Lung Diseases Classification using X-ray images and machine learning

|  |  |
| --- | --- |
| 1st Ahmad Saigol  *dept. name of organization (of Aff.) name of organization (of Aff.)*  Hamburg, Germany | 2nd Joel Romero  *dept. name of organization (of Aff.) name of organization (of Aff.)*  Hamburg, Germany email address or ORCID |

*Abstract*—COVID-19 continues to be a global health affair. Early detection through CRX Images is an important countermeasure. For this study, feature extraction with traditional machine learning methods, as well as deep learning methods, have been used. The experimental results show that Convolutional Neuronal Networks performed better than Machine Learning models with feature extractors as input.

# I. INTRODUCTION

## A. Background and motivation

On March 2020, the COVID-19 outbreak was declared a pandemic. Ever since, vaccines have been developed, but they have an efficacy of 95% only in terms of preventive action rather than a cure per se. Therefore, early diagnosis of COVID-19 is still important, as it can be traced, and further spread can be avoided. Computed tomography (CT) and chest X-ray (CXR) images are used in the detection of lung infection of any type. Healthcare professionals are focusing efforts on early detection of coronavirus: only in the first 80 days of 2020, about 1245 academic articles were written. Deep Learning is a commonly method for COVID-19 detection using CXR and CT Images. [1]

## B. Problem overview

This study aims to perform COVID-19 detection and differentiate it from other respiratory conditions, by contrasting or combining feature extraction with traditional machine learning methods, and deep learning. Deep learning offers the advantage of feature selection not been hand-picked, in contrast to traditional machine learning methods. A specific CXR database has been provided.

## C. Objectives

Classification of patients with healthy lungs or respiratory diseases from X Ray Images was sought. Two types of Classification were required: Binary and Multiclass Classification. Binary Classifications would determine whether a patient had Covid or not. Multiclass Classification would diagnose 4 different conditions: Healthy (Normal) lungs, Covid-19, Pneumonia and Lung Opacity. To tackle with this classification problem, this study was divided into two phases. During Phase 1, the use of feature extractors and supervised machine learning models was demanded for Classification. During Phase 2, the use of Deep Learning models was required.

# II. Materials and Methods

## A. Dataset Description

To perform this study a dataset of Chest X-Ray images inside different folders was provided [2]. A folder with 16930 X-Ray Images and its labels were provided to train the Machine Learning models. A folder with 4235 images was also provided to test the accuracy of the model. Finally, a folder with noisy dataset 4235 was provided to test the DL models developed in Phase 2.

## B. Data Splitting

The images of the training set were split into training and validation set.

## C. Data Preprocessing

To effectively process the information, different operations where performed to the images.

Normalization: It is commonly applied so feature values can be comparable. Min-max method takes the pixel values of the image to a range from 0 to 1, which is given by (1) [3].

Resizing: Images of the same dataset must be converted to the same size, so machine learning algorithms can handle them in a consistent manner. Neuronal Networks (NN) required a fixed input size which are optimal to the pretrained NN to be used. Making the images smaller may also be used to reduce the computational complexity of the model, and thus reducing the processing time. [4]

Bilateral Filter: Filtering denoises an image and better preserves the details. Bilateral filter performs a pixelwise operation that uses the neighborhood pixels to determine the spatial and intensity distance between them. [5]

Augmentation: The noisy data test provided was mainly composed of rotated X-Ray images (90° and 180°), which represented an additional challenge to solve for the Deep Learning models. To overcome this, Augmentation was applied to the training dataset. It consists of randomly applying different operations to existent images and storing the results. Operations include Translation, Flip, Intensity Changing, Cutout, Scale, Color Space, etc. It should be noted that Augmentation is usually used when there is an insufficient volume of data to train the model. This was not the case for the present study. [6] Online and Offline Augmentation were applied to train the Deep Learning models. Online consisted of applying Augmentation to the input of the model while training. Offline consisted of generating additional images by augmenting the training dataset, preserving the tags from the original images, and adding the results to indicated dataset.

## D. Feature extraction

During Phase 1, Machine Learning models required an input with information summarized from images. Feature extractors were used to obtain this representative information.

Histogram: It is a way to quantify the intensity of pixels and cluster them groups representing ranges, called bins. As a result, a vector of accumulated intensities was obtained. It was used as it is a more quantifiable representation of a images, and further feature extractors can work over it, such as Skewness and Kurtosis.

Skewness: It is a measure of asymmetry of an image’s intensity distribution given the center as a reference point. It can be measured with the Fisher-Pearson coefficient as indicated in (2):

Kurtosis: It is a measure of whether the data is heavy tailed or light-tailed distributed compared to a normal distribution [7]. The formal definition is showed in (3):

Contrast: Refers to the amount of color or grayscale differentiation that exists within an image. It can be calculated

Haralick: As healthcare professionals visually identify different areas and textures within the lungs to assess their condition, feature extractors that can analyze such patterns, provide significant information for the stated problem. Haralick extracts features related to the textures of the images through Co-occurrence matrices. The elements of these matrices represent the probability of going from one pixel with intensityto another pixel of intensity , according to a certain distance and angle of the neighborhood.

Zernike: It captures information related to shape or geometry from an image. Zernike moments are resistant to rotation, non-redundant, and tolerant to noise. These moments are computed from the image’s intensity function projections on the orthogonal base functions. To do so, the image’s center is considered as the center of a unit disk. Each

of the 64 moments are then calculated from the equations (4) and (5) [8]:

## E. Phase 1: ML-based Classification

A first approach for the classification task implied using feature extractors to feed and test different models to achieve the highest performance possible.

Support Vector Machine: SVM creates hyperplanes, which partition the data to classify it, as shown in Fig. X. SVMs were chosen as they allow to model highly complex relationships.



Fig 1 Data classification with hyperplane

Random Forest: Also called Decision Tree Forests, combine bagging (generation of datasets by sampling original training data) and random feature selection to add diversity to the decision tree models. After trees are generated, the model uses a vote to combine trees’ predictions. It was chosen as an alternative model as they can handle large datasets, as only a small portion of the full feature set is used.

Boosting: It boosts performance of weak learners using models trained on resampled data and a weighted vote to determine the final prediction. Contrary to bagging, resampled datasets are constructed specifically to generate complementary learners. A common boosting algorithm is AdaBoost or adaptative boosting, which is tree-based implemented. This model can also effectively handle large datasets, and it was chosen for this reason.

## F. Phase 2: NN-based Classification

Neuronal Networks (NNs) use artificial neurons, shown if Fig. X, as building blocks to model complex data. The neuron is modeled mathematically as indicated by (4):



Fig 2 Typical aritificial neuron

Neuronal Networks can be defined in terms of activation function, network topology and training algorithm, which result in an infinite variety of models that can be trained.

Convolutional Neuronal Networks (CNNs) are NNs that include at least one convolutional layer. A typical CNN consists of the combination of convolutional layers, pooling layers, and dense layers. For the objective of this study, two pre-trained CNNs typically mentioned in the Literature were tested: MobileNet V2 showed in Fig. X [9] and Resnet 50 showed in Fig. X [10].

A picture containing funnel chart

Description automatically generated

Table

Description automatically generated

Fig 3 MobileNet V2’s structure

Table

Description automatically generated

Fig 4 Resnet50’s structure

## G. Metrics

To measure the performance of different preprocessing techniques, data and models, some common metrics where used.

Confusion matrix: It is a table that categorizes predictions according to whether they actually belong or not to a class. The size of the matrix depends on the number of classes. The class of interest is known as positive class, and the others are negative. The matrix will be composed of the following categories, as shown in Fig. X for a two classes example:

* True Positive (TP): Correctly classified as class of interest
* True Negative (TN): Correctly classified as not class of interest
* False Positive (FP): Incorrectly classified as class of interest
* False Negative (FN): Incorrectly classified as not class of interest



Fig 5 Two-class confusion matrix

The importance of this matrix falls in all the performance metrics derived from it.

Accuracy: Also called success rate, It is defined by (5):

Precision: Also known as positive predictive value, measures the proportion of positive examples that are truly positive as defined by (6):

Sensitivity: Also called true positive rate, measures the proportion of positive examples correctly classified, as defined by (7):

Specificity: Also called true negative rate, measures the proportion of negative examples correctly classified, as defined by (8):

Balanced accuracy:

F measure: Also called F1-score, describes de model performance, by combining recall and precision, as defined in (10):

Matthews Correlation Coefficient: It is a statistical tool for model evaluation, is defined by (11):

It is worth to mention that this indicator, even though it is widely used, it does not perform well in imbalanced datasets [11].

## H. Loss function

A loss function measures the difference between predicted and true probability distributions. Weighted cross entropy loss is a variation of cross entropy loss. It is especially useful for class imbalance, as it assigns different weights to different classes, as shown in (12) [12]:

## I. Optimizer

An optimizer is required to update the weights of a neuronal network. Adam optimizer avoids oscillations and convergence problems during training. It can be tuned with various parameters, as shown in Fig. X [13].

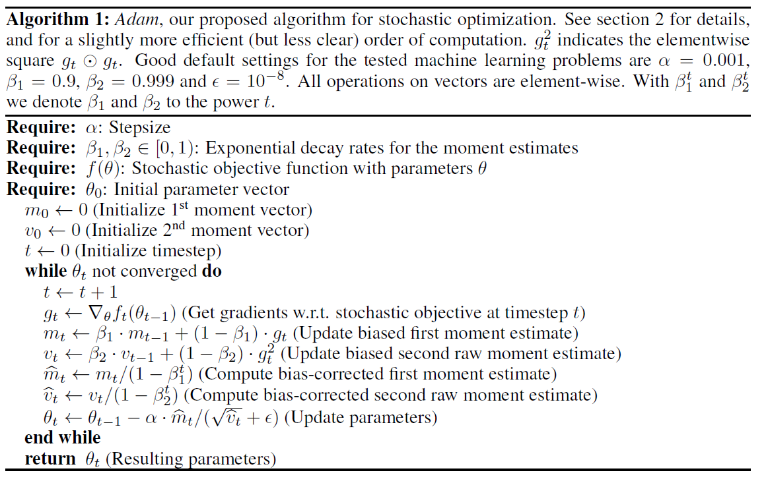
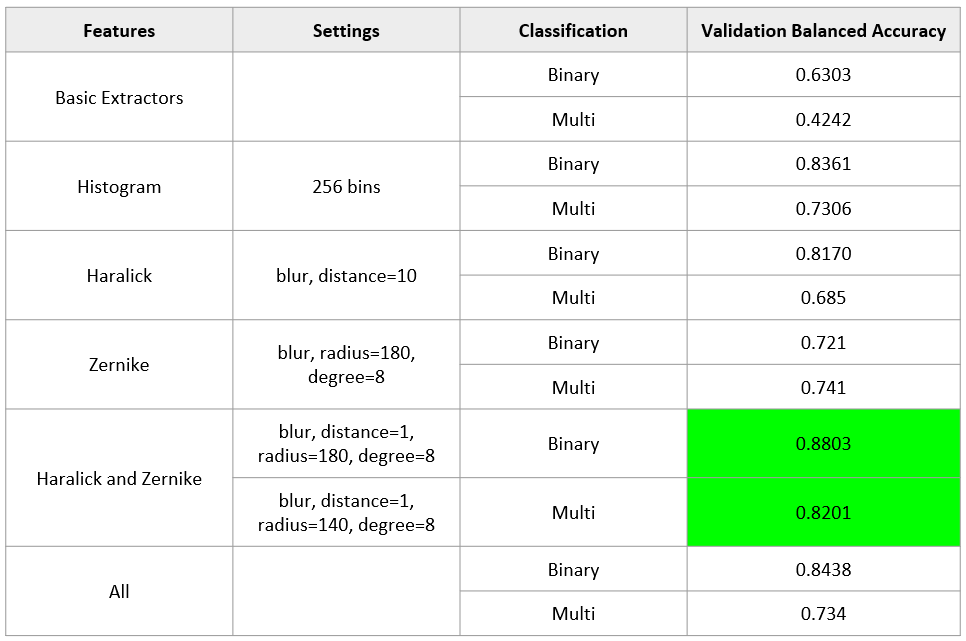
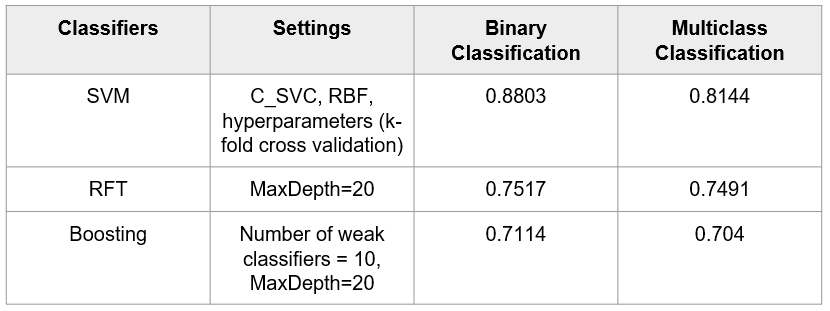


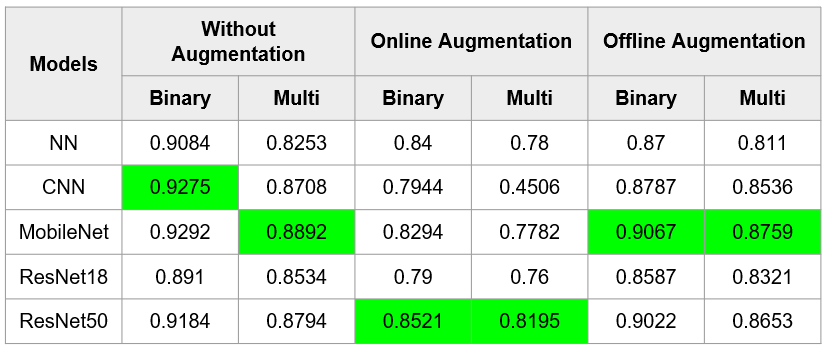
Fig 6 Adam algorithm

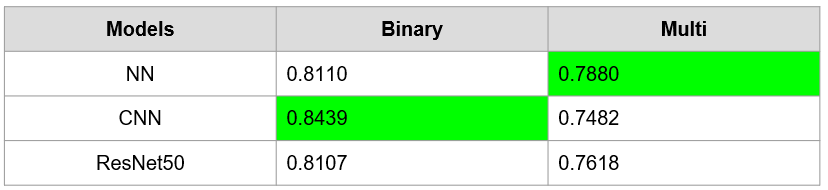
# III. Results and Discussion

Main inconvenient was computational complexity of the models. Deeper analysis would require more time to train models. Optimal parameters for feature extractors such as Haralick and Zernike require further analyses. Based on the literature review, additional steps such as lung segmentation or combination of pretrained NNs would allow more precise results.









# IV. Conclusion

Haralick and Zernike were the feature extractors that performed the best.

Randomly generated CNN performed the best.

References

[1] e. a. Nandhini Subramanian, “A review of deep learning-based detection methods for COVID-19,” 2022.

[2] ISM, *Chest X-Ray images dataset.* [Online]. Available: http://​gofile.me​/​5AIED/​xNqNoKPxu

[3] Brett Lantz, *Machine Learning with R,* 2nd ed. Packt Publishing, 2015.

[4] Emre AVUÇLU, “COVID-19 detection using X-ray images and statistical measurements,” 2022.

[5] Sidheswar Routraya, Arun Kumar Raya, Chandrabhanu Mishrab, “Image denoising by preserving geometric components based on weighted bilateral filter and curvelet transform,” 2018.

[6] Mingle Xu, Sook Yoon, et al., “A Comprehensive Survey of Image Augmentation Techniques for Deep Learning,” 2023.

[7] NIST/SEMATECH, *Handbook of Statistical Methods*. NIST/SEMATECH. [Online]. Available: https://​www.itl.nist.gov​/​div898/​handbook/​index.htm

[8] The Florida State University, *Contrast Manipulation in Digital Images.* [Online]. Available: https://​micro.magnet.fsu.edu​/​primer/​java/​digitalimaging/​processing/​contrast

[9] Pytorch Team, *Mobilenet V2.* [Online]. Available: https://​pytorch.org​/​hub/​pytorch\_​vision\_​mobilenet\_​v2/​

[10] Pytorch Team, *ResNet.* [Online]. Available: https://​pytorch.org​/​hub/​pytorch\_​vision\_​resnet/​

[11] Qiuming Zhu, “On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset,” 2020.

[12] e. a. Sheng Lu, “Dynamic Weighted Cross Entropy for Semantic Segmentation with Extremely Imbalanced Data,” 2019.

[13] Diederik P. Kingma, Jimmy Lei Ba, “Adam: A method for Stochastic Optimization,” 2015.