Chest X-Ray Image Pneumonia Classification - model performance study

1st Fabiana Rodrigues da Silva
FEUP, University of Porto
Porto, Portugal
up202100810@fe.up.pt

2nd Gabriel Copolecchia Carvalhal
FEUP, University of Porto
Porto, Portugal
up202103616@fe.up.pt

3rd Guilherme Carlos Salles
FEUP, University of Porto
Porto, Portugal
up202100811@fe.up.pt

***Abstract*—This study explores CNN's applications on chest X-ray images for a classification task, including full implementation of Convolutional Neural Networks and the Vision Transformer. Also, it was applied techniques of transfer learning and fine-tuning for the models VGG-16, VGG-19, and ResNet-50. Therefore, it was suggested an ensemble method for increasing the robustness and performance of Pneumonia detection and it was evaluated all the models based on accuracy, precision, recall, and f1-score.**

***Index Terms*—medical image classification, convolutional neural network (CNN), supervised learning, chest radiography, transfer learning, visual transformer (ViT)**

I. Introduction

Healthcare professionals make use of machine learning technologies to help with identifying, diagnosing, and predicting diseases, as well as personalizing treatments for patients. One of the most common viral infections is pneumonia [1], which is a serious infection of the human respiratory system and, if not treated well, can cause illness followed by severe health damage. An important diagnostic method for this disease is chest X-rays, which offer the possibility of identifying the location and extent of the lung inflammation. Although using chest radiography images can also be a challenge in identifying pneumonia due to its similarity to other pulmonary diseases. This paper presents a computer-aided supervised classification task to predict pneumonia based on chest X-ray images. The proposal is based on Convolutional Neural Network models and Transformers for exploring different techniques in order to compare results: CNN-trained end-to-end, ViT trained end-to-end, VGG-16, VGG-19, ResNet-50 and an Ensemble method.

II. Related Works

Researchers have used deep learning over the last ten years to detect lung infections and diseases from chest X-ray [2].

Stephen et al. [3] trained a CNN from scratch to extract features from chest X-rays to identify if the patient has pneumonia or not. The study reached a great classifier performance compared to their previous studies based on traditional manual features.

Ikechukwu et al. [4] compared pre-trained models such as VGG-19 and Resnet-50 against CNN training from scratch, applying data augmentation and dropout regularization to avoid overfitting. They achieved an accuracy of 97.3 in VGG-19 and 96.3 in ResNet-50, giving better results as compared to training from scratch.

Training CNN models require a large amount of labeled data, which is computationally expensive and requires advanced machines for processing. In order to solve these problems, the transfer learning (TL) approach has been proposed. In [5], the authors explain that the TL has become very popular, and as a consequence, it requires fewer inputs in the CNN model, reduces costs, and also brings more efficiency.

The authors in [6] used the TL approach by applying four pre-trained CNN architectures on ImageNet [7] for pneumonia detection with three augmentation strategies (rotation, scaling, and translation) to generate new training sets to classify chest X-ray images.

Rajaraman et al. [8] developed a new CNN-based approach that uses only the lung images rather than the entire image, which is known as a "region of interest (ROI)." The CNN learns better characteristics from a specific location in the image in this manner. Although, according to Mabrouk et al. [2], detecting pneumonia with a high degree of efficiency is a challenge using these approaches. The authors also affirm that, besides those interesting approaches, using vision transformer (ViT) brings optimistic results with a small number of features and layers.

III. Methodology

A. Dataset

The dataset used for the study [9] is composed of chest X-ray images from pediatric patients from one to five years of age at the Guangzhou Women and Children's Medical Center. The dataset is organized by two sets of chest x-ray images: NORMAL and PNEUMONIA, divided by subclasses of training, testing, and validation images. Table 1 shows the number of patients divided by each class and subset. The research focuses on whether the patient has pneumonia or not.

TABLE I
THE COMPOSITION OF CHEST X-RAY DATASET

	Training	Testing	Validation
PNEUMONIA	3.875	390	8
NORMAL	1.341	234	8

B. Preprocessing

According to the authors in [10], the preprocessing step has the purpose to enhance information and simplify data. Therefore, the preprocessing steps are essential before applying the images to the CNN model because they can have different sizes and formats, which makes it difficult for the model to learn the patterns in order to extract features. To begin, two functions were created: "get data" and "get training data color." The first one lists the available images, transforms them to grayscale with one channel, and resizes them to (224x224x1). The second one has the same process but keeps three channels of colors, creating an output image of (224x224x3). The function "get data" is required for training CNN end-to-end and ViT. Respectively, "get training data color" is a requirement for the pre-training algorithms Resnet-50, VGG-16, and VGG-19.

As the study is a supervised learning classification task, features and labels are separated to train the models. Thus, features went through normalization, changing the intensity level of pixels by converting them to values between 0 and 1, which facilitates deep learning models to learn.

Figure 1 shows samples of normal and pneumonia chest X-ray images from the selected dataset.

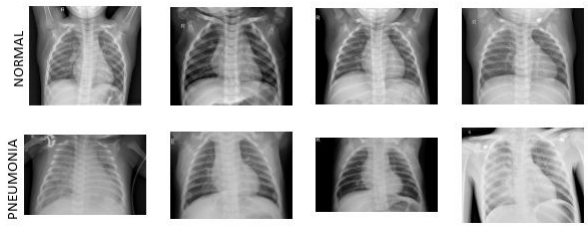

Fig. 1. Samples of chest x-ray from the selected dataset

C. Data augmentation

Data augmentation is an approach of enriching a dataset by producing new data samples from the original dataset. Shorten and Khoshgoftaar [11] declare that data augmentation helps to avoid overfitting from the root of the problem, once it is applied to the training dataset. For this study, the data augmentation approach is also applied to handle the imbalanced dataset, using the following techniques to generate new images:

1) rotation_range = randomly rotate images in the range of 30 degrees
2) zoom_range = shear intensity of 0.2 angles in a counter-clockwise direction in degrees
3) width_shift_range = shift the image to the left or right (horizontal shifts), considering the percentage of total width as a range of 0.1
4) height_shift_range = shift the image to the up or down (vertical shifts), considering the percentage of total height as a range of 0.1
5) horizontal_flip = True. Randomly flip inputs horizontally.

IV. Problem implementation

Five different experiments using Convolutional Neural Networks (CNNs) were applied in this study, and the following sections will detail each one.

A. CNN End-to-End

The CNN end-to-end was created based on research made on Kaggle [12] for benchmarking CNN's with great accuracies, in order to compare with traditional architectures. The CNN architecture used as an example achieves an accuracy of 92.6%. Although, for this study, for testing purposes, some additional changes were applied based on the example of a CNN architecture, for example decreasing the number of layers, changing the number of image channels, and applying specific data augmentation transformations. In Figure 2 can be seen the architecture of this CNN.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 224, 224, 32)	320
batch_normalization (BatchNormalization)	(None, 224, 224, 32)	128
max_pooling2d (MaxPooling2D)	(None, 112, 112, 32)	0
conv2d_1 (Conv2D)	(None, 112, 112, 64)	18496
dropout (Dropout)	(None, 112, 112, 64)	0
batch_normalization_1 (BatchNormalization)	(None, 112, 112, 64)	256
max_pooling2d_1 (MaxPooling2D)	(None, 56, 56, 64)	0
conv2d_2 (Conv2D)	(None, 56, 56, 64)	36928
batch_normalization_2 (BatchNormalization)	(None, 56, 56, 64)	256
max_pooling2d_2 (MaxPooling2D)	(None, 28, 28, 64)	0
conv2d_3 (Conv2D)	(None, 28, 28, 128)	73856
dropout_1 (Dropout)	(None, 28, 28, 128)	0
batch_normalization_3 (BatchNormalization)	(None, 28, 28, 128)	512
max_pooling2d_3 (MaxPooling2D)	(None, 14, 14, 128)	0
conv2d_4 (Conv2D)	(None, 14, 14, 256)	295168
dropout_2 (Dropout)	(None, 14, 14, 256)	0
batch_normalization_4 (BatchNormalization)	(None, 14, 14, 256)	1024
max_pooling2d_4 (MaxPooling2D)	(None, 7, 7, 256)	0
flatten (Flatten)	(None, 12544)	0
dense (Dense)	(None, 128)	1605760
dropout_3 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 1)	129

Fig. 2. Convolutional Neural Network (CNN) Model

The neural network was compiled and made use of optimizer 'rmsprop' and binary_crossentropy as a loss. RmSprop optimizer maintains a moving discounted average of the square of gradients, dividing the gradient by the root of this average

and using it to estimate the variance. Figure 3 shows details about how a CNN architecture is organized.

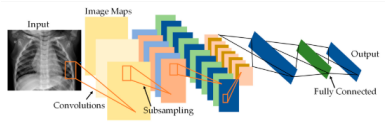

Fig. 3. Convolutional Neural Network (CNN) Architecture
(Rahman et al., 2020)

B. VGG-16 and VGG-19

The Visual Geometry Group (VGG16 and VGG19) concepts from the Convolutional Neural Network (CNN) [13] were used to classify whether a patient has pneumonia or not through chest X-ray images.These pre-trained models are characterized by its simplicity, using only 3×3 convolutional layers stacked on top of each other in increasing depth. Both models reduce volume size with max pooling and use two fully-connected layers, followed by a sigmoid classifier (project approach). The "16" and "19" stand for the number of weight layers in the network. Figure 4 exemplifies the VGG-16 architecture.

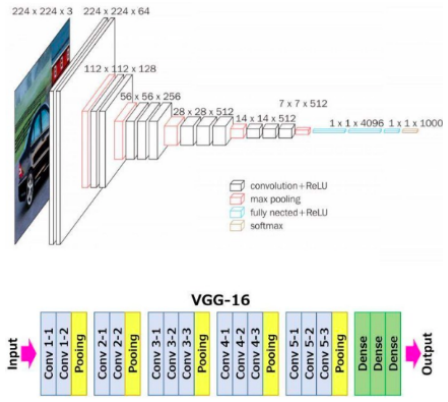

Fig. 4. VGG-16 architecture
(Rohini, 2021)

C. ResNet-50

The ResNet-50 is a convolution neural network that has 50 layers and uses skip connections to avoid vanishing and exploding gradients. This model was pre-trained on ImageNet [15] also uses 3 channels as input. Besides having 50 layers, the uses of global average pooling help to optimize the model. The Figure 5 presents the ResNet-50 architecture.

D. Vision Transformer (ViT)

The ViT model consists of blocks of multiple transformers, which use a few layers with different purposes. The patch encoding layer as shown in Figure 6, splits the input images as a series of patches which, once transformed into vectors, are seen as sequences of images with positional embeddings. Then, a Multi-Head Attention layer with a self-attention mechanism is applied to the sequence of patches to generate the final tensor that will be processed by a classifier head with a sigmoid function. For this model, it was used resized images to 224x224, 36 patches per image, and 8 transformers layers.

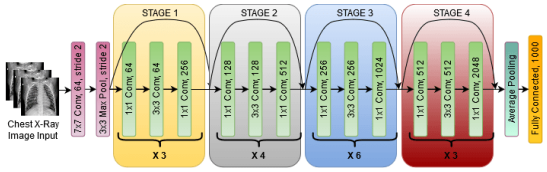

Fig. 5. ResNet-50 architecture
(Abu, 2021)

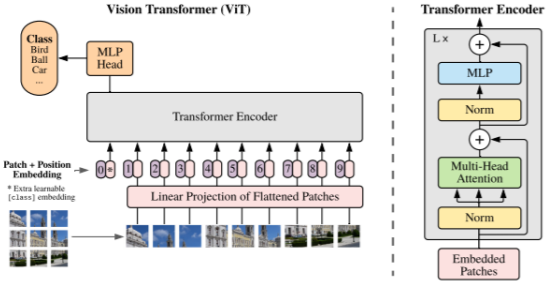

Fig. 6. ViT Model overview
(Dosovitskiy et al., 2020)

E. Ensemble proposal

This section describes the implementation of an ensemble-based strategy to increase the performance and robustness of the predictions. The goal of the proposed method is to use the best models on the extracted medical images, to perform an individual prediction based on the majority-voted results to reach a classification.

The models used in the ensemble strategy were three of the five models that presented the best results in terms of Recall and Accuracy. As shown in Figure 7, the medical images go through the ensemble method composed of three models: CNN end-to-end, VGG-16, and VGG-19. Thus, it computes the output of each model individually, combining them and evaluating them with a majority vote system. Eg. If two of three models predict an output one (Pneumonia), the result of the ensemble will be one. In this way, the final output gets more robust on image prediction, since each prediction has an agreement of at least two different models.

V. Results and Discussions

This study evaluated the representation learning capabilities of a CNN end-to-end and Vision Transformer (ViT) end-to-end, which were implemented manually and trained for all the network layers. The VGG-16, VGG-19, and ResNet-50 were pre-trained models and the main task of this study was to perform the fine-tuning. Figure 8 shows the metrics evaluation of all proposed methods.

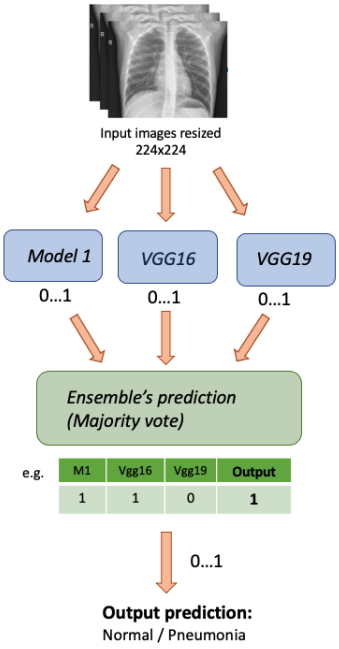

Fig. 7. Ensemble strategy overview

For model evaluation, it was used the metrics: Precision, Accuracy, Recall, and F1-Score. However, it was not weighed equally the metrics. More focus was given to Accuracy since it evaluates all the classifications (TP, FP, FN, TN), and Recall because uses the formula (TP / TP + FN), which focuses on False Negative. In pneumonia detection, it is very important to minimize the False Negatives, because it can represent several consequences in medical terms when the prediction says it is a normal lung, but actually is Pneumonia. In this way, the correct treatment and protocols may not start on time. Figure 10 exemplifies how good or bad each model has classified a patient with pneumonia or normal.

Model	Precision	Accuracy	F1_score	Recall
Model 1	0.920716	0.902244	0.921895	0.923077
Model ResNet	0.861538	0.806090	0.861538	0.861538
Model VGG16	0.933842	0.910256	0.937420	0.941026
Model VGG19	0.957386	0.887821	0.908356	0.864103
Model ViT	0.901734	0.820513	0.847826	0.800000
Ensemble	0.916870	0.921474	0.938673	0.961538

Fig. 8. Models results

Considering the time and scope of the study, overall the VGG-16 architecture presented the best results compared to other proposed methods, as can be seen in Figure 8 and also in Figure 9.

For the ViT model was expected better results. However, it is understandable as it was not used as a trained model. In this way, to improve the model quality without pre-training,

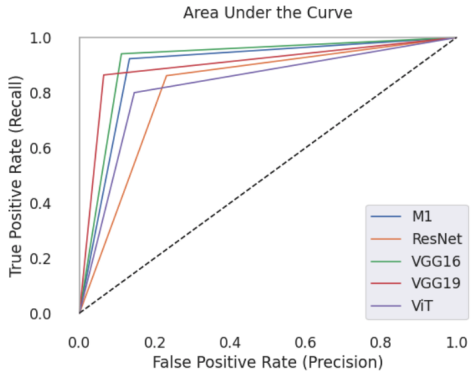

Fig. 9. ROC Models overview

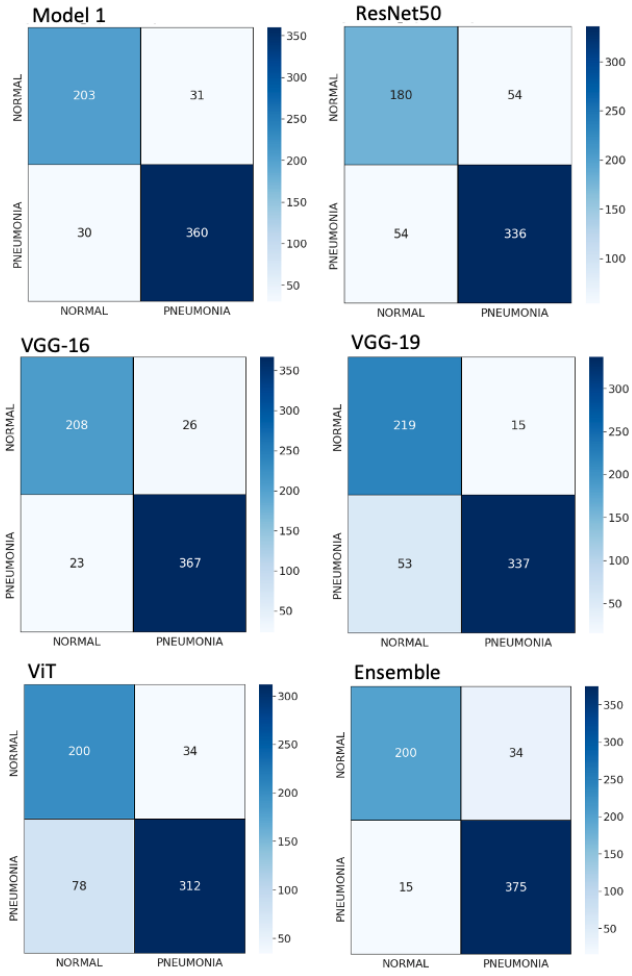

Fig. 10. Confusion Matrix overview

they could be trained the model using more transformation on data augmentation, increasing the number of transformer layers, resizing the input images, changing the patch size, or increasing the projection dimensions, for example. Also, it could be used as a trained model and fine-tuning for this dataset.

The ensemble method presented an interesting result, as it combines the output of the three best models, and is more reliable and robust than one specific architecture. In this way, was possible to get the best of three models and minimize a potential bias from one specific model. Figure 11 presents the classification report for the ensemble model. As can be seen, it shows an overall accuracy of 0.92 and a macro average recall of 0.91. In Figure 12, is also possible to verify the ROC curve from the ensemble and top 3 models, which proves that the ensemble curve is the closest to the benchmark.

	precision	recall	f1-score	support
Normal (Class 0)	0.93	0.85	0.89	234
Pnenomia (Class 1)	0.92	0.96	0.94	390
accuracy			0.92	624
macro avg	0.92	0.91	0.91	624
weighted avg	0.92	0.92	0.92	624

Fig. 11. Classification report from ensemble strategy

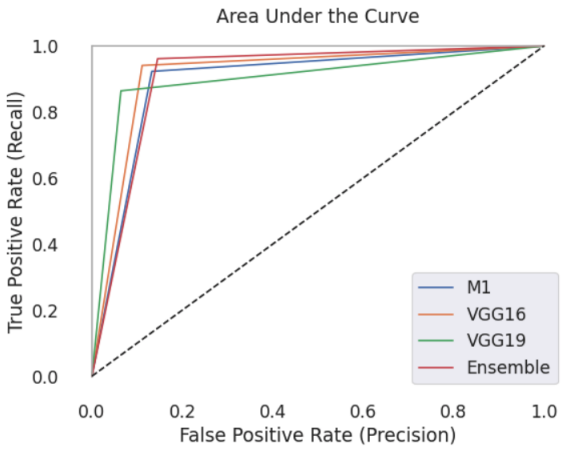

Fig. 12. ROC from Top 3 models and ensemble

VI. Conclusion

Pneumonia is a life-threatening lung infection and identifying if the patient has this disease or not is really important for decision-making on treatments. For pneumonia detection, chest radiography is a common test recommended by doctors, and the detection process in advance can save lives. In this way, technology has been helping the medical domain on identifying and predicting diseases, and artificial intelligence works in collaboration with humans in healthcare. By creating models applying CNN on images is a good approach to fastly classify if patients are suffering from pneumonia or not, and might support the clinician and doctor on the diagnostic. Besides the CNN approach, the Vision Transformer model presented good results and is shown as a promising alternative to explore in future works. The proposal ensemble strategy has shown that is a better approach for the medical domain. It presents the results of three models, which are similar to having the opinions of various doctors to achieve a consensus on patient diagnoses and treatments. In this way, it presents a good strategy to be applied in the field and more developed in near future.

References

[1] Ortiz-Toro, C.; García-Pedrero, A.; Lillo-Saavedra, M.; Gonzalo-Martín, C. Automatic pneumonia detection in chest X-ray images using textural features. Comput. Biol. Med. 2022, 145, 105466. https://doi.org/10.1016/j.compbiomed.2022.105466

[2] Mabrouk, A.; Díaz Redondo, R.P.; Dahou, A.; Abd Elaziz, M.; Kayed, M. Pneumonia Detection on Chest X-ray Images Using Ensemble of Deep Convolutional Neural Networks. Appl. Sci. 2022, 12, 6448. https://doi.org/10.3390/app12136448

[3] Stephen, O.; Sain, M.; Maduh, U.J.; Jeong, D.U. An efficient deep learning approach to pneumonia classification in healthcare. J. Healthc. Eng. 2019, 2019, 4180949. https://doi.org/10.1155/2019/4180949

[4] A.Victor Ikechukwu, S.Murali, R. Deepu, R.C.Shivamurthy (2021). ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. Global Transitions Proceedings, Volume 2, Issue 2, November 2021, Pages 375-381. https://www.sciencedirect.com/science/article/pii/S2666285X21000558

[5] Cheplygina, V.; de Bruijne, M.; Pluim, J.P. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal. 2019, 54, 280–296. https://doi.org/10.1016/j.media.2019.03.009

[6] Rahman, T.; Chowdhury, M.E.; Khandakar, A.; Islam, K.R.; Islam, K.F.; Mahbub, Z.B.; Kadir, M.A.; Kashem, S. Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. Appl. Sci. 2020, 10, 3233. https://doi.org/10.3390/app10093233

[7] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255). 10.1109/CVPR.2009.5206848

[8] Rajaraman, S.; Candemir, S.; Kim, I.; Thoma, G.; Antani, S. Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs. Appl. Sci. 2018, 8, 1715. https://doi.org/10.3390/app8101715

[9] Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, V2, https://data.mendeley.com/datasets/rscbjbr9sj/2

[10] W. Zhou, X. Ma, and Y. Zhang, "Research on image preprocessing algorithm and deep learning of Iris recognition," Journal of Physics: Conference Series, vol. 1621, Article ID 012008, 2020. https://iopscience.iop.org/article/10.1088/1742-6596/1621/1/012008

[11] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0

[12] Mathur M (2020). Pneumonia Detection using CNN(92.6% Accuracy). https://www.kaggle.com/code/madz2000/pneumonia-detection-using-cnn-92-6-accuracy/notebook

[13] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

[14] Rohini, G. Everything you need to know about VGG16, 2021. https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778. https://doi.org/10.48550/arXiv.1512.03385

[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[17] Abu. Overview of Residual Neural Network (ResNet), 2021. https://open-instruction.com/dl-algorithms/overview-of-residual-neural-network-resnet/