Convolutional neural networks with approximation of Shapley values for the classification and interpretation of pneumonia in X-ray images

Arthur Gabriel Mathias Marques[1], **Alexei Manso Correa Machado**[1,2]

[1]Departamento de Ciência da Computação - PUC Minas
[2]Departamento de Anatomia e Imagem - Universidade Federal de Minas Gerais

arthurgmm516@hotmail.com, alexeimcmachado@gmail.com

Abstract. *Pneumonia is a lung disease responsible for the highest number of deaths from infection in children and adults. Its diagnosis must be fast and accurate so that procedures are taken as soon as possible to combat the disease. In this work, Convolutional Neural Networks were explored for the classification of chest radiography images in the context of pneumonia diagnosis. Although these models are highly effective, their predictions are difficult to interpret. Therefore, the proposed method additionally aims at presenting an explainable model based on Shapley approximation values to perform the diagnosis of pneumonia with higher robustness. Results show that the model achieves competitive accuracy when compared to other architectures, and overcome them with respect to interpretation abilities.*

1. Introduction

Pneumonia is an infection that is among the most commonly diagnosed lung diseases. It can be fatal, and is still the main cause of death for children up to 5 years of age. According to UNICEF [2022], more than 2000 children die every day in the world. Infection mortality before Covid pandemics was estimated in 2.5 million deaths and the fight against the disease continues to require important attention.

Pneumonia diagnosis can be made by tests such as chest X-ray, computer tomography (CT), magnetic resonance (MRI) and ultrasound. Currently, the most common screening method is X-ray, as it is a less expensive technique that can be performed in seconds, despite producing a simpler visualization compared to other exams. The interpretation of chest radiography is however time consuming, so that it has become a potential target for machine learning techniques. Convolutional Neural Networks (CNNs) have proved to be effective in several medical applications with image classification despite providing little understanding of how decision process is carried out. It is worth mentioning the importance of data interpretability, especially in the medical field, so that health professionals feel confident to incorporate diagnosis through deep learning models as an ally in their work. Interpretability also allows the discovery of other important information in the image data that could otherwise go unnoticed.

The objective of this work is to propose a CNN architecture to accurately classify pneumonia from chest X-ray images, along with the application of an interpretability technique based on the approximation of *SHapley Additive exPlanations* (SHAP) values to provide visual explanations for the models' decisions. Based on comparisons between the classification metrics of the networks and their interpretation results, the best interpretable model can be chosen to perform more reliable pneumonia diagnoses.

The work is structured in five sections. Section 2 presents a literature review with an analysis of background work. Section 3 presents the theoretical framework for carrying out the proposed methodology. Section 4 presents the method with experiments described in section 5. Finally, section 6 presents the conclusion of this article and the trends for future work.

2. Related works

In order to propose a method for the interpretability of neural networks, Shrikumar et al. [2017] present the *Deep Learning Important Features* (DeepLIFT). In the approach, the activation of each neuron is compared with a reference activation and a contribution score is generated. According to the authors, the use of the difference of references guarantees the propagation of the information regarding the importance of the neurons, thus avoiding misinterpretations. This approach is further improved by Lundberg and Lee [2017] that create a unified framework for the interpretation of predictions. The work implements an integration of interpretability methods using the approximation of SHAP values, that assigns to each feature an importance value for a specific prediction. One of the integrated methods is based on DeepLIFT, where the authors provide a new value approximation approach resulting on the *Deep SHAP* algorithm.

In view of the impact caused by the COVID-19 pandemics, Ravi et al. [2020] proposed a visual data clustering structure (ViDi) to assist radiologists in the diagnosis of COVID-19. The authors employed the *Deep SHAP* algorithm to generate clusters that were described as interpretable saliency maps and managed to achieve a grouping homogeneity of 80% in a dataset. Also in the context of the COVID-19 pandemics, Panwar et al. [2020] proposed a deep transfer learning algorithm that accelerates the detection of chest X-rays and CT scans, together with the use of a visualization approach based on Gradient-weighted Class Activation Mapping (Grad-CAM) for interpreting the classifications performed by the model. The authors concluded that for the accurate detection of cases of COVID-19 it is also necessary to train the model for the detection of pneumonia in order to avoid false negatives. Using the same approach of activation maps for the diagnosis of gliomas in magnetic resonance images, Pereira et al. [2018] proposed a Convolutional Neural Network model for brain tumor classification. The approach has two main steps: the extraction of regions of interest and the prediction of the degree of glioma. The Guided Backpropagation and Grad-CAM methods were used for the interpretation of the predicted results. The method achieved an accuracy of 0.895 in the complete brain analysis and an accuracy of 0.9298 in the analysis of regions of interest in the tumor. The approach was further addressed by Saleem et al. [2021] with the objective of generating visual explanations for the segmentation of 3D brain tumors. It also presented a technique based on class activation maps (CAMs), for the interpretation of the segmentation. According to the authors, the use of a methodology for interpretability that does not depend on gradients to generate visual explanations, yields better results for segmentation, surpassing other techniques such as Grad-CAM.

Another group of algorithm based on the Prototype Classification approach Chen et al. [2019] has also pursued more interpretable diagnoses. In the XProtoNet proposal, an interpretable framework is presented for diagnosing chest radiographs that uses probable areas of occurrence of a certain disease as prototypes with discriminative characteristics, and compares them with the characteristics of the query image. The authors claims that XProtoNet achieves better performance for 10 out of 14 diseases, when compared to competitor classifiers. Another work based on Prototype Classification, the Gen-ProtoPNet Singh and Yow [2021], presents an interpretable deep learning model for detecting Covid-19 in chest radiographs. Unlike the ProtoPNet approach, the model generalizes the L2 distance function and uses any type of spatial dimension in its prototypes to ensure greater precision and more correct interpretation reasoning. More recently, Singla et al. [2021] present an approach with counterfactual explanations, where regions that the classifier considers to be associated with the disease are highlighted. The authors concluded that the explanations generated by the technique were consistent with the diagnosed diseases.

3. Shapley Values

SHAP values were introduced by game theory economist Lloyd Shapley. The values represent, in a game, the contribution of each player of a team in a fair way.

The calculation for the importance of a player *i* is made from the difference in the contribution of a team *S* to which the player is added, by the contribution of the same team *S* without the player. The difference is obtained by a function *c* that calculates the resulting contribution:

$$c_{\text{result}} = c(S \cup \{i\}) - c(S)$$

For each player, contributions are calculated for the combination of all players in *S* that can be formed by the total set of players. The final contribution for each member of the set is given by the average of its contribution with respect to all possibilities of sets that do not include them.

In this work an improved version of DeepLIFT, the *Deep SHAP* algorithm, was used as a method of interpreting recursive predictions for deep learning. In the algorithm, resource contributions are calculated from how an output of a neuron *t* changes as the difference between the activation of an input x_i and the activation of a reference entry. The reference comes from a single input and its choice depends on each domain in which the algorithm is applied.

To calculate the contribution, the algorithm compute multipliers for the entries, given by $m_{\Delta \mathrm{x_i} \Delta \mathrm{t}}$, which are like partial derivatives that inform how much an output changes from a small change in the input. Input multipliers are computed from output neurons using backpropagation. Finally, the contribution to entry *i* is given as the product of its multiplier by its Δi, the difference between the activation and the reference entry:

$$C_{\Delta \mathrm{i_i} \Delta \mathrm{t}} = m_{\Delta \mathrm{i_i} \Delta \mathrm{t}} \times \Delta i.$$

Deep SHAP, emerging as an evolution of the approach, is proposed as a high-speed algorithm for SHAP values. Its implementation differs from DeepLIFT in that it defines its multipliers using Shapley values from not just a reference input, but using a selection of background samples. The algorithm estimates approximate SHAP values by summing the difference between the expected output of each sample and the actual output $f(x) - E[f(x)]$.

Figure 1 presents an example for the explanation of the classification of four inputs taken from the MNIST dataset. The input images are on the leftmost column and are followed by the explanations for each of the classes. In the explanation columns, red pixels indicate an increase in their contribution to the specific class, while blue pixels indicate a decrease in contribution to the same class. For the image with the digit zero, for example, the central part contributes positively to the classification. As for the digit four, the explanation shown in the rightmost column class indicates that the lack of connection at the top of the digit decreases the contribution to classifying the image as a nine.

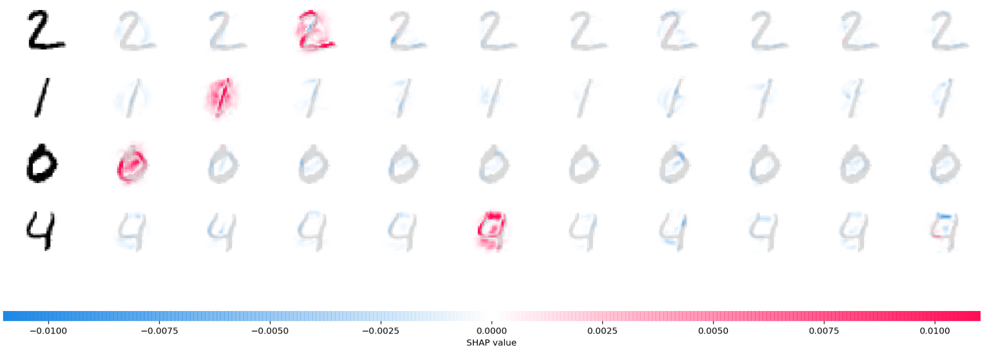


Figure 1. Examples of explanation in the classification of the MNIST dataset.

4. Materials and Methods

The Kermany et al. [2018] database used in this work consists of 5,856 x-rays divided into two categories, normal and pneumonia, selected from pediatric patients aged one to five years. For the database, the images were diagnosed by three physicians for quality control before being made available. Figure 2 presents an example of a database image for each class.

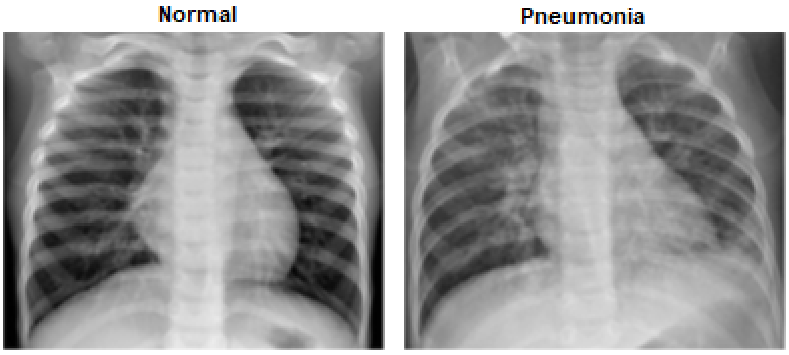


Figure 2. Examples of normal and pneumonia images.

The database images have varying dimensions and an unbalanced proportion between classes. The database consists of 4,273 images with the diagnosis of pneumonia and 1,583 images diagnosed as normal.

The proposed method, *ShapCNN*, is composed of a customized CNN of 12 layers that is trained to perform the first step of the diagnosis. The architecture is shown in Figure 3, and is based on the model described in Giełczyk et al. [2022], where the authors present the impact of pre-processing methods on classification results with a data base of x-rays from the chest for classes normal, COVID-19, and pneumonia. Due to the fact that the database in this work had only two classes, the last fully connected layer of the model was adjusted to 1 unit and consequently the activation function changed from *Softmax* to *Sigmoid* to perform binary classification. In addition, four different pre-trained networks were investigated: Visual Geometry Group 16 (VGG16), Visual Geometry Group 19 (VGG19), MobileNetV2 and DenseNet121. In order to validate the results obtained by ShapCNN, a comparative analysis was performed between the results obtained from each model.

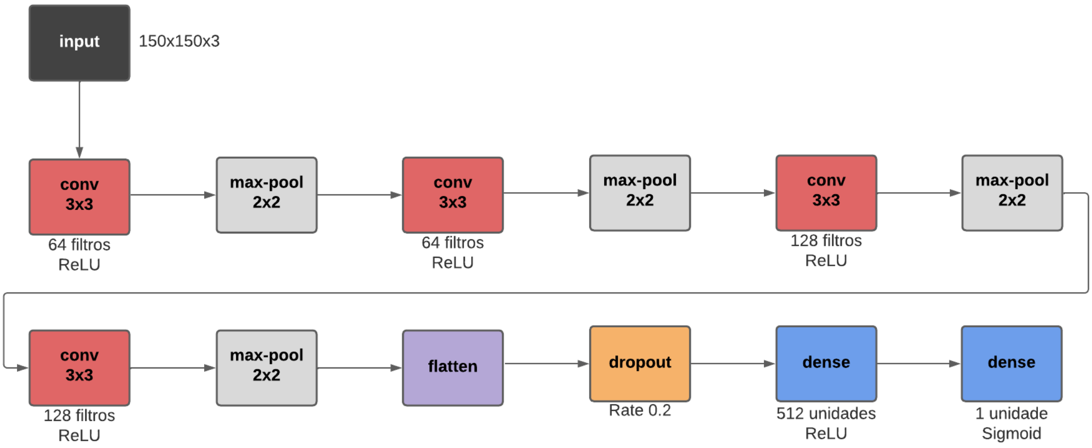


Figure 3. Proposed network architecture of the ShapCNN.

For the training step, as there is an imbalance in the proportion of the database classes, pre-processing was performed to increase the training data. In this process, random transformations are performed on the images such as rotations and displacements to create a new data set.

For the second stage of the process, after training the ShapCNN, visual explanations are generated for the classification of x-rays using the Deep SHAP algorithm that, from each of the previously trained models and a set of sample images, generates an explanation module for each of the networks.

5. Experimental results

The experiments were carried out using the Google Colab service in the PRO version, which provides remote access to machines with a Nvidia Tesla K80 GPU, two vCPUs and 32GB of RAM memory. Python was used in version 3.7.14, the core libraries Tensor Flow with Keras in version 2.8.0, and the SHAP library in version 0.41.0.

5.1. Effectiveness evaluation

The data set was divided into image sets used for training, testing and validation, with 5,216, 624 and 16 images respectively. A pre-processing step was applied to the images.

In order to standardize the input data, the images were resized to 150x150 pixels and normalized so as to speed the convergence of the CNNs. The models were configured with the following parameters: binary cross entropy for the loss function, Adam optimizer with learning rate of 1×10^{-3}, a batch size of 32 and 10 training epochs. The results were obtained from a 10-fold cross-validation.

In order to evaluate the impact of pre-processing by the increase of the sample size, training tests were initially performed without using data augmentation. Table 1 presents the results for the metrics and running time for training and prediction for each CNN.

Table 1. Classification performance in 10-fold validation without data augmentation.

Model	Accuracy	Precision	Recall	F1-Score	Training*	Predict*
VGG19	0,77	0,86	0,69	0,70	487.24	0,21
ShapCNN	0,76	0,86	0,68	0,69	488,28	0,04
DenseNet121	0,74	0,85	0,66	0,66	889,59	0,03
VGG16	0,73	0,84	0,64	0,62	494,89	0,70
MobileNetV2	0,62	0,31	0,50	0,38	514,92	0,25

*Time in seconds

In the first set of tests, the models showed satisfactory accuracy rates when dealing with unbalanced data, with the exception of MobileNetV2 that seems to suffer more from unbalancing. In the other models, despite having good accuracy, the Recall and F1-Score metrics showed the difficulty of the models to perform classification with unbalanced data.

Table 2 presents the results of the models using data augmentation. For the transformations, a new dataset was generated with images created from 20º rotations, displacements of 10% of the total width, and zoom of 20%. The model parameters were maintained and the results were also obtained from a 10-fold cross-validation.

Table 2. Classification performance in 10-fold validation with data augmentation.

Model	Accuracy	Precision	Recall	F1-Score	Training	Predict
DenseNet121	0,93	0,93	0,94	0,93	988,09	0,04
VGG16	0,90	0,90	0,90	0,90	699,68	0,19
VGG19	0,90	0,89	0,90	0,90	681,38	0,09
ShapCNN	0,89	0,91	0,85	0,87	696,93	0,17
MobileNetV2	0,80	0,82	0,75	0,77	702,94	0,01

*Time in seconds

The tests showed improvements in the results of the metrics in all models. DenseNet121 presented the best accuracy with 93%, followed by VGG16 and VGG19, where both presented an accuracy rate of 90%, and ShapCNN with 89%. The model that presented the lowest performance was the MobileNetV2, with an accuracy rate of 80%. For Recall and F1-Score metrics, MobileNetV2, despite obtaining better results compared to the first tests, was still slightly below the other networks, which indicates that the model still suffers from the issue of class imbalance. In addition, changes in running times can be observed. As expected, training times were higher with increasing data for all models. However, the prediction times showed variations with different behaviors, where some models showed an increase while others showed a decrease in the elapsed time. Therefore it is important to consider that the execution environment may have influenced running time.

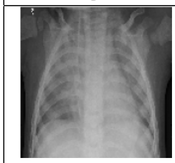
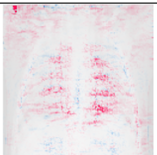
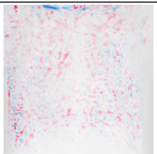
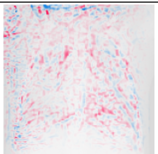
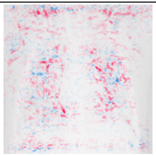
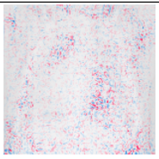
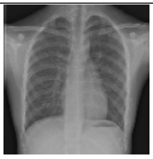
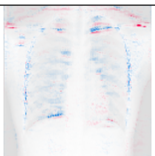
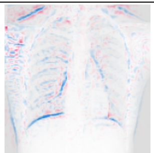
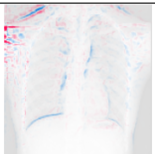
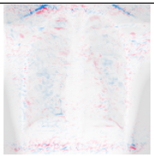
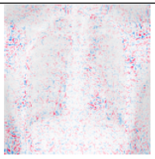
5.2. Interpretability

For the interpretation of the models, the Deep SHAP algorithm was used to generate an explaining module for each of the networks. As previously presented, the algorithm needs a trained model, and a sample dataset. For this latter parameter, it should be noticed that the algorithm complexity scales linearly with the number of images in the set and, therefore, although the use of the entire training database would provide more accurate results, groups of 100 and 200 images randomly chosen from the database were defined for the generation of explainers for each of the models.

As the models carry out a binary classification, the outputs of the explainer modules present, in addition to the original image, an image with the interpretations, where red pixels indicate a contribution from that region in favor of the pneumonia class, and blue pixels indicate a contribution from the region in favor of the normal class. For comparison purposes, the indicated regions with greater contributions were analyzed considering the diagnosis of the original image. In order to validate the generated explanations, it is important to say that for an accurate evaluation of the results, an analysis together with specialists in the field of radiology would be necessary.

Several tests were carried out with the two sizes of sets for all models. Table 3 presents the results of explainers generated using a set of 100 sample images. The first x-ray is from a patient diagnosed with pneumonia, and the second is from a healthy patient. The choice of the exemplified images was based on the most common pattern of the results obtained in the tests.

Table 3. Examples of explanations for pneumonia and normal samples.

Original	ShapCNN	VGG16	VGG19	DenseNet121	MobileNetV2
					
					

In the case of the image with the diagnosis of pneumonia, ShapCNN mostly presented regions of interest that contributed to the correct classification of the image, in addition to the fact that its emphasized regions were presented mainly in lung regions, indicating patterns of pneumonia. Among the pre-trained networks, VGG16 and VGG19, did not clearly show patterns that could generate some explanation, as well as the MobileNetV2, that presented the worst interpretation results. The DenseNet121, on the other hand, demonstrated some patterns in inner regions of interest of the lungs.

For the results of the second image shown in Table 3 referring to healthy subject, ShapCNN and VGG16 emphasized regions of interest in internal regions and at the edges of the lungs, which contributed to the correct diagnosis. The VGG19 presented few regions that lead to the interpretation of the diagnosis. The MobileNetV2 still showed the worst results for interpretability with an absence of patterns in important regions for classification purposes. Finally, the DenseNet121 highlighted some small areas in the the lungs that contributed to the normal diagnosis, but not in a significant way, like the explanations presented by ShapCNN and VGG16.

For the tests using a set of 200 sample images, the results did not show noticeable differences compared to the explainers generated with 100 sample images for the same radiographs. Numerous tests were performed and no improvements were obtained for none of the five CNNs. In the following section, a discussion of the experiments and results is presented.

5.3. Discussion

In this work, the performance of Convolutional Neural Networks and their interpretations for the classification of x-rays into pneumonia and normal classes were evaluated.

The database used has problems regarding to the balancing of classes. For this reason, a data augmentation procedure was used to generate a new training set. The technique proved to be effective as expected, causing significant improvements in the results, in which the DenseNet121 presented the highest accuracy with 93%. It is possible that the exploration of other pre-processing techniques may improve even more the results, mainly for the MobileNetV2 that obtained the lowest accuracy. In general, the training process of the CNNs showed to be effective for the task, although much improvement can be further achieved.

For the interpretation aspect, an explainer was generated for each of the models trained using different sample sizes. It was possible to observe a difference in the interpretations generated for ShapCNN in comparison with the pre-trained networks. During the performance of numerous tests, ShapCNN demonstrated more defined regions of interest in the lung area for the interpretation of the classification of both classes. The interpretation of the model can also be improved, as some attention regions still fall out of the lungs portion, which, in principle, should not be used for diagnosis. Figure 4 exemplifies these findings, where the model appears to detect patterns that contribute positively to the correct classification for pneumonia, but also with focuses on areas outside the region of interest.

Figure 5 depicts another classification result for pneumonia from ShapCNN. In this case, the explanation map emphasizes small areas in the regions of interest, and also other areas with more subtle contributions to the correct class, that cannot be considered a well-defined explanation. This example demonstrates that the model can still be adjusted to provide better interpretation, despite its already high levels of accuracy.

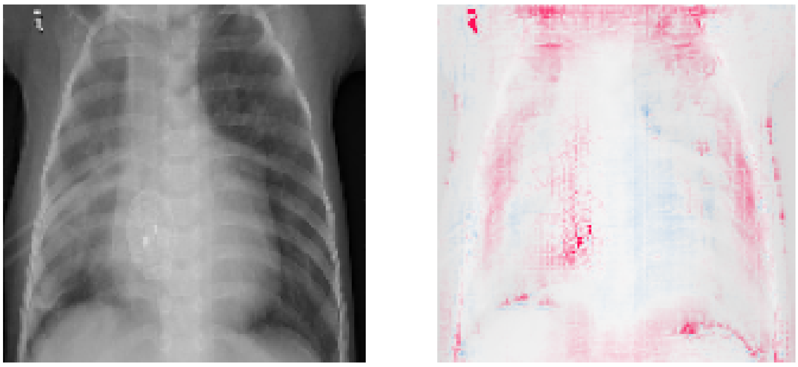


Figure 4. Example of pneumonia image with the explanation map computed by ShapCNN in which the attention regions are well defined.

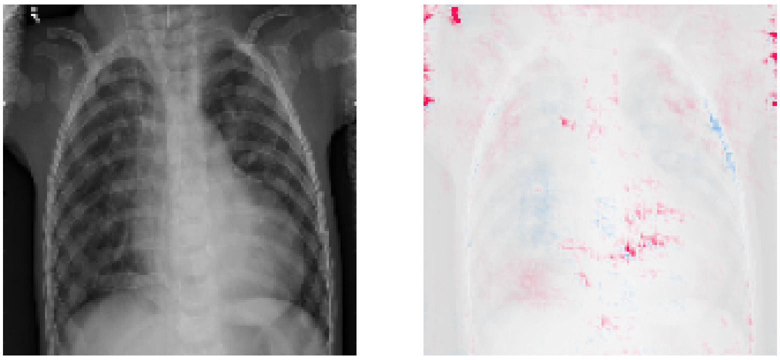


Figure 5. Example of pneumonia image with the explanation map computed by ShapCNN in which the attention regions are not well defined.

Limitations in the computational environment prevented the use of larger samples in the implementation of the explainer modules with consequent decrease in precision. Changing from 100 to 200 images did not produce noticeable changes in the explanation ability of the networks. Thus, there is a need to consider other test environments in order to overcome this limitation and to use a larger number of samples in future investigation.

From the analyses and comparisons of the models mentioned in section 4 for classification and interpretation of x-ray images, with the aim of identifying a pattern that indicates the presence of pneumonia in these images, it can be concluded that ShapCNN obtained a competitive performance, in comparison with the other networks, by jointly carrying out the diagnoses and the explanations for their decisions. The model presented good results for the classifications, despite not obtaining the best accuracy. However, it demonstrated to be superior with respect to interpretation, presenting better defined regions related to the disease effects when compared to other models.

6. Conclusion

In this work, a new method based on deep learning and the approximation of SHAP values was proposed to perform an accurate and clear diagnosis of pneumonia in X-ray images, with explained results. Even though the accuracy obtained by the DenseNet121 was higher, the proposed method was able to provide better explanations during the data interpretation phase. Considering the criteria of accuracy together with the ability to explain the results, the ShapCNN proved to be competitive, so that this combination of techniques should deserve attention in the development of computer-aided diagnosis applications.

Further improvement of this work should consider a systematic way to evaluate the interpretations, in addition to qualitative visual inspection. Future implementations of the architecture using large-scaled computational resources may overcome limitations in the number of images used in the experiments. It is also possible to expand the database to classify bacterial and viral types of pneumonia, as well as to explore other pre-processing techniques and model configurations.

References

C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

A. Giełczyk, A. Marciniak, M. Tarczewska, and Z. Lutowski. Pre-processing methods in chest x-ray image classification. *Plos one*, 17(4):e0265949, 2022.

D. Kermany, K. Zhang, M. Goldbaum, et al. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2), 2018.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh. A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. *Chaos, Solitons & Fractals*, 140: 110190, 2020.

S. Pereira, R. Meier, V. Alves, M. Reyes, and C. A. Silva. Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment. In *Understanding and interpreting machine learning in medical image computing applications*, pages 106–114. Springer, 2018.

S. Ravi, S. Khoshrou, and M. Pechenizkiy. Vidi: Descriptive visual data clustering as radiologist assistant in covid-19 streamline diagnostic. *arXiv preprint arXiv:2011.14871*, 2020.

H. Saleem, A. R. Shahid, and B. Raza. Visual interpretability in 3d brain tumor segmentation network. *Computers in Biology and Medicine*, 133:104410, 2021.

A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

G. Singh and K.-C. Yow. An interpretable deep learning model for covid-19 detection with chest x-ray images. *Ieee Access*, 9:85198–85208, 2021.

S. Singla, B. Pollack, S. Wallace, and K. Batmanghelich. Explaining the black-box smoothly-a counterfactual approach. *arXiv preprint arXiv:2101.04230*, 2021.

UNICEF. A child dies of pneumonia every 43 seconds, 2022. `https://data.unicef.org/topic/child-health/pneumonia/`, Last accessed on 2022-10-30.