Detection of Pathologies in X-Rays Based on a Deep Learning Framework

Jhonatan Camasca
5100036@ue.edu.pe / Universidad ESAN

Marks Calderón-Niquin
mcalderon@esan.edu.pe / Universidad ESAN

Wilfredo Mamani-Ticona
wmamani@esan.edu.pe / Universidad ESAN

Recepción: 1/9/2020 Aceptación: 1/10/2020

Abstract. The diagnostic process of respiratory diseases requires experience and skills to assess the different pathologies that patients may develop. Unfortunately, the lack of qualified radiologists is a global problem that limits respiratory diseases diagnosis. Therefore, it will be useful to have a tool that minimizes errors and workload, improves efficiency, and speeds up the diagnostic process in order to provide a better healthcare service to the community. This research proposes a methodology to detect pathologies by using deep learning architectures. The present proposal is divided into three types of experiments. The first one evaluates the performance of feature descriptors such as SIFT, SURF, and ORB in medical images with machine learning models as an introduction to the last experiment. The second one evaluates the performance of deep learning architectures such as ResNet50, Alexnet, VGG16, and LeNet. Finally, the third one evaluates the combination of deep learning and machine learning classifiers. Furthermore, a novel chest X-ray dataset called PathX_Chest, which contains 2,200 images of ten different classes, is presented. In contrast with the state of the art, good results were obtained in the three different approaches. However, the best performance was achieved by combining deep learning and machine learning: a 99.99 % accuracy was obtained with the combination of ResNet50 and SVM classifier. This methodology may be used to develop a CAD system to help radiologists have a second opinion and a support during the diagnostic procedure.

Keywords: chest X-ray / deep learning / CNN / computer vision / computer-aided diagnosis

Detección de presencia patológica en radiografías basada en un marco de *deep learning*

Resumen. El proceso de diagnóstico de las enfermedades respiratorias requiere experiencia y habilidades para evaluar las diferentes patologías que pueden desarrollarse en los pacientes. Desgraciadamente, la falta de radiólogos cualificados es un problema global que limita el diagnóstico de las enfermedades respiratorias. Por lo tanto, será útil contar con una herramienta que minimice los errores, la carga de trabajo, mejore la eficiencia y agilice el proceso de diagnóstico para brindar un mejor servicio de salud a la comunidad. Esta investigación propone una metodología para la detección de presencia patológica utilizando arquitecturas de *deep learning*. La presente propuesta se divide en tres tipos de experimentos. El primero evalúa el rendimiento de descriptores de características como SIFT, SURF y ORB en imágenes médicas con modelos de *machine learning* como introducción al último experimento. A continuación, se evalúa el rendimiento de arquitecturas de *deep learning* como ResNet50, Alexnet, VGG16 y LeNet. Por último, se evalúa la combinación de clasificadores de aprendizaje profundo y aprendizaje automático. Además, introducimos un nuevo conjunto de datos de rayos X de tórax que se llama PathX_Chest y que contiene 2200 imágenes de diez clases. En contraste con el estado del arte, se obtuvieron buenos resultados en tres enfoques diferentes. Sin embargo, podemos ver que el mejor rendimiento se logró en la mezcla entre *deep learning* y *machine learning*, obteniendo una precisión del 99,99 % en la combinación de ResNet50 y el clasificador SVM. Esta metodología puede ser utilizada para desarrollar un sistema CAD con el fin de ayudar a los radiólogos permitiéndoles tener una segunda opinión y como apoyo durante el procedimiento diagnóstico.

Palabras clave: radiografía de tórax / aprendizaje profundo / CNN / visión por ordenador / diagnóstico asistido por ordenador

1. INTRODUCTION

Pathology detection is a time-consuming process that involves knowledge, experience, concentration, and a patient's medical history. On the other hand, according to the American College of Radiologist (ACR), radiologist shortage is seen in developed countries as well as the least developed countries. For instance, in the USA, UK, and Australia, this problem affects hospital care and service delivery in some medical areas. In Peru, according to the Ministry of Health (MINSA), 66.4 % of the radiologists are in Lima and the rest are in provinces, where most of the pathologies occur. Due to the shortage of specialists and the complexity of the diagnostic process, a tool that could help radiologists and give them a second opinion might improve their performance in terms of speed, efficiency, and error detection. Meanwhile, according to The Journal of Health (2020), the potential of the AI in medical imaging could accelerate the diagnostic process, provide target-focused treatments, and enhance human-led clinical decision. In pathology detection, several approaches were developed using chest X-rays: The most common ones involved feature descriptors, and the most advanced ones involved deep learning architectures known as convolutional neural networks (CNNs). However, due to the nature of the problem, it is necessary to focus more on precision.

France & Jaya (2019) used patch and SIFT as feature extraction process. Thus, through clustering models such as bag of words (BOW) and histogram of bag of words (HOG), features were obtained. In the end, SVM was applied for the classification into normal and abnormal. On the other hand, Saric et al. (2019) trained VGG16 and ResNet50 for detecting lung cancer, which reached up to 0.75% accuracy in cancer classification. X-ray image classification is a difficult task if there are a few images. Rahman et al. (2020) proposed three classes: normal, pneumonia, and viral pneumonia. Afterwards, during the classification, AlexNet, ResNet18, DenseNet201, SqueezeNet architectures, and their respective weights were trained. Dong, Y. (2017) trained VGG-16 and ResNet-101 for binary and multi-classification tasks. They reached up 82.2% accuracy in binary classification. Object detection models are widely used for detecting elements in images and classifying them. In Rahmat et al. (2019), Faster R-CNN was used for binary classification. Its results showed average accuracy, sensitivity, specificity, and precision levels. Srinivas et al. (2016) proposed a discriminative feature extraction using deep CNNs.

In this research, we propose a methodology for pathology detection divided into three types of experiments. The first one evaluates the performance of feature descriptors such as SIFT, SURF, and ORB in medical images with machine learning models as an introduction to the last experiment. Then, we propose to use deep learning architectures such as ResNet50, Alexnet, and VGG16. Finally, we propose to use the combination of deep learning and machine learning classifiers. Additionally, we present a novel chest X-ray dataset designed by radiologists called PathX_Chest, which contains 2,200 images divided into ten different classes. This paper is organized as follows: A brief review of related work and the methodology

implemented in the present work are described in Section 2. The results are discussed in Section 3. The conclusions of using CNNs and machine learning models to detect pathologies and the future work to improve the classification task are presented in Section 4.

2. METHODOLOGY

Our methodology is divided into three different approaches (see Figure 1): The first approach is done with machine learning and feature extractors such as SIFT, SURF, and ORB. The second approach is done by applying different CNNs such as AlexNet and VGG16. Finally, the third approach merged machine learning algorithms and deep learning architectures in order to obtain state-of-the-art performance in the classification task.

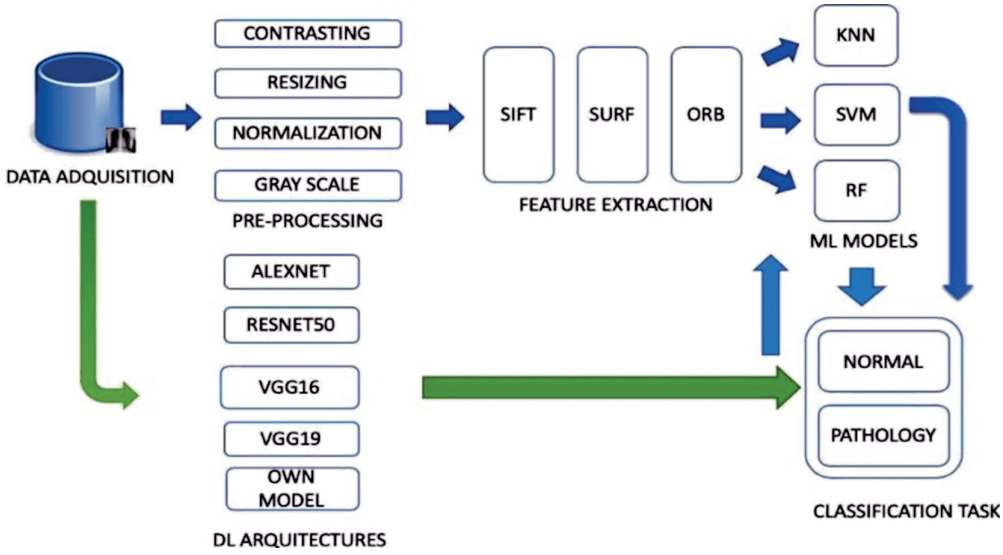

Figure 1. Methodology of three different approaches to detect pathologies
Source: Own elaboration

2.1 Preprocessing

All images were preprocessed with the following techniques: grayscale conversion, contrast enhancement, normalization, and resizing, as we can see in Figure 2. In this stage, a data augmentation technique was applied with filters such as median, mean, and brightness which increased by 0.25: These data were used by CNNs. Additionally, other data augmentation techniques such as vertical and horizontal flip could affect the classification performance. Therefore, the type of data augmentation techniques was determined empirically.

Resizing was set to 700 x 700 based on the average image resolutions.

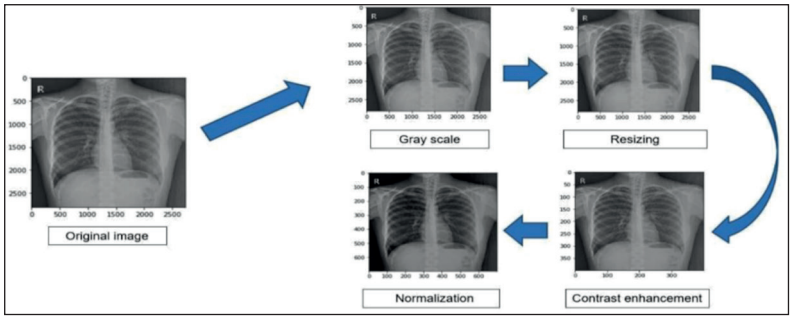

Figure 2. Data preprocessing steps

Own elaboration

2.2 First Approach

Feature extraction

Since the images have been processed and there is already a complete database, they are ready to be used to obtain characteristic vectors, which are the inputs for our classification models. In the present work, the SURF, SIFT, and ORB feature extraction techniques were used. These algorithms extract features of an image. Among its main outputs are the key points in the image: These points are known as descriptors. SURF, SIFT, and ORB provide 128, 64 and 32 descriptors, respectively.

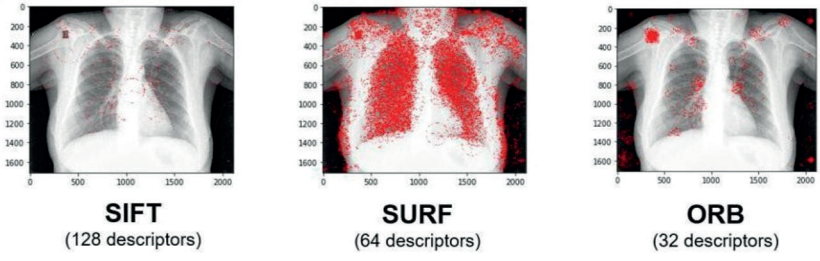

Figure 3. Feature extraction techniques SIFT, SURF, and ORB

Source: Own elaboration

However, these descriptors are transformed into feature vectors using techniques to be inputs for classification models. Therefore, a technique known as bag of visual words (BOVW) was applied. This technique is based on BOW and uses descriptors to represent the images in histograms according to the frequency of its descriptors (Davida, 2018). First, K-means

clustering, where each centroid is used as a vocabulary for the visual dictionary, is applied. Finally, feature vectors are normalized to have entries with the same weight. The K clusters that were used were 10, 50, and 100. KNN has the following hyperparameters to optimize: n_neighbors, weights, and metrics. The n_neighbors parameter can iterate on 3, 5, 11, and 19; the weights parameter can iterate on uniform and distance. Euclidean and Manhattan distances were used. Also, random forest (RF) needs to tune up the following hyperparameters: n_estimators, max_features, max_depth, and criterion. The n_estimators parameter can iterate between 200 and 500; max_features parameter can iterate between auto, sqrt, and log2; the max_depth parameter can iterate between 4, 5, 6, 7, and 8; and the criterion parameter can iterate between Gini and entropy. In this stage, machine learning techniques such as support vector machine (SVM), random forest, and K-NN will be applied. Thus, GridSearch will be applied for finding each algorithm. SVM has hyperparameters to optimize such as kernel, C, and gamma; GridSearch needs a range of values to evaluate the best performance of SVM. The kernel must iterate in linear, RBF, and sigmoid; and C (cost) must iterate between 1, 10, 100, and 1000.

2.3 Second Approach

This approach was built with CNNs such as ResNet50, VGG19, and VGG16 with transfer learning; and LeNet, AlexNet, and an own model without transfer learning. The convolutional networks extract features and, with the sigmoid function in the last layer, obtain the probabilities that allow an image to be classified as normal or pathological. The more layers a CNN has, the more characteristics can be extracted. However, several layers can be misclassifying instead of providing a better performance; the solution is a deeper network with skip connections in order to avoid this problem (Simonyan, Zisserman, 2015).

Transfer learning was used, since the weights of ImageNet were set in all the CNNs, thus achieving better results than the CNNs without pre-trained weights. Subsequently, data augmentation increased substantively the database to train CNN architectures for binary classification. For each image in the database, three more images were created with characteristics such as increased brightness, and filters such as median and average. It is important to remark that the batches, epochs, and learning rates were set empirically.

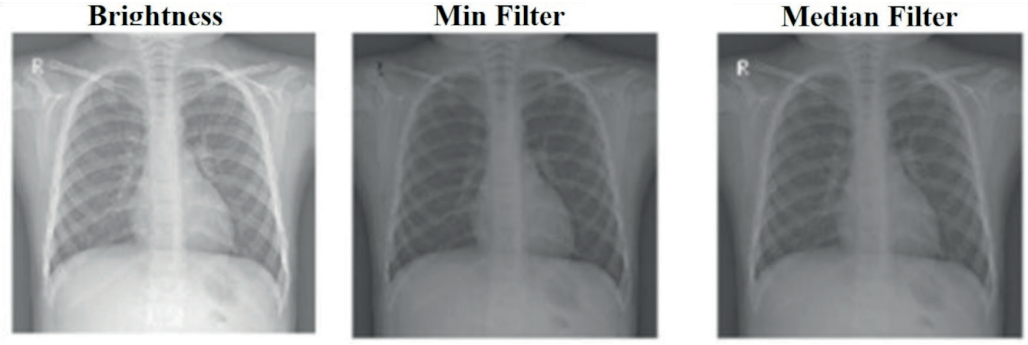

Figure 4. Examples of data augmentation from an original image
Own elaboration

2.4 Third Approach

The feature vectors obtained by the CNN filters (kernels) were extracted to be used as inputs for machine learning models. These CNNs were the same as those used in the second approach. All the features were saved in a CSV file. The number of columns were 10, 50, and 100 (due to the K clusters). The files were used as inputs for the classification models. The feature vectors were split on train, test, and validation set to carry out the training and validation with machine-model learning (SVM, KNN, RF). During the training process, GridSearch was implemented with the same parameters as the first approach.

3. RESULTS

3.1 Database Acquisition

At this stage, the existence of chest X-ray datasets was evaluated. It was concluded that, for the present work, the best option was to create a dataset because there are a lot of misclassified images. The dataset had 2,200 images of normal chest X-rays and 2,200 of pathological chest X-rays, which are detailed in Table 1. Ten (10) pathologies were observed within the image dataset: 220 images were collected per each pathology. These images were provided by the Hospital San José, in Callao. The images were saved in DICOM format with high resolution. Each image was labeled manually by three radiologists. In assessing whether there were different diagnoses, it should be noted that no differences were found in each specialist's diagnosis and that all of them validated the correct labeling of the images.

Table 1
Dataset Distribution

Image Type	Description	No. of Samples	Description	No. of Samples
Pathology	Cardiomegaly	220	Lung consolidation	220
	Emphysema	220	Infiltration	220
	Pleural effusion	220	Fibrosis	220
	Pulmonary nodule	220	Mass	220
	Pneumothorax	220	Edema	220
Normal	Normal cases	2200		

Own elaboration

4. EXPERIMENTAL RESULTS

4.1 First Approach

The results of the three extraction techniques will be presented as descriptors (SIFT, SURF, and ORB), algorithms (SVM, KNN, and random forest) and K clusters (K = 10, 50, 100).

Table 2
First Methodology Results

Classifiers	METRICS	10	50	100	10	50	100	10	50	100
SVM	ACCURACY	0.73	0.79	0.83	0.74	0.8	0.82	0.65	0.71	0.76
	PRECISION	0.76	0.78	0.83	0.77	0.84	0.85	0.69	0.74	0.81
	RECALL	0.71	0.8	0.84	0.71	0.77	0.8	0.59	0.70	0.70
	F1	0.73	0.79	0.83	0.74	0.80	0.82	0.64	0.72	0.79
K-NN	ACCURACY	0.74	0.80	0.81	0.7	0.71	0.77	0.63	0.66	0.68
	PRECISION	0.79	0.83	0.86	0.73	0.76	0.84	0.62	0.65	0.73
	RECALL	0.70	0.76	0.73	0.62	0.66	0.66	0.63	0.65	0.55
	F1	0.73	0.79	0.79	0.67	0.70	0.74	0.62	0.65	0.68
RANDOM FOREST	ACCURACY	0.71	0.74	0.77	0.70	0.72	0.75	0.61	0.63	0.67
	PRECISION	0.72	0.75	0.78	0.70	0.73	0.74	0.62	0.64	0.70
	RECALL	0.70	0.73	0.76	0.69	0.71	0.71	0.60	0.63	0.72
	F1	0.72	0.75	0.78	0.71	0.74	0.75	0.61	0.64	0.68

Own elaboration

Table 2 shows that the SIFT feature extractor has a better performance with respect to machine learning algorithms. SIFT is the feature extractor that gets the most points (128 key points). Furthermore, SVM is the best algorithm in terms of classification of images in normal or pathological scenarios, obtaining the best results in the four metrics (accuracy, sensitivity, precision, and recall) of the three extractors (SIFT, SURF, and ORB). In terms of classification, the best algorithm was SVM with XY accuracy, followed by KNN and random forest.

4.2 Second Approach

The VGG19 model obtained an accuracy of 97 % and a loss of 0.05. When evaluating the other metrics, a precision of 97 %, a recall of 99 %, and an F1-score of 98 % were obtained. The VGG16 model obtained an accuracy of 97 % and a loss of 0.07. When evaluating the other metrics, a precision of 97%, a recall of 98 %, and an F1-score of 97 % were obtained. The ResNet model obtained an accuracy of 97 % and a loss of 0.08. When evaluating the other metrics, a precision of 97 %, a recall of 99%, and an F1-score of 98 % were obtained. The AlexNet model obtained an accuracy of 94% and a loss of 0.18. When evaluating the other metrics, a precision of 95 %, a recall of 90 %, and an F1-score of 92 % were obtained. The LeNet model obtained an accuracy of 71 % and a loss of 0.51. When evaluating the other metrics, a precision of 92 %, a recall of 58 %, and an F1-score of 71 % were obtained. Our own model obtained an accuracy of 94 % and a loss of 0.25. When evaluating the other metrics, a precision of 94 %, a recall of 90 %, and an F1-score of 92 % were obtained.

Table 3
Second Methodology Results

	ACCURACY	PRECISION	RECALL	F1-SCORE	LOSS
VGG19	0.97	0.97	0.99	0.98	0.05
VGG16	0.97	0.97	0.98	0.97	0.07
ALEXNET	0.94	0.95	0.90	0.92	0.18
RESNET 50	0.97	0.97	0.99	0.98	0.08
LENET	0.71	0.92	0.58	0.71	0.51
OWN MODEL	0.94	0.94	0.90	0.92	0.25

Own elaboration

From Table 3, the VGG16, VGG19, and ResNet models obtained the best performance. The AlexNet model got a good performance in terms of accuracy but did not show a good performance when evaluating the loss. The LeNet model got a lower performance because

it had fewer layers than other CNN models. Our own model had a similar performance to AlexNet but it was not the best in general, so the model probably needs to be optimized.

4.3 Third Approach

Table 4
Third Methodology Results

CNN	Classifiers	ACCURACY	PRECISION	RECALL	F1-SCORE
ALEXNET	SVM	0.9893	0.99	0.99	0.99
	KNN	0.9800	0.99	0.98	0.98
	RF	0.9950	0.99	0.98	0.98
VGG16	SVM	0.9954	0.99	0.99	0.99
	KNN	0.9924	0.99	0.99	0.99
	RF	0.9928	0.98	0.99	0.99
VGG19	SVM	0.9999	0.98	0.99	0.98
	KNN	0.9987	0.98	0.99	0.98
	RF	0.9936	0.99	0.98	0.99
RESNET	SVM	0.9998	0.99	0.99	0.99
	KNN	0.9897	0.99	0.98	0.99
	RF	0.9945	0.98	0.99	0.98

Own elaboration

In Table 4, VGG16 + SVM is the model with the highest result. This methodology had results > 90% in the four metrics (accuracy, precision, recall, and F1-score). It can also be concluded that the results were similar, demonstrating that CNN architectures extract feature vectors efficiently. The results shown above allow us to conclude that the best results come from the third methodology, that is, they come from the combinations of CNN architectures and machine learning algorithms (to be more specific, CNN + SVM). In the case of the present research work, it should be considered that the importance of detecting a pathology implies the highest precision, so the most suitable model would be CNN + SVM. Moreover, the results obtained in the third approach could be improved using more images.

4. CONCLUSIONS

There are several factors which help to optimize medical diagnoses. First, there is a shortage of radiologists in various countries of the world, as well as in the Peruvian departments with cold

climates. Second, there is a high incidence of severe respiratory pathologies in Peru. Third, the medical diagnostic process is long and complex. In such a situation, the need for finding a solution to improve the medical diagnostic process arose. Such improved process would allow a prefiltered chest X-ray for doctors to focus only on the diagnosis of radiographs, thus taking advantage of their knowledge and experience to make good diagnoses and cover more medical examinations. Therefore, the present research work aims to develop theoretically and practically a computational vision system design for the prediction of pathologies from chest X-rays to support medical diagnoses. In order to develop this research, we used computer vision techniques for chest X-rays of patients with pathologies and healthy people to extract feature vectors and predict pathologies from chest X-rays.

In conclusion, it is expected to offer an alternative that allows optimizing medical diagnostic services, thus improving medical service in Peru. To obtain better results, it is important to do a proper preprocessing. As a future work, it will apply models of object detection such as YOLO, Fast R-CNN, RetinaNet, among others, for the segmentation of regions and the multiclassification of ten pathologies.

Acknowledgements

We would like to thank radiologists Armando Camasca and Carmen Huamán for their help during the classification task and assessing the veracity of the results.

REFERENCES

Bay, H., Tuytelaars, T & Van Gool, L. (2006). SURF: Speeded Up Robust Features. Retrieved from https://www.vision.ee.ethz.ch/~surf/eccv06.pdf

Davida, B. (2018). Bag of Visual Words in a Nutshell. Towards Data Science. Retrieved from https://towardsdatascience.com/bag-of-visual-words-ina-nutshell- 9ceea97ce0fb

Dong, Y, Pan, Y., Zhang, J., & Xu, W. (2017) Learning to Read Chest X-Ray Images from 16000+ Examples Using CNN. *IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 51-57.

France, K., & Jaya, A. (2019). Classification and retrieval of thoracic diseases using patch-based visual words: A study on chest x-rays. Biomedical Physics & Engineering Express. https://doi.org/10.1088/2057-1976/ab5c7c

Goodfellow, I., Bengio, Y. & Courville, A. (2016). Deep Learning. MIT Press. Retrieved from http://www.deeplearningbook.org

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. Retrieve from https://www.cvfoundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Resid ual_Lear ning_ CVPR_2016_paper.pdf

Imágenes Médicas Diagnósticas. (17 de febrero del 2017). La escasez de radiólogos a nivel mundial. https://www.grupoimd.com.co/blog/escacezradiologos- mundial/ Kreisman

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep

Convolutional Neural Networks. Retrieved from https://papers.nips.cc/paper/4824-imagenet-classification-with-deepconvolutional-neural-networks.pdf

Lowe, D. (2004). Distinctive Image Features from Scale-Invariant Keypoints. Retrieved from: https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf

Rahman, T., Chowdhury, M.E.H., Khandakar, A., et al. (2020) Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray. Retrieved from https://arxiv.org/ftp/arxiv/papers/2004/2004.06578.pdf

Rahmat, T., Ismail, A., & Sharifah, A. (2019). Chest X-ray Image Classification using Faster R-CNN. https://doi.org/10.1016/j.imu.2020.100405.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: an efficient alternative to SIFT or SURF. Retrieved from http://www.willowgarage.com/sites/default/files/orb_final.pdf

Šarić, M., Russo, M., Stella, M. & Sikora, M. (2019). CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images. *2019 4th International Conference on Smart and Sustainable Technologies (SpliTech)*, 1-4.

Simonyan, K. & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. Retrieved from https://arxiv.org/pdf/1409.1556.pdf

Srinivas, M., Debaditya, R., & Krishna M. (2016). Discriminative Feature Extraction from X-Ray Images Using Deep Convolutional Neural Networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 917-921, 10.1109/ICASSP.2016.7471809.