost as the main point, the best results were compared with the next-generation dense pneumonia classification networks. In recent times, deep learning algorithms based on Convolutional Neural Networks (CNN) [Shen et al. 2017] have become the standard choice for medical image classifications, although such techniques are computationally expensive, swallowing a ton of processing power. As an alternative, our study proposes to solve a more complex problem than current studies that seek to classify whether the image is pneumonia or not (2 classes). It was proposed to classify the image into 3 classes, separating pneumonia into 2 groups, viral and bacterial. Something extremely relevant for medical decision making when choosing which drug will be used in the treatment. A conceptually simple but efficient model for dealing with the pneumonia classification problem shown in Figures 1, 2 and 3.



(a) (b)

(c) (d)

Figure 1. X-ray images without pneumonia.



(a) (b) (c) (d)

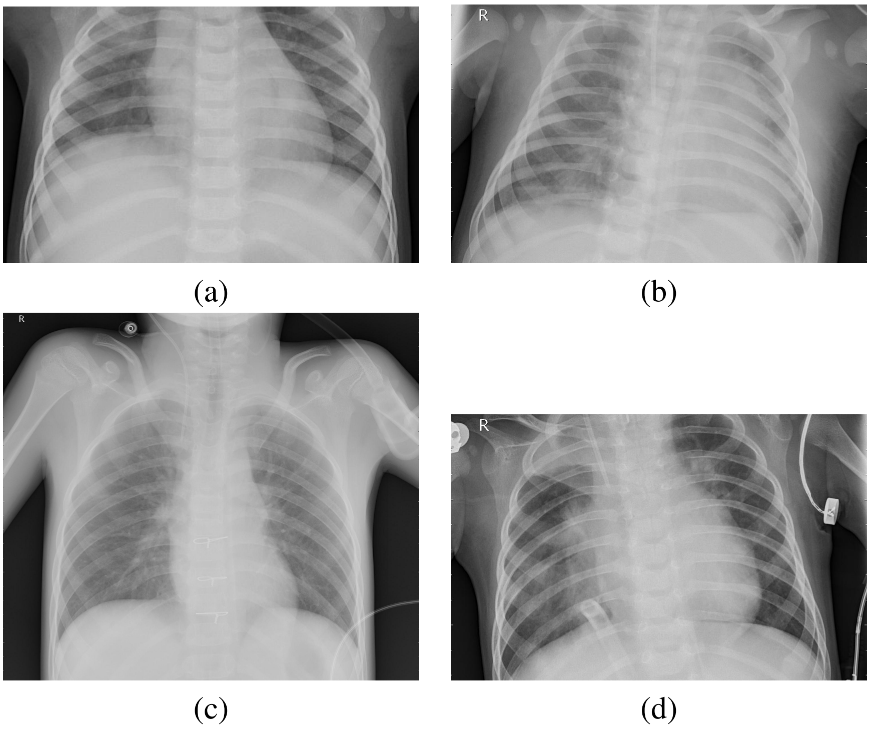
Figure 2. X-ray images with viral pneumonia.

2. Related Work

Recently, the use of algorithms capable of classifying chest x-ray images has been widely applied in the literature. These algorithms are increasingly being used for the detection of pulmonary nodules [Huang et al. 2018] and classification of pulmonary tuberculosis [Lakhani and Sundaram 2017] as well as their use for the detection of COVID-19 [Ali et al. 2021]. The performance of multiple convolutional models on various human health abnormalities using X-ray images with the publicly available dataset [Demner-Fushman et al. 2016], it was found that the same deep convolutional network architecture does not perform well on all abnormalities [Islam et al. 2017].

Convolutional Neural Networks (CNNs) have an advantage over deep neural network (DNNs) in that they have a human-equivalent visual processing scheme and an extremely optimized structure for handling 2D and 3D images and Shapes, as well as the ability to extract abstract 2D features through learning. Most importantly, gradient-based learning algorithms are employed in training CNNs and are less prone to decreasing gradient problems. Since the gradient-based algorithm is responsible for training the entire network to directly lower the error criterion, highly optimized weights can be produced by CNNs. Knowing this, it is common to observe the use of gradient-based algorithms to classify X-ray images [Hamadi 2021], [Ali et al. 2021], [Sirazitdinov et al. 2019].

Among the studies that seek to classify pneumonia using X-ray images, the vast majority are concerned with classifying the problem into two classes. The work of Gu and Pan [Gu et al. 2018] sought to classify X-ray images of children into three classes (Normal, Viral Pneumonia and Bacterial Pneumonia). For this, a CNN was used to extract characteristics from the images and classification. This work obtained an accuracy of



(a) (b)

(c) (d)

Figure 3. X-ray images with bacterial pneumonia.

0.8048 ± 0.0202.

3. Histogram of Oriented Gradients

Objects in images have shape, color and texture. Such features can usually be grouped into scalar arrays, called image descriptors. In this sense, if we consider that each object is represented by a point in a space, where each point has a total of n characteristics, it is desirable that the array that describes this point is invariant to transformations (rotation, scale, luminosity). The Oriented Gradients Histogram (HOG) algorithm [Dalal and Triggs 2005] is a descriptor that calculates the histogram of the gradients orientation in the image. The final descriptor is an array of histograms extracted from the image. The algorithm is based on the idea that the shape and appearance of an object can often be described by the intensity of the gradients or the direction of the edges, without prior knowledge of the position of such edges. The method initially consists of normalizing the image according to the lighting. Next, the gradients are computed.

Next, the image is divided into small spatial regions called cells. For each cell, a local histogram of the cell pixel orientations is calculated. Cells can be rectangular or radial in shape. After the histograms are computed, they are normalized. Normalization is done by an accumulation of local histograms in slightly larger spatial regions called blocks. The accumulated histograms are then used to normalize all cells in that block. After normalization, a detection window is collected on the generated histograms, which consists of the output of the HOG descriptor. The HOG descriptor uses the array of gradient orientations to describe the shape of the analyzed image. Such description is invariant to the position of the gradient

4. Materials and Methods

All experiments were based on a chest X-ray image dataset proposed in [Stephen et al. 2019]. We used the OpenCV library [Bradski 2000] to extract the characteristics of the images and the Weka framework [Hall et al. 2009] to execute the classification algorithms. All experiments were run on a desktop with 16GB of ram and Intel Xeon processor e5 2620 v3.

4.1. Database

The original dataset [Jaeger et al. 2013] consists of three main folders (Training, Testing, and Validation) and two sub folders containing pneumonia and normal chest X-ray images. Altogether the database has a total of 5856 X-ray images of the anterior-posterior chest chosen from retrospective pediatric patients between 1 and 5 years old.

All data has been reorganized into a one set only. The pneumonia image group was separated into two, viral and bacterial pneumonia, this information was contained in the image metadata file. Our dataset has a total of 5216 images, where 1341 images were normal, 2530 bacterial pneumonia and 1345 viral pneumonia.

In order to train the machine learning model, the used dataset represents the radiographic images where each instance represents an image, each attribute represents the response of the HOG descriptor and the class labeled according to the metadata of the original dataset. As a way to reduce the resulting HOG's array each image was resized to 512×512 . The cell size used was 8×8. Thus, the generated base, which is called Original Base, now has 5216 Instances (total number of images) and 1765 attributes (size of the generated HOG descriptor response for each resized image).

4.2. Pre-processing the original database

Three other datasets were created in order to reduce and investigate the behavior of the analysed classifiers, since the original base had too many instances and attributes. The first reduction (Reduced Base 1) occurred using the Resample technique, thus balancing the number of classes since the number of instances with the bacterial pneumonia class was greater. This process reduced the original base by 1046 instances. The second reduction (Reduced Base 2) was using a feature selection method, Correlation-based Feature Selection (CFS) [Hall 1999] with the best-first-search heuristic. This process reduced the number of attributes to 123. The third reduction (Reduced Base 3) in the original database was the use of Principal Component Analysis (PCA) [Wold et al. 1987]. This process extracted 572 attributes from the original base. Table 1 shows number of instances and attributes for each base.

Dataset	Number of Instances	Number of Attributes
Original Base	5216	1765
Reduced Base 1	4170	1765
Reduced Base 2	5216	123
Reduced Base 3	5216	572

Table 1. Number of attributes and instances for each dataset.

5. Results

In order to evaluate and to validate the effectiveness of this proposal, two groups of classifiers were defined, a group containing supervised learning classifiers and a group with ensemble systems. Several results were obtained, however, for simplicity reasons, this study reports only the most important ones. All results shown are mean accuracy of 3×10-*fold cross-validation*.

5.1. Supervised Learning

From the experiments performed with supervised learning techniques, 4 algorithms were used: k-NN, Naive Bayes, Decision Tree (J48) and the neural network Multilayer Perceptron (MLP) [Faceli et al. 2021]. The experiment with k-NN used a k ranging from 1, 3, 5 and 10. The best results were with k equal to 1. The decision tree was executed with and without pruning, the bests results were with pruning. For the experiment with neural networks, several combined parameters were used. The amount of network interactions at 100, 500 and 1000. The learning rate ranged between 0.1, 0.01 and 0.001. In addition number of neurons also varied in number of attributes+classes/2, number of attributes+classes and number of classes. Table 2 shows the mean accuracy for each classifier.

Dataset	k-NN	J48	Naive Bayes	MLP
Original Base	66.10	55.83	57.51	67.33
Reduced Base 1	80.44	76.16	55.15	**80.93**
Reduced Base 2	65.64	56.19	60.16	63.21
Reduced Base 3	34.85	50.36	53.31	56.25

Table 2. Performance of classifiers for all datasets.

In the k-NN method to Original Base it was possible to observe that the higher the value of k, more accurate the algorithm was. The Reduced Base 1 obtained the best results for this experiment. The value of K as it increased, decreasing the accuracy of the algorithm, with the best value of K being equal to 1. In one of the 3×10-*fold cross-validation* executions, the algorithm had an accuracy of 83.21%. The Reduced Base 2, as the original, it was observed that the higher the value of K, the more accurate the algorithm was. Reduced Base 3 was imprecise in all runs.

In the Decision Tree method, the Reduced Base 3 it was not bad in relation to the other bases as k-NN. Furthermore, it was possible to observe that the reduction of bases was beneficial (disregarding the Reduced Base 3).

In the Naive Bayes classifier the Reduced Base 2 obtained the best results due to its smaller amount of attributes. However the MLP and the others classifiers the order of the best accuracy were to Reduced Base 1, Original Base, Reduced Base 2 and finally the Reduced Base 3.

5.2. Classifier Ensembles

Still aiming to investigate the behavior of traditional machine learning methods, experiments with classifier ensembles were also performed. The AdaBoost, Bagging and Stacking techniques were performed using the same classifiers presented in Subsection 5.1.

The AdaBoost and Bagging members varied between 10, 15 and 20. Nevertheless, no significant performance variation was detected in the results. Among the obtained results, the ranking of classifiers for AdaBoost and Bagging was as follows: the highest average accuracy was obtained by MLP, followed by k-NN, J48 and Naive Bayes.

Stacking execution was divided into homogeneous and heterogeneous. For the homogeneous one, the number of classifiers varied between 10, 15 and 20. The heterogeneous stacking was divided in 4 groups all composed of 20 classifiers in total. The first group was composed by k-NN and J48 classifiers, the second group was composed by J48 and Naive Bayes, the third group was composed by Naive Bayes and k-NN and the last group was composed by k-NN, J48 and Naive Bayes.

Based on the obtained results, we can observe that the homogeneous Stacking obtained better results than the heterogeneous one. However, among the heterogeneous Stacking, k-NN combined with J48 was better. Table 3 shows the best average accuracy of each ensemble for the different analysed datasets.

Dataset	**AdaBoost**	**Bagging**	**Stacking**
Original Base	70.21	70.11	62.50
Reduced Base 1	89.52	91.32	75.77
Reduced Base 2	64.11	66.37	60.62
Reduced Base 3	59.03	62.77	50.59

Table 3. Performance of the classifier ensembles for all datasets.

Analyzing the results of AdaBoost, Bagging and Homogeneous Stacking , there was no significant improvement in the algorithm for the amount of classifiers used, however the highest number (20) was more accurate.

Analyzing the individual results for each dataset, the ratings of the classifier ensembles were better than the individual classifiers. The algorithm that achieved a significant improvement was the Decision Tree, while Naive Bayes had a gets worse. Considering the mean accuracy, the ranking of classifier ensembles was Bagging in first position followed by AdaBoost, Homogeneous Stacking and finally Heterogeneous Stacking. For all bases Heterogeneous stacking with k-NN, J48 and MLP, was more satisfactory than the other groups.

In the Original Base use Naive Bayes with Heterogeneous Stacking was not very beneficial. For the Reduced Base 1 the J48 obtained a very considerable increase. Just like in the Original Base, the Naive Bayes got worse. The results for Reduced Base 2 were not as expressive as in the previous bases. However, if we take into account the individual classifiers for this base, the classifier ensembles obtained a significant improvement for all the methods, none of them decreased. For the results of Reduced Base 3 was possible observe some particularities. The k-NN decreased significantly compared to the individual, while the decision tree and Naive Bayes was improved.

6. An Analysis of the Obtained Results

With the results seen in the previous section, it is possible to reach the conclusion that the dataset reduction had a positive effect in the classification process, specially the balancing of classes with Resample (reduced base 1). The use of the PCA, on the other hand, did not bring any benefits to the classification process, obtaining unsatisfactory results when compared to the other datasets.

In supervised methods k-NN with k equal to 1 and MLP with learning rate in 0.001, number of neurons equal to number of attributes and number of interactions at 1000 had the best results. Analyzing the methods of classifier ensembles, Bagging and AdaBoost had better results than the Stacking methods. The classifiers used in the classifier ensemble methods, k-NN and MLP had better accuracy.

In order to analyse the statistical validity of the obtained results, the non-parametric Friedman test was performed, which revealed significant differences between all binding criteria ($p - value < 0.001$), with a significance level of $\alpha = 0.05$. As a result, a p-value of 0.0010317868098302344 was obtained, which means that the results are significantly different. Next, the post-hoc Nemenyi test was performed. Figure 4 illustrates the Critical Difference (CD) diagram of the results of the Nemenyi test. As it can be seen in this figure, Bagging with MLP was statistically better than all other analysed classification models.

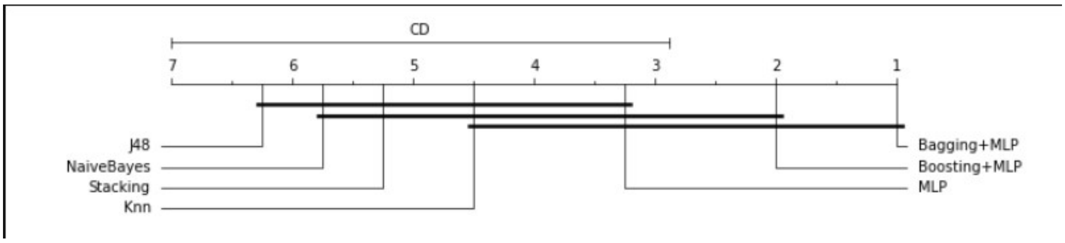


Figure 4. Nemenyi test of the bests classifiers.

The obtained results and the analysis of the statistical tests allow us to state that the Bagging method with MLP was statistically superior to the others. Furthermore, it was also possible to observe that the K-NN classifier stood out considerably in relation to J48, Naive Bayes and Stacking.

When comparing the results obtained in this work with the results of similar studies such as the work of Gu and Pan [Gu et al. 2018] which obtained an accuracy of 80.42% to classify X-ray images into three classes (Normal, Viral Pneumonia and Bacterial Pneumonia), using deep networks. Additionally, in [Stephen et al. 2019], the authors obtained an average accuracy of 94.18% to classify X-ray images into two classes (Normal, Pneumonia) using deep networks. It is possible to notice that while the cited studies had a great effort to idealize a deep learning model, as well as the increase of the data set (*data augmentation*), the present work had its best accuracy of 91.32%. Therefore, we can observe that the obtained results were superior to the ones obtained by Gu's [Gu et al. 2018], which has the same complexity, and a result close to the ones obtained by [Stephen et al. 2019]. This indicates that there is a great influence of the used classifiers and also the set of features extracted with the HOG descriptor.

With the results obtained, it is believed that a fine adjustment in the extraction of features using the HOG descriptor and the use of classifier ensembles, as well as the use of neural networks and deep networks can bring even better results for the classification of pulmonary X-ray images .

7. Conclusion

This paper presented an investigation of machine learning models to the classification of pneumonia using x-ray images. In the proposed methodology, four classification algorithms and three ensemble structures were assessed in an empirical analysis. Additionally, three different dataset reductions were also assessed. Through this analysis, we can observe that the use of MLP achieved the highest accuracy levels, among all analysed classification algorithms. Among the classifier ensembles, the AdaBoost method obtained the highest overall accuracy level. Finally, a comparison with two existing studies showed that the obtained results were very competitive, which is promising.

References

Ali, S. S. M., Alsaeedi, A. H., Al-Shammary, D., Alsaeedi, H. H., and Abid, H. W. (2021). Efficient intelligent system for diagnosis pneumonia (sars-covid19) in x-ray images empowered with initial clustering. *Indonesian Journal of Electrical Engineering and Computer Science*, 22(1):241–251.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Faceli, K., Lorena, A., Almeida, T., de Carvalho, A., and Gama, J. (2021). *Inteligência Artificial: uma abordagem de Aprendizado de Máquina (2a edição)*. LTC.

Gu, X., Pan, L., Liang, H., and Yang, R. (2018). Classification of bacterial and viral childhood pneumonia using deep learning in chest radiography. In *Proceedings of the 3rd International Conference on Multimedia and Image Processing*, ICMIP 2018, page 88–93, New York, NY, USA. Association for Computing Machinery.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hall, M. A. (1999). Correlation-based feature selection for machine learning. *University of Waikato Hamilton*.

Hamadi, A. B. (2021). Interactive automation of covid-19 classification through x-ray images using machine learning. *Journal of Independent Studies and Research Computing*, 18(2).

Huang, P., Park, S., Yan, R., Lee, J., Chu, L. C., Lin, C. T., Hussien, A., Rathmell, J., Thomas, B., Chen, C., et al. (2018). Added value of computer-aided ct image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study. *Radiology*, 286(1):286–295.

Islam, M. T., Aowal, M. A., Minhaz, A. T., and Ashraf, K. (2017). Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*.

Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R. K., Antani, S., et al. (2013). Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245.

Lakhani, P. and Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582.

Naicker, S., Plange-Rhule, J., Tutt, R. C., and Eastwood, J. B. (2009). Shortage of healthcare workers in developing countries–africa. *Ethnicity & disease*, 19(1):60.

Narasimhan, V., Brown, H., Pablos-Mendez, A., Adams, O., Dussault, G., Elzinga, G., Nordstrom, A., Habte, D., Jacobs, M., Solimano, G., et al. (2004). Responding to the global human resources crisis. *The Lancet*, 363(9419):1469–1472.

Rudan, I., Tomaskovic, L., Boschi-Pinto, C., and Campbell, H. (2004). Global estimate of the incidence of clinical pneumonia among children under five years of age. *Bulletin of the World Health Organization*, 82:895–903.

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248.

Sirazitdinov, I., Kholiavchenko, M., Mustafaev, T., Yixuan, Y., Kuleev, R., and Ibragimov, B. (2019). Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database. *Computers & electrical engineering*, 78:388–399.

Stephen, O., Sain, M., Maduh, U. J., and Jeong, D.-U. (2019). An efficient deep learning approach to pneumonia classification in healthcare. *Journal of healthcare engineering*, 2019.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

World Health Organization (2018). Household air pollution and health. *Disponível em: `https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health`*.