PnuemoNet: An Introductory Study On Pneumonia Detection From Chest X-Rays Through The Use of CNN.

LaTeX template adapted from:
European Conference on Artificial Intelligence

Stefano Veglia[1]

Other group members:
Francesco Murgioni[2], **Aniston Anton Thiruchelvam**[3] **Nir Peretz**[4]

Abstract. Pneumonia is still today one of the leading causes of death in underdeveloped countries and among young human beings. The use of machine-based automatic classification from X-rays is ideal for disease screening. We use a data set with more than 5873 images. Our data set can be considered small and unbalanced, having more than three times the cases of pneumonia over regular X-rays. We use weight balancing, data augmentation, oversampling and cross-validation methods to improve the neural network performance. The final results are considerably high and are achieved with the help of an algorithm we developed to combine the prediction of four different models, bringing the accuracy close to perfect. Our study may provide an essential basis for helping third-world countries, and fastening up the system to give diagnoses could also help prevent death, guaranteeing a fast and reliable source for diagnosis of pneumonia.

1 Introduction

Worldwide, pneumonia is thought to be the leading cause of death for children. Pneumonia claims the lives of almost 1.4 million kids annually. Approximately two billion individuals contract pneumonia worldwide on an annual basis. [10] As of right now, chest X-rays are the most effective way to diagnose pneumonia . Pneumonia is not always evident on X-rays, and it is sometimes mistaken for other benign abnormalities or other diseases. [9] Low-resource countries need more qualified radiologists as well, particularly in rural areas. Consequently, there is an urgent need for computer-aided diagnosis (CAD) systems that can assist radiologists in quickly identifying various forms of pneumonia from chest X-ray pictures.

Convolutional deep neural networks have emerged as the industry standard for supervised image classification, demonstrating exceptional performance across various tasks, but their result can be affected by the class imbalance. [1] In order to address this unbalances there are several methods that can be used and can be categorized into data level and algorithm level.

Data-level approaches attempt to reduce the level of imbalance through data sampling methods. We explore two strategies: Oversampling and Synthetic Minority Over-sampling Technique (SMOTE). The first strategy was applied to the minority class. In our case, the X-ray represented patients with no pneumonia and, through the application of geometrical transformation, generated new images to add to the dataset. A study by Mohammed, Roweida, Jumanah Rawashdeh, and Malak Abdullah on unbalanced datasets suggests that the best approach could be to apply oversampling and undersampling on the majority class. [8] Due to the small dimensions of our data set, we decided to apply only the oversampling strategy. SMOTE takes the whole data set, removes the main features from the images, and creates synthetic photos to balance the data. [2]Unfortunately, we encountered some difficulty and obtained very inconsistent results and decided not to proceed on this route.

In order to mitigate bias towards the dominant group, algorithm-level techniques for imbalanced deep learning were also developed. These techniques are typically applied with a class weight or penalty for handling class imbalance. We tested a simple application assign weights to the two classes based on the number of elements.

Our research also focused on the tuning of the training parameters like data augmentation during the training. The use of data augmentation during the training add robustness to the model and reduce over fitting. [12] Other strategies adopted in the training are the use of ReLU activation and Kernel regularizer allowing the model to learn complex patterns and relationships within the data.

We train the model using cross validation in order to have a reliable estimate of the model's performance and to verify any sign of over fitting or under fitting. The idea of cross-validation is to divide the data into subsamples. Each subsample is predicted using the ML model learnt from the remaining subsamples, and the estimated error rate is the average error rate from these subsamples. The ML model finally used is calculated from all the data. Cross-validation gives a better estimate of the error rate than train/test at the cost of more computation. [7]

At last we try to combine the four different model developed in order to achieved a better accuracy.

[1] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: sv3511y@gre.ac.uk
[2] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: fm1087f@greenwich.ac.uk
[3] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: aa5315i@gre.ac.uk
[4] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: np9769y@greenwich.ac.uk

2 Background

Our research is based on the data set developed by Mendeley data [5]. The data set is original divided in three categories: viral pneumonia, bacterial pneumonia and normal chest x-ray and has 5873 images. We joined the two pneumonia categories to reach a class of 4282 images, while the class of normal images has 1591 samples. We can consider this data set small and unbalanced for the difference in size between the two classes. Figure 1 show the data distribution.

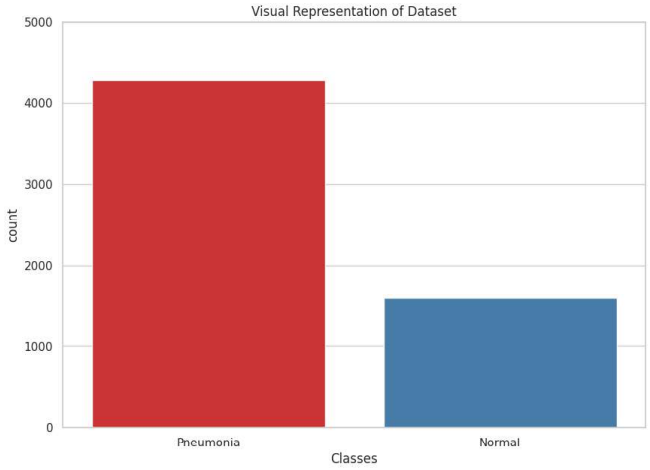

Figure 1. Visual Dataset Representation

We use the research made by Khanfashee posted on Kaggle.com that utilize the previous mentioned data set the detect pneumonia with CNN using transfer learning from Mobilenet. This project was very help full for us in order to understand image pre-processing, data visualization and application of class weights. [6] Another project from Kaggle.com that has been very important to us is the work of Jedrzej Dudzicz. Jedrzey use a different approach from the previous work mentioned, using a CNN created layer by layer, but still applying transfer learning between two different models. From this project we took inspiration from the structure of our model and for the use of the callback function to reduce the learning rate and stop the training with the function "EarlyStopping". [4]

3 Experiments and results

The first step in our project was to import the data and split it into training, testing, and validation. We decided to use a classic division of 80/15/5 to have a larger slice of images for our movement. The size of our database dictates this choice, and being small forces us to use as many images as possible to reinforce the training and create a robust model. We visualize samples from each one of these groups and check if the number of images is congruent. The second step is the pre-processing phase; we resize the images to a standardized dimension while converting them to grayscale. Following this, we shift the images from grayscale to a three-channel representation, effectively creating RGB images. Integral to this process is the normalization of pixel values, ensuring their uniformity within a defined range. We also label each image by scrutinizing the image filenames and accurately assigning labels; images bearing 'NORMAL' or 'IM' are categorized as representing normal conditions (marked as 0), and images containing 'virus' or 'bacteria' are identified as indicative of pneumonia (labelled as 1). The next step is to prepare the model for the training. First, we define the class weight to supply the imbalance between the two classes. These weights will be used during the training. After that, we created an image generator to augment the images during the exercise. We test the model with different parameters to find the best setting for the model to learn efficiently, creating three different augmentation settings. We train the same model to understand which combination gives better results. The three augmentation settings are zero, moderate, and high. As shown in Figure 2 3, the model learns fast and better with a medium setting. Considering the nature of our problem, we were naive to create an image generator for testing and validation because new images would always be presented to the model in a regulated way. We started with a simple model called Giovanna 1 with two convolutional layers presenting batch normalization, max pooling and dropout, followed by two dense layers with sigmoid activation. Figure 2 and Table 1 show the testing and training results.

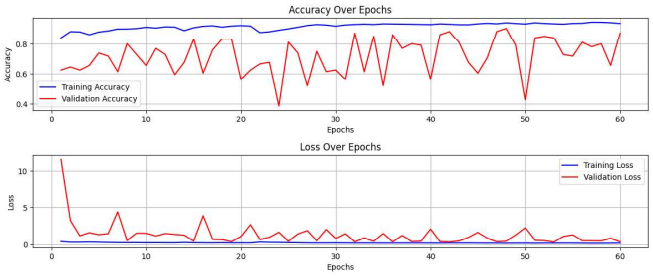

Figure 2. Training and validation, loss and accuracy for Giovanna 1

[H]

Model	Test Loss	Test Accuracy
Giovanna 1	0.49	0.88

Table 1. Test results of Giovanna 1

Given the previous result, we updated the model, adding three convolutional layers and two dense layers to capture more intricate and complex features present in the data and achieve better accuracy. Figure 3 and Table 2 show the results of the testing and the training.

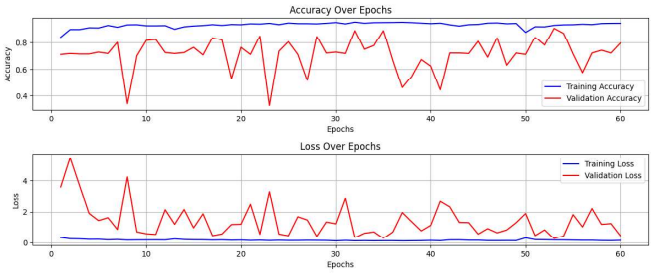

Figure 3. Training and validation, loss and accuracy for Giovanna 2

[H]

Model	Test Loss	Test Accuracy
Giovanna 2	0.38	0.86

Table 2. Test results of Giovanna 2

The last modification to the model was to double the filters on the convolutional layers; adding two identical convolutional layers allows the network to learn more complex and diverse features by

processing the output of the first layer in the subsequent layer. We also added a kernel regularizer and the callback function to reduce the learning rate and handle overfitting better. Figure 4 and Table 3 show the testing and training results.

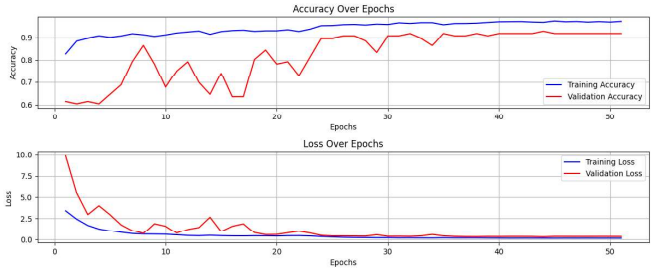

Figure 4. Training and validation, loss and accuracy for Giovanna 3

[H]

Model	Test Loss	Test Accuracy
Giovanna 3	0.157	0.964

Table 3. Test results of Giovanna 3

Following the previous results, we adopted oversampling to balance the data set. We first convert the images into a format suitable for processing. We define two sets of augmentation parameters for image transformation, splitting the photos into halves and applying different augmentation techniques to each half. We generate a specified number of augmented images based on two other augmentation parameters, ensuring a blend of consistency and diversity in the expanded data set—Figure 5.

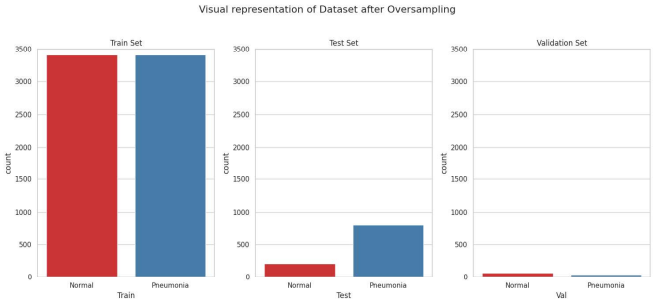

Figure 5. Visual representation of the data after oversampling

We than train the model on this new dataset obtaining slightly worst results how shown in the Figure 6 and Table 4.

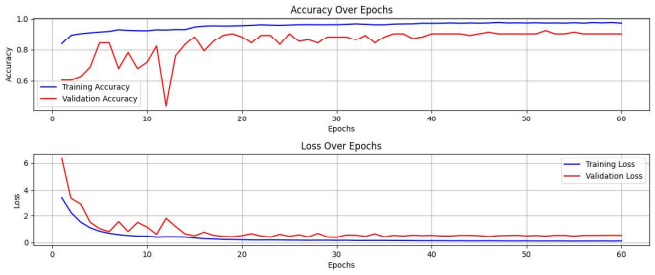

Figure 6. Training and validation, loss and accuracy for Giovanna 3 on oversample dataset

[H]

Model	Test Loss	Test Accuracy
Giovanna 3	0.157	0.964

Table 4. Test results of Giovanna 3

The next step is cross-validation. We employ a five-fold stratified cross-validation approach. We initialize our model each time and train it on augmented data from the training set. The model's performance is then evaluated on the corresponding test set. Table 5 shows the accuracy and loss for each k-fold in the following table. The result showed a slight variation in effects, probably caused by overfitting.

Subsequence	Test Loss	Test Accuracy
1	24.32%	91.71%
2	18.68%	95.59%
3	22.23%	93.73%
4	15.04%	95.88%
5	17.35%	94.06%

Table 5. Cross-Validation results with five different subsequences

The last step of our journey is the combination of the models trained during the research. Thanks to the confusion matrix, we assume the strengths and weaknesses of each algorithm when it comes to identifying one of the two classes. The combinations of the four algorithms, as shown in Table 6, help us to achieve higher accuracy.

	Bagley	Mk54	Lola	Giovanna	Mean	M and W
w.p	44	35	33	30	25	27

Table 6. Combination of models - w.p = wrong prediction, m = mean, m and w = majority and weights

4 Discussion

The experiment process undertaken followed mainly a logic pattern. We first consider the nature of our data set and explore solutions to adjust the imbalances and provide a better starting point for our training. In the article [3], the authors give an excellent overview of data augmentation that helps us to define the parameter set for the generators and also provides us with the idea to balance the data set with the use of oversampling. The results are similar, but the oversampling process could have been done better with advanced augmentation or synthetic data creation. During the creation of the model, we found difficulties in designing a model that could be complex enough to capture all the hidden features from the images while avoiding overfitting. Using the callback function and applying kernel regulation was pivotal in the attempt to avoid overfitting in a model with a large number of layers. The choice of creating a new model and not using well-known transfer learning techniques from the pre-trained model was dictated by the will to understand better the impact of every single layer on the model's training and be able to adjust the whole model. The use of cross-validation to ensure robust model evaluation, leveraging data augmentation for improved generalization and reliability in the classification task, showed the model's good reliability. From the article [11], we deduct that cross-validation is critical when working on a small data set. Also, we followed the author's suggestion, used only the training data set for the cross-validation, and reserved the testing data for a second validation. We consider the option out of simple intuition for the last and final step of model combination. In the article [13], a similar technique is presented, calculating the mean of the model's predictions and assigning weights

to each model. In our model combination algorithm, the choice will be based on most of the output instead of the mean of the outcomes. In case of a draw between the four models, we prioritize the models with a higher accuracy to detect a particular class.

5 Conclusion and future work

The training of convolutional neural networks can be costly in computing units and time-consuming, significantly if not aided by transfer learning. We had many challenges, starting from the imbalanced dataset and experimenting with different techniques to mitigate biases. Our rigorous approach included parameter tuning, cross-validation, and model amalgamation, culminating in enhanced accuracy. Overall, we train a reliable model with reasonable accuracy that could work as a hospital support tool. Moving forward, exploring more sophisticated data augmentation techniques, refining model complexity, exploring advanced model combinations and working on more extensive and diverse data sets could push the boundaries of accuracy in pneumonia detection.

ACKNOWLEDGEMENTS

I want to thank my teammates. They have been an incredible source of motivation and support throughout the project. Also, I'd like to thank everyone involved in the module "Introduction to AI" at the University of Greenwich for the excellent learning opportunity and the passion transmitted to us during the lectures and labs.

REFERENCES

[1] Lei Cai, Jingyang Gao, and Di Zhao, 'A review of the application of deep learning in medical image classification and segmentation', *Annals of translational medicine*, **8**(11), (2020).

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, 'Smote: synthetic minority over-sampling technique', *Journal of artificial intelligence research*, **16**, 321–357, (2002).

[3] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth, 'A review of medical image data augmentation techniques for deep learning applications', *Journal of Medical Imaging and Radiation Oncology*, **65**(5), 545–563, (2021).

[4] Jedrzej Dudzicz. Pneumonia detection on chest x-ray, 2020.

[5] Daniel Kermany, Kang Zhang, Michael Goldbaum, et al., 'Labeled optical coherence tomography (oct) and chest x-ray images for classification', *Mendeley data*, **2**(2), 651, (2018).

[6] Khanfashee. Medical image classification for beginner, 2020. Medical Image Classification For Beginner.

[7] Ross D King, Oghenejokpeme I Orhobor, and Charles C Taylor, 'Cross-validation is safe to use', *Nature Machine Intelligence*, **3**(4), 276–276, (2021).

[8] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah, 'Machine learning with oversampling and undersampling techniques: overview study and experimental results', in *2020 11th international conference on information and communication systems (ICICS)*, pp. 243–248. IEEE, (2020).

[9] Mark I Neuman, Edward Y Lee, Sarah Bixby, Stephanie Diperna, Jeffrey Hellinger, Richard Markowitz, Sabah Servaes, Michael C Monuteaux, and Samir S Shah, 'Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children', *Journal of hospital medicine*, **7**(4), 294–298, (2012).

[10] Tawsifur Rahman, Muhammad E. H. Chowdhury, Amith Khandakar, Khandaker R. Islam, Khandaker F. Islam, Zaid B. Mahbub, Muhammad A. Kadir, and Saad Kashem, 'Transfer learning with deep convolutional neural network (cnn) for pneumonia detection using chest x-ray', *Applied Sciences*, **10**(9), (2020).

[11] R Bharat Rao, Glenn Fung, and Romer Rosales, 'On the dangers of cross-validation. an experimental evaluation', in *Proceedings of the 2008 SIAM international conference on data mining*, pp. 588–596. SIAM, (2008).

[12] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann, 'Data augmentation can improve robustness', *Advances in Neural Information Processing Systems*, **34**, 29935–29948, (2021).

[13] Jamal Zaherpour, Nick Mount, Simon N Gosling, Rutger Dankers, Stephanie Eisner, Dieter Gerten, Xingcai Liu, Yoshimitsu Masaki, Hannes Müller Schmied, Qiuhong Tang, et al., 'Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models', *Environmental modelling & software*, **114**, 112–128, (2019).