UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Analyzing the effect of artificial data used in training CNN for COVID-19 detection

Ivan Mateo Hidalgo Davila

Juan José Murillo Celi

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO INDUSTRIAL

Quito, 20 de diciembre de 2021

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Analyzing the effect of artificial data used in training CNN for COVID-19 detection

Ivan Mateo Hidalgo Davila

Juan José Murillo Celi

Nombre del profesor, Título académico **Maria Gabriela Baldeón Calisto, PHD**

Quito, 20 de diciembre de 2021

© DERECHOS DE AUTOR

Nombres y apellidos: Juan José Murillo Celi

Código: 00200323

Cédula de identidad: 1724061740

Lugar y fecha: Quito, 20 de diciembre de 2021

Nombres y apellidos: Ivan Mateo Hidalgo Davila

Código: 00139389

Cédula de identidad: 1727276295

Lugar y fecha: Quito, 20 de diciembre de 2021

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La detección de enfermades mediante imágenes de radiografias pulmonares como el COVID-19 se ha vuelto una de las principales herramientas de apoyo para tener un segundo diagnóstico en la comunidad médica. En este trabajo se hace uso de Redes Neuronales Convolucionales (CNN) para la clasificación de imágenes de radiografias pulmonares CXR en las categorías COVID-19, neumonia y normal. Debido a la escaces de imágenes para poder entrenar la red neuronal, se utilizo una técnica llamada Data Augmentation, que es empleada en Deep Learning para aumentar el tamaño del set de entrenamiento y aumentar la precisión de la base de datos. Data Augmentation incluye una variedad de ténicas que pueden ser modificadas dependiendo de las necesidades del experimento. El objetivo de este estudio es hallar las técnicas de Data Augmentation que influencian en la precisión de la red neuronal, mediante el uso de Diseño de Experimentos. Se logró demostrar que las técnicas de Data Augmentation que si afectan a la precisión de la red neuronal son cambio de altura y zoom. Estos resultados permiten evidenciar que los parámetros ingresados para generar imágenes artificales deben ser previamente analizados para garantizar una mejor precisión de las redes neuronales que se emplean para la detección de enfermedades mediante imágenes de radiografías pulmonares. Se establece como perspectivas futuras utilizar nuevas técnicas de Diseño de Experimentos que permitan encontrar cual es el valor numérico que maximice la precisión de las Redes Neuronales Convolucionales.

Palabras clave: Clasificación de imágenes médicas, COVID-19, Redes Neuronales Convolucionales, Imágenes de rayos X, Image Data Augmentation, Diseño de Experimentos

ABSTRACT

The detection of diseases through lung X-rays images such as COVID-19 has become one of the main support tools to have a second diagnosis in the medical community. In this work, Convolutional Neural Networks (CNN) are used to classify images of CXR lung radiographs into the COVID-19, pneumonia and normal categories. Due to the scarcity of images to train the neural network, a technique called Data Augmentation was used, which is applied in Deep Learning to increase the size of the training set and increase the precision of the database. The data augmentation includes a variety of techniques that can be modified according to the needs of the experiment. The objective of this study is to find the techniques of data augmentation that influence the precision of the neural network, through the use of Design of Experiments. It was possible to show that the techniques of data augmentation that affect the precision of the neural network are height shift and zoom. These results make it possible to show that the parameters entered to generate artificial images must be previously analyzed to guarantee better precision by means of the neural networks that are used for the detection of diseases with lung radiographs images. Future perspectives are established to use new techniques of Design of Experiments that will be able to find which is the numerical value that maximizes the precision of the Convolutional Neural Networks.

Key words: Medical Image Classification, COVID-19, Convolutional Neural Network, X-ray images, Image Data Augmentation, Design of Experiments

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

The novel Coronavirus disease (COVID-19) first appeared in Wuhan, China, in January 2020 [1]. The rapid spread of this virus, and the thousands of deaths it caused, led the World Health Organization to declare it a global pandemic on March 12, 2020 [2]. The need to develop new methods to detect the SARS-CoV-2 became a challenge for the scientific community. From nucleic acid amplification tests to antibody detection assays, the first methods appeared to be effective in the rapid identification of COVID-19 patients [3]. Nevertheless, these methods came with its own limitations like low detection sensitivity, long detection times, frequent false negative nucleic acid results and the need to be performed by professional technicians [3].

Viral pneumonia is one of the main causes of death of COVID-19. This led to the use of an alternative screening technique that allows to detection of the SARS-CoV-2 virus by analyzing the chest radiography (X-ray) of a possibly infected individual [4]. Various international guidelines recommend chest X-rays to detect pneumonia because of the worldwide availability, affordability, and fast obtention. However, the interpretation of the image can be very difficult and many times inconsistent, having a high inter- and intra-observer variability between practitioners. Thus, requiring of a radiologist with expert knowledge and experience to give the diagnosis [5]. In light of these complications, computer-aided diagnosis (CAD) has become a popular tool to aid doctors in the detection and differential diagnosis of abnormalities in medical images, including X-rays. CAD uses computerized analysis and algorithms to automatically provide a diagnosis, serve as a second opinion to physicians, and eliminate inter-observer variation [6].

Deep Neural Networks are considered one of the most powerful tools in machine learning when handling huge amounts of data. In the past decade, deep neural networks have been widely applied in CAD and medical decision support systems [7]. Moreover, in the field of medical

image analysis, Convolutional Neural Network (CNN) have gained a lot of popularity due to their excellent performance and state of the art results in various computer vision tasks [8]. CNNs have the advantage of automatically learning a set of feature detectors from labelled data and requiring minimal pre-processing operations [9]. Considering that an effective and rapid screening of infected patients is a pivotal step in winning the fight against the Covid-19 pandemic, various researchers have proposed CNNs tailored for covid-19 chest X-ray classification. These CNNs receive as input the chest X-ray image of a patient and predict whether the person is or not infected with the SARS-CoV-2 virus. Wang et. al developed the COVID-Net [10], a CNN architecture that introduced a lightweight projection-expansion-projection-extension design that enables an enhanced representation capacity. The model was tested in the open access COVIDx dataset and achieved a 93% accuracy on the test dataset. In [11], Mahmoud et al. developed the CovidXrayNet , a CNN based on the EfficientNet-B0 with low cycle process need, and optimized the hyperparameters and data augmentation strategy for the Covid-19 detection task. CovidXrayNet obtained a 95.82% accuracy on the COVIDx dataset with only 30 training epochs. A study developed by El-Shafai et al. [12] focused on finding the best activation function and optimizer to build a model to classify COVID-19 from CXR and CT images. The experiments showed that a combination of the Stochastic Gradient Descent with momentum and ReLU activation functions give the best accuracy. Weiss et al. [13] proposed a hyperparameter optimization method, where three different CNN architectures were tested by varying the hyperparameter values of the learning rate, batch size and number of epochs. The result showed that the Xception architecture [14] had an optimal performance with an epoch size of 40, learning rate of 0.000005, and batch size of 32. The study on which this work is based, CoroNet [15], achieved an overall accuracy of 89.6%, taking into consideration that this CNN is computationally less expensive and achieved promising results.

CCNs usually have tens of millions of parameters and require a considerable amount of data to avoid overfitting the training set. Overfitting happens when a neural network perfectly models the training set and suffers of high prediction variance [10]. Hence, the model is good at predicting known data but has a bad performance predicting unknown data. One of the most used techniques to avoid overfitting and improve the generalization of a model is data augmentation. Data augmentation increases the size and diversity of the training set by modifying the appearance of the original images or generating artificial ones. This technique is especially important in fields where large datasets are not available, such as in medical imaging where acquiring well-annotated data can be time-consuming, very costly, and in some pathologies even impossible. Data Augmentation can be divided into three different techniques: basic augmentation techniques, deformable augmentation techniques, and deep learning augmentation techniques [11]. Basic augmentation techniques apply a transformation that maps points of the image to different positions and also the manipulate the intensity of the pixel values. These techniques are generally simple and fast to implement, which is why researchers frequently apply them when training deep learning models [16]. On the other hand, deformable augmentation techniques are implemented when basic augmentation techniques do not provide sufficient variability to make the generalizable. The scale of deformation is defined by the user to ensure that the result is clinically plausible. Finally, in deep learning techniques networks automatically learn representations of images and artificially generate realistic images.

Given the scarcity of publicly available Covid-19 datasets, most of the CNNs developed for Covid-19 detection on chest X-rays apply various data augmentation techniques to improve the classification accuracy. Basic augmentation techniques were the most commonly reported operations. Nevertheless, almost none of the models analyze how the data augmentation operations affected the model prediction, and if they were needed at all. In [17] , Elgendi et al.

empirically examined the effectiveness of 4 geometric data augmentation strategies in Covid-19 detection and concluded that these strategies significantly decrease the accuracy. Nevertheless, the authors did not study how each individual data augmentation operation affected the detection power of the model. Considering that data augmentation can be a very powerful method to improve the generalization and robustness of a model, it is important to have a better understanding the impact each operation has when training a CNN.

In this work, we will test the effect basic augmentation techniques have in the classification of Covid-19 chest X-Rays using experimental design. For this study, the following basic augmentation techniques will be tested out to run the model: vertical flip, horizontal flip, rotation, zoom, width shift, height shift and shear. In this study, CoroNet CNN model developed by Khan et al. [13] will be used to test out different approaches related to the characteristics that the model receives as input and extract different accuracies [13]. Since the main objective of this study is to find which are the most important Data Augmentation techniques that allow to attain a better accuracy, we use design of experiments as the main tool to determine which factors and interactions of the basic augmentation techniques are significant in the experiment.

DESARROLLO DEL TEMA

1. Related work

In this section, we provide a brief review about data augmentation techniques for covid-19 classification, and experimental design.

1.1. Data Augmentation Techniques for Covid-19 Classification

Data augmentation is a method that artificially inflates the training set, which is used when training CNNs to prevent over-fitting and improve performance [10]. In medical image analize this technique is crucial because the availability of images, such as CXR or CT, is scarce due to the acquisition cost, cost of obtaining a diagnosis by professionals, patient confidentiality, among others. Surveys that present methods on image data augmentation for deep learning are presented in Connor et al. [18] work, where evidence is presented that shows that data augmentation in Deep Learning helps constructing better datatasets, as it prevents overfitting by modifying limited datasets to possess special factors that describe big data. Other works focus on the use of other techniques enabled by CycleGAN to increase the accuracy of classification tasks. This study is based in the combination of data augmentation techniques aiming to increase the quality of the dataset [19]. More specifically in [11], Chlap et al. provides a review centered on data augmentation techniques employed in state-of-the-art deep learning models for medical image data. This study divides data augmentation into three categories, basic augmentation techniques, deformable augmentation techniques, and deep learning augmentation techniques. The basic data augmentation techniques can be defined as techniques that apply simple transformation to the image. Deformable augmentation techniques defines scales of deformation stablished by an user to ensure result is clinically plausible. Deep learning techniques are based in automatically learn representations of the images given. Basic augmentation techniques are the mostly used type of operations, and they

include geometric transformation, cropping, occlusion, intensity operations, noise injection, filtering, and combination. Although data augmentation has shown to improve the test set accuracy in some tasks, other works have shown that the use of certain data augmentation techniques might negatively affect the model´s performance [17, 18].

In the previous year, various works developed CNN architectures for the automatic diagnosis of COVID-19 from chest X-ray images. In **Table 1**, we present a comparative study of highly cited studies in this manner, including the name of the neural network, type of data augmentation operations applied, size of the training set, and accuracy achieved. Since the aim of this study is analyzing how basic data augmentations affect the accuracy of a the model, if mentioned in the manuscript, we include the ranges in which each data augmentation operation was applied. The most repeated Data Augmentation types found in related works were chosen to be analyzed in this study which are vertical flip, horizontal flip, rotation, zoom, width shift, height shift and shear.

Table 1: Data Augmentation Techniques for COVID-19 detection in Chest X-ray images

Title	Neural Network Architecture	Data Augmentation Types	Training set size per class	Accuracy achieved
A Modified Deep Convolutional Neural Network For Detecting Covid-19 And Pneumonia From Chest X-Ray Images Based On The Concatenation Of Xception And Resnet50v2	Xception and ResNet50V2 concatenated	Rotation Re-scale	Covid 149 Normal 1634 Pneumonia 2000 Total: 3783	Xception: 91.31 ResNet50V2: 89.79 Concatenated: 91.40
Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network	EfficientNetB	Re-scale	Covid 404 Normal 404 Pneumonia 404 Total: 1212	Binary classification: 99.62 Multi-class classification: 96.7

Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network	AlexNet VGG19 ResNet GoogleNet SqueezeNet	Flipping: up/down, right/left Translation Rotation using random five different angles	Covid 105 Sars 11 Normal 80 Total: 196	AlexNet 89.1 VGG19 93.1 ResNet 93.1 GoogleNet 89.65 SqueezeNet 82.75
OptCoNet: an optimized convolutional neural network for an automatic diagnosis of COVID-19	OptCoNet GWO-based CNN		Covid 900 Normal 900 Pneumonia 900 Total: 2700	97.78
SARS-Net: COVID-19 detection from chest x-rays by combining graph convolutional network and convolutional neural network	SARS-Net	Elastic Translation Rotation Random horizontal flip Zoom MotionBlur Intensity shift	Covid 258 Normal 7966 Pneumonia 5451 Total: 13675	97.6
CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images	CoroNet	Re-scale Rotation Width shift Height shift Shear Zoom	Covid 284 Normal 310 Pneumonia Bacteria 330 Pneumonia Viral 327 Total: 1251	89.6
PDCOVIDNet: a parallel-dilated convolutional neural network architecture for detecting COVID-19 from chest X-ray images	PDCOVIDNet	Rotation: 30 Height shift: 0.15 Width shift: 0.15 Shear: range 0.10 Zoom: range 0.10	Covid 175 Normal 1072 Pneumonia viral 1076 Total: 2323	96.58

2.2 Experimental Design

Experimental design is one of the ways by which we learn about how systems or process work. It is used as a tool for engineers and scientists to use for product design and development, which can be related to this work as one of the main objectives is to find the Data Augmentation

techniques that allows the neural network to reach a better accuracy [20]. As mentioned in [21], Lujan-Moreno et. al the experimenter is allowed to choose any suitable screening method based upon the number of available runs, the desired precision of the results, and reuse of previous runs. In this experiment, since some factors like time and precision desired play a crucial importance, it was decided to use a fractional factorial design.

During this investigation a 2^K factorial designs is implemented. These designs have factors which are the parameters that will be varied and levels which are the different values that the parameters or factors will have. In this factorial design used there will be several factors and only two different levels for each factor. These factorial designs are used in the early stages of research when there are several factors with two levels each and it is desired to know which ones positively affect the response variable [20]. In these factorial models there is the possibility of reducing the runs. These runs refer to the possible combinations that each instance can have by varying the different factors at different levels. Decreasing these runs will create a fractional model which helps to have conclusions similar to a full factorial design, which is defined as an experiment design where all combinations of the parameter levels are tested in order to analize the results [22]. Fractional models are used when you do not have enough resources to complete a full model. The disadvantage of using fractional models is that allied terms can be created, that is, certain combinations of factors are combined with main factors. For this reason, it is sought to have a fractional model with a high resolution which will try to eliminate these allied terms and have more concrete conclusions.

Previous works have used design of experiments to determine the optimal hyperparameters in machine learning algorithms. In [21], Lujan-Moreno et. al proposed to use the design of experiments methodology as the first step to screen for the most significant hyperparameters and a Response Surface Methodology to further tune their value. Staelin et al. [23]developed a

design of experiments inspired algorithm that iteratively refines the boundaries and resolution of a search grid. The algorithm was tested on the hyperparameter optimization of a least squares SVM regression. A study developed by F.Chou et al. [24] focused in finding the combination of hyperparameters that enhances the performance of a CNN for image identification. They introduce the concept of uniform experimental design (UED), which is defined as a space filling design that can be used when the underlying model is unknown [25].

2. Methodology

In this paper, we analyze how basic data augmentation techniques affect the classification accuracy of CNNs on Covid-19 classification from chest X-ray images. These experiments are performed on 2 publicly available Covid-19 dataset of different sizes. Moreover, the CoroNet architecture [15] has been selected because its code is open source, has a good prediction performance, and makes an efficient use of the model parameters. Finally, to have statistically significant results of the effect of each data augmentation operation, a 2^K factorial experimental design is implemented. In this section we describe the Covid-19 datasets used, review the CoroNet architecture, explain the data augmentation factors analyzed, and finally present the 2^k factorial design.

a. Description of datasets

We select two publicly available Covid-19 chest X-rays datasets for the present study. The first dataset is presented by Khan et. al [15] and is a recollection of images from Github repository by Joseph et. al [26]. This repository is a collection of images from the RSNA, Radiopedia, and Kaggle database [27]. The database has images from 4 different classes, namely Covid-19 positive, normal, bacterial pneumonia, and viral pneumonia. To get a balanced dataset, the authors randomly eliminated images from classes with a higher frequency of observations. For

this study, the dataset was modified to have only 3 classes (covid-19 positive, normal and pneumonia) by combining the observations from bacterial and viral pneumonia. The dataset contains 1678 images, from which 15% are used for testing, 17% for validation, and 68% for training. The second dataset was made available by Khuzani et. al [28] and has images from 3 different classes that correspond to covid-19 positive, normal, and pneumonia. The dataset is already balanced and contains 381 images. 10% of the images are used for testing, 18% for validation and 72% for training. A summary of number of images per class and between set is presented in Table 2. Furthermore, in **Figure 1** we present examples of images from each dataset and class. Images of both datasets were preprocessed by recalling with a factor of 1/255, applying ZCA whitening, dividing inputs by the standard deviation of the dataset and setting input mean to 0 over the dataset [29].

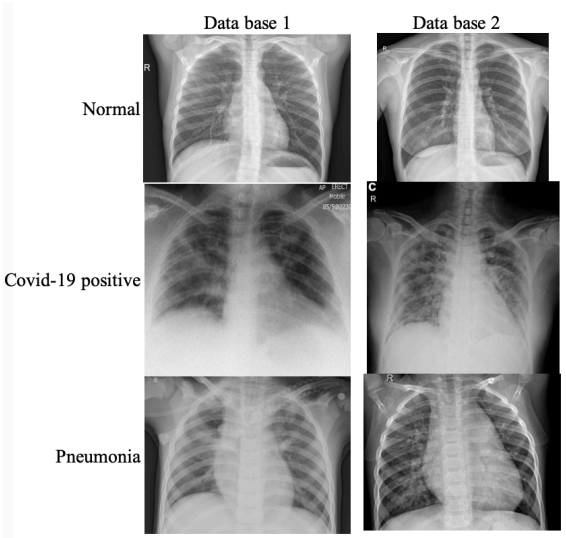

Figure 1. Images from each dataset and class

Table 2. Summary of sizes of Data Bases

Repository	Classes	Test Size	Train and Validation size
Khan et. al	Covid-19	20	300
	Normal	145	768
	Pneumonia	77	368

Khuzani et. al	Covid-19	14	126
	Normal	10	90
	Pneumonia	15	126

b. Model architecture

The CoroNet architecture [15] was selected for this study, which is a CNN architecture specifically designed for the detection of Covid-19 from chest X-ray images. The CoroNet is based on the Xception architecture [14], which is an extreme version of the Inception model. The CoroNet borrows the base Xception architecture and adds a dropout layer and two fully-connected layers at the end. Also, it has optional batch normalization layer to improve the training time. The model has a total of 33,969,964 parameters, from which 33,915,436 are trainable and 54,528 that are non-trainable. [15]. The architectural details of the CoroNet are presented in **Table 3**.

The code for the CoroNet is open source and is available on the Github repository of the author [15]. The architecture is fairly simple in comparison with competitor CNNs proposed for the same task, which will allow to better evidence the impact the different data augmentation operations have on the classification accuracy. Furthermore, the CoroNet obtained a superior performance in comparison to other studies in literature in the task of multi-class classification.

Table 3. CoroNet architecture details [15]

Layer type	Number of Parameters
Xception	20861480
Flatten	0
Dropout	0
dense (Dense)	13107456

dense_1 (Dense)	1028

c. Image Data Augmentation

The most commonly used basic data augmentation techniques found in literature (refer to Table 1) for the detection of Covid-19 were selected for testing in this study. The operations considered with their corresponding definition are presented in **Table 4,** while in **Figure 2** we show how each operations affects the appearance of the original images. Moreover, in **Table 5** the values for the data augmentation operations are shown. The values have also been set based on the literature review. To implement an experimental design, two parameters need to be set for each augmentation operation. In the experimental design terminology these parameters are known as a low level and high level of a factor. For an specific run, either the high or low level will be selected to train the CNN. The data augmentation is implemented on the fly with the Data Image Generator tool from the Tesorflow library.

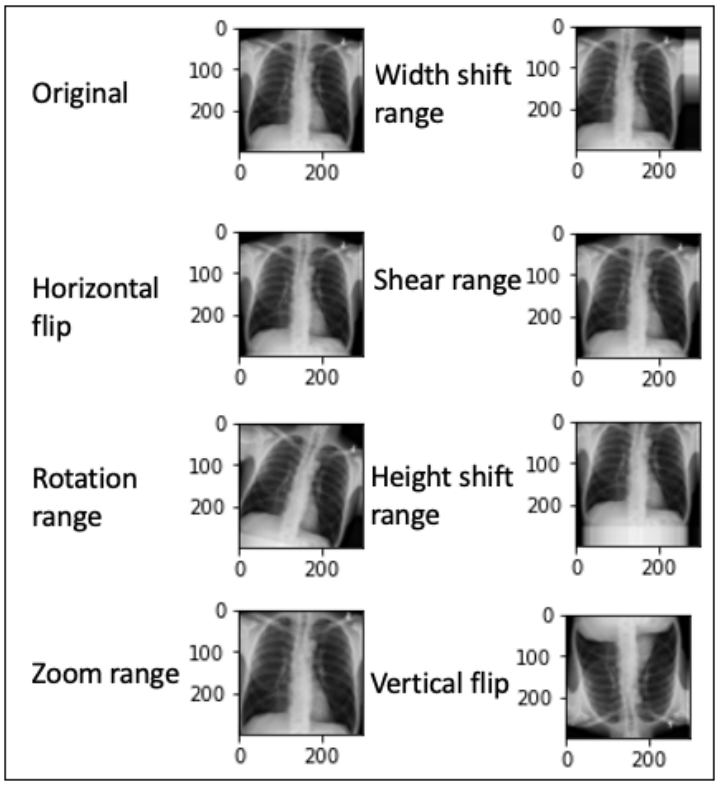

Figure 2. Appearance of images altered by data augmentation

Table 4. Definition of Data Augmentation operations [29]

Data Augmentation Operation	Definition
Vertical_flip	It receives a boolean input data and performs a vertical random flip.
Horizontal_flip	It receives a boolean input data and performs a horizontal random flip.
Rotation_range	It receives an input angle in the form of an integer value and performs a random rotation between zero and the given angle.
Zoom_range	It receives a float value and performs a random zoom between zero and the given value.
Width_shift_range	It receives a float value and represent the fraction of total width.
Shear_range	It receives an input angle in the form of a float value and represents shear angle in counter-clockwise direction in degrees.
Height_shift_range	It receives a float value and represent the fraction of total height.

Table 5. Data Augmentation Operations and levels for the factorial design.

Data Augmentation Operation	Low Level	High Level
A: Vertical_flip	FALSE	TRUE
B: Horizontal_flip	FALSE	TRUE
C: Rotation_range	0	15
D: Zoom_range	0	0.15
E: Width_shift_range	0	0.15
F: Shear_range	0	0.25

G: Height_shift_range	0	0.15

d. Design of Experiments

The aim of this work is to gain insights into the effects the data augmentation operations have on the classification accuracy of a CNN for Covid-19 detection. A factorial experimental design is a technique commonly used to determine how each factor from a set of factors and its combinations affect a response variable. In this work the factors we want to analyze are the data augmentation operations presented in **Table 4**, and the response variable is the classification accuracy in the test set. **Table 5** shows the levels of the factors within our experimental model. As previously mentioned, the factors and corresponding levels were set based on a literature review of the most commonly used data augmentation operations on CNNs for Covid-19 detection. The workflow used to implement the experimental design is composed of 6 steps as shown in **Figure 3.** The steps are explained next.

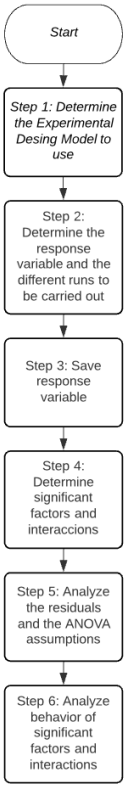

Figure 3. Workflow of the Design of Experiments

Step 1: The resulting experimental model is a 2^7, which leads to 128 different combinations of data augmentation parameters. Hence, due to the limitation of resources and time, a fractional factorial model is implemented. This new model will be a factorial $2^{(7-1)}$, which results in 64 different combinations. This model has a VII resolution, which means that it does not have aliased terms. Allied terms refer to factors that within the model are confused with other interactions, causing a term to represent a factor and an interaction.

Step 2: For each of the 64 combinations obtained in Step 1, a CNN is trained with the specific data augmentation parameters and the classification accuracy on the test set, as displayed in equation 1, measured as the response variable. The models are trained using the Adam optimizer with a learning rate of 0.0001, batch size of 25 and 300 training epochs. The models are implemented in Keras with tensorflow 2.0 backend. Furthermore, the experiments are carried out using an Nvidia Docker VM, on a Linux Ubuntu 18.04 platform equipped with an Nvidia Tesla V100 graphics card.

$$Accuracy = \frac{No.of\ images\ \ correctly\ cassified}{Total\ no.\ of\ images} \quad (1)$$

Step 3: The classification accuracy on the test set of the classification model of each of the Design of Experiments runs were saved and transferred to the Design Expert software in order to perform an accuracy analysis.

Step 4: The significant factors and interactions were chosen based on a half-normal plot and a pareto graph, where it is clearly seen which terms cause a significant change on the response variable.

Step 5: The graphs of the residuals were analyzed to verify that all the ANOVA assumptions (normality, independence and homogeneity) are fulfilled. If the assumptions are not met, data transformations must be carried out until the assumptions are met.

Step 6: The ANOVA table was analyzed to determine which terms are significant and if the mathematical model that tries to predict the response variable using these terms is significant. Subsequently, the cube plot was analyzed to understand the behavior of the interactions in relation to the response variable.

3. Results and discussion

In this section we present the results of the experimental design on the two datasets and the comparison of the model with the "optimal" data augmentation technique against other models in literature.

a. Classification accuracy on Dataset 1

In **Figure 4** a pareto chart with the results of the experimental design is presented. In this graph, the y-axis shows the t-value and the x-axis shows the rank of the terms ordered from the term with the highest t-value to the lowest. This t-value refers to the influence that each term has on the response variable. There is a limit marked with a black line that represents the level of significance (alpha). If the t-value of a term exceeds this limit, it means that it is a statistically significant term, and it is graphically evidenced by seeing that the bar of the term exceeds the limit marked with the black line. The data augmentation operations zoom_range (factor D), height_shift_range (factor G) and the third level interaction ABF (combination of factors A (vertical_flip), B (horizontal_flip) and F (shear_range)) have a statistically significant influence over the classification accuracy. Due to the hierarchy assumptions, the factor F (shear_range) and the second-level interaction BF must be included even though these terms are not significant but improve the mathematical prediction of the design of experiments.

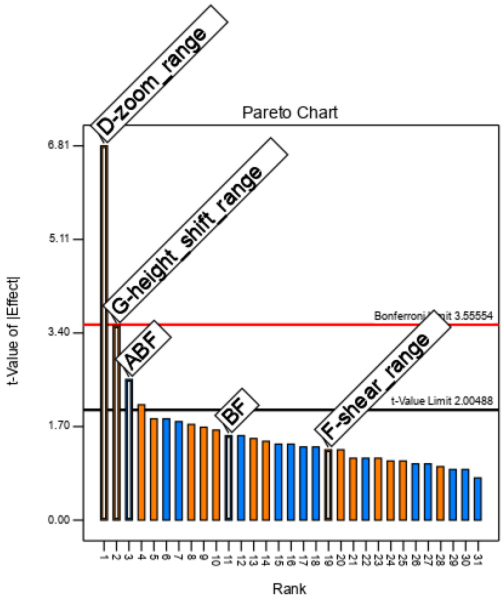

Figure 4. Pareto Chart for Dataset 1

The significant factors and interactions with the suggested hierarchy terms are selected. With this terms an ANOVA is performed and analized. The results indicate that the data augmentation operations zoom_range (factor D) and height_shift_range (factor G), and the ABF interaction are significant and positively affect the accuracy of the model. To analyze the interaction, we proceeded to obtain a cube analysis of the accuracy of the model, and it is shown in the **Figure 5**.

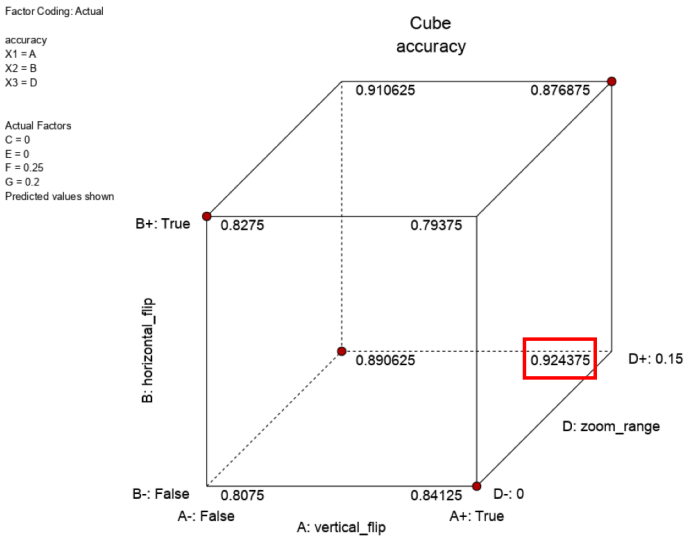

Figure 5. Cube analysis of Accuracy of Data Set 1

By analyzing all the vertices, which represent all the possible combinations for the ABF interaction at their high and low levels, the optimal combination of data augmentation operations can be found. The optimal point is framed in red on the **Figure 5** and its optimal data augmentation operations configuration is when vertical_flip (factor A) is true, horizontal_flip (factor B) is false, zoom_range (factor D) at its high level, shear_range (factor F) at its high level and height_shift_range (factor G) at its high level. All other operations are irrelevant and do not cause a significant change in the accuracy of the model.

b. Classification accuracy on Dataset 2

As established in step 5, the ANOVA analysis is correct if all the assumptions about the residuals are satisfied. In this database, the assumptions were not met, so a data transformation was implemented. This data transformation was a ArcSin of the square root of the response variable. The pareto chart for this dataset is presented in **Figure 6**, which show that the height_shift_range (factor G) and the third level interactions BFG, ADG and BDG are significant. Due to hierarchy assumptions, more factors must be included in the model even though these terms are not significant but improve the mathematical prediction of the design of experiments.

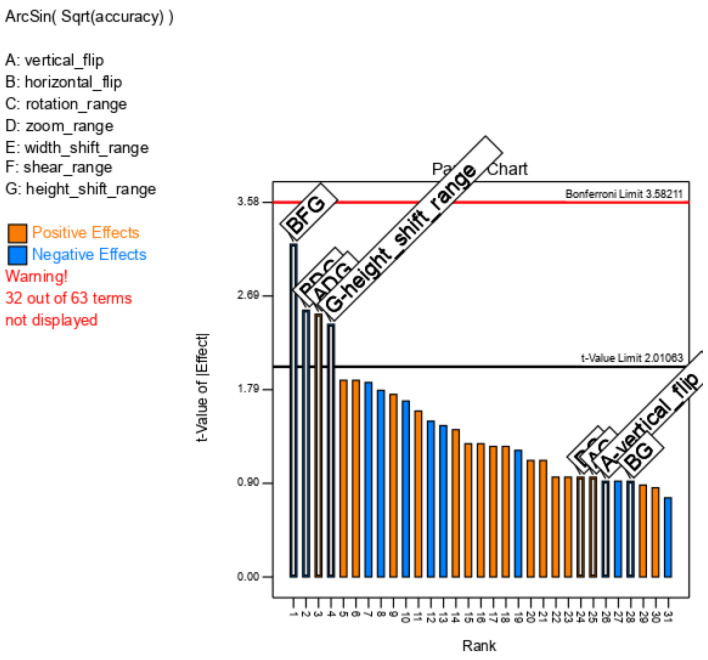

Figure 6. Pareto Chart of Data Set 2

Afterwards, the ANOVA is performed with the significant factors and interactions. The results show that the factor G, and the interactions BFG, ADG and BDG, are significant and negatively affect the accuracy of the model. Finally, a cube analysis is completed to analyze the interactions as demonstrated in **Figure 7**. By analyzing the vertices of the cube, the optimal data augmentation strategy is obtained which is when the horizontal_flip (factor B), zoom_range (factor D), shear_range (factor F) and height_shift_range (factor G) are at their low level. All other data augmentation operations are irrelevant and do not cause a significant change in the accuracy of the model.

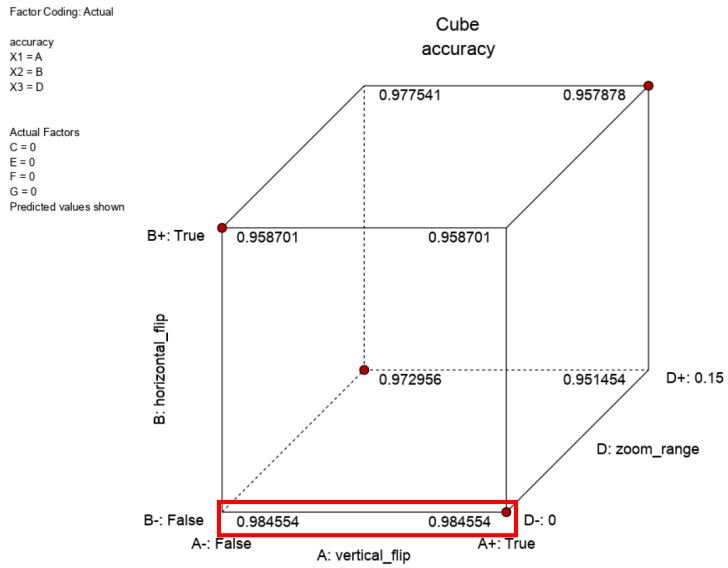

Figure 7. Cube analysis of Accuracy for Dataset 2

c. Data Augmentation analysis on Data Sets

In **Table 6** we present the results of the experimental design analyses for both datasets. In database 1 of around 1300 images, data augmentation is significant and has a positive effect. On the other hand, in database 2 of around 400 images, data augmentation is significant and has a negative effect. This means in database 2 the best thing would be not to use data augmentation since it decreases the precision of the model.

Table 6. Data Augmentation Operations effect on datasets

Data Augmentation Operation	Database 1	Database 2
A: Vertical_flip	Positively affect	Does not affect
B: Horizontal_flip	Negatively affect	Negatively affect
C: Rotation_range	Does not affect	Does not affect
D: Zoom_range	Positively affect	Negatively affect
E: Width_shift_range	Does not affect	Does not affect
F: Shear_range	Positively affect	Negatively affect

G: Height_shift_range	Positively affect	Negatively affect

Table 7. Optimal configuration of Data Augmentation Operations on datasets

Data Augmentation Operation	Database 1	Database 2
A: Vertical_flip	TRUE	TRUE or FALSE
B: Horizontal_flip	FALSE	FALSE
D: Zoom_range	0.15	0
F: Shear_range	0.25	0
G: Height_shift_range	0.2	0

For the two databases there are different optimal configurations that maximize the classification accuracy. For database 1, the operations of zoom, with a value of 0.15 and height shift with a value 0.2, are significant and have a positive effect on the accuracy. On the other hand, for database 2, only height shift is significant but has a negative effect. Hence, it should not be included as a data augmentation operation. In relation to significant interactions, it is known that there are optimal configurations, and these configurations are shown in **Table 7**. For both databases rotation and width shift are not significant, which mean they do not affect either positive or in a negative way the model´s accuracy.

d. Benchmark comparison

The CoroNet with the optimal data augmentation settings obtained with the experimental design is fully-trained and its performance compared with competing models in the corresponding datasets. The results are shown in **Table 8** for database 1 and **Table 9** for dataset 2.

Table 8. Performance comparison on data base 1

	Original CoroNet	Optimized CoroNet
Accuracy	95%	97%

Table 9. Performance comparison on data base 2

	Original COVID-Classifier	Optimized CoroNet
Accuracy	97%	100%

Although the objective of this research was to analyze the behavior of the different data augmentation techniques in the different sizes of databases, it was also possible to improve the results of the CoroNet neural network. In dataset 1 the accuracy achieved by Khan et. al (95% accuracy) was compared with the optimal configuration of the same neural network (97% accuracy), and an increase of approx. 2% was achieved. This means that by improving the accuracy of the neural network, 20,000 more patients can be correctly diagnosed for every 1 million patients.

For database 2, the precision achieved in the COVID-Classifier architecture proposed by Khuzani et. al (94% accuracy) versus CoroNet architecture with the optimal configuration of data augmentation (100%) was compared, and an increase of 6% was achieved. This means that by improving the accuracy of the neural network, 60,000 more patients can be correctly diagnosed for every 1 million patients.

Subsequently, another performance analysis was carried out where the same neural network (CoroNet) is compared with and without data augmentation and the results are shown in **Table 10**. It was found that the neural network trained with data augmentation achieved 97% accuracy versus the neural network without data augmentation reached 79% accuracy. This means that by applying data augmentation in the neural network, 180,000 more patients can be correctly

diagnosed for every 1 million patients. It is evidenced that data augmentation is a technique that significantly improves the accuracy of this neural network, and it is recommended to use it in neural network training for image classification.

Table 10. Performance comparison of Data Augmentation Operations on data base 1

	Real Data	Data Augmentation Data
Accuracy	79%	97%

CONCLUSIONES

In this study, we proposed a 2^7-1 experimental design to determine the combination of Data Augmentation factors than can improve the accuracy obtained from a CNN named CoroNet, which was specially designed to automatically detect COVID-19 infection from chest X-ray images. The 2^7-1 experimental design allows to introduce 7 different factors, each one with a low and high level that leads to a Resolution IV Design. The number of runs required for this design is 32. The proposed experimental design was tested on two datasets that reached a model p-value less than 0.05, which means that the models were significant. The best combination of Data Augmentation Parameters for the first dataset was: Zoom range: 0.15, height shift range: 0.2, shear range: 0.25, vertical flip: TRUE, and horizontal flip: FALSE, with an accuracy of approximately 92.44%. On the other hand, the optimized combination of Data Augmentation Parameters for the second dataset was: Height shift range: 0, shear range: 0, zoom range: 0, and horizontal flip: FALSE, with an accuracy of approximately 98.46%.

In further work an accuracy analysis will be performed on larger databases to find if there is similar behavior in relation to the data augmentation settings. Subsequently, another Design of Experiments will be applied where the optimal configuration of the significant factors can be found, and the accuracy of the neural network can be maximized. It is expected that in the future the same methodology can be applied to other hyperparameters of the neural network and to be able to increase the accuracy of the model. It is also expected to be able to apply this on other simple neural networks to be able to show if this can be replicated on these models.

An additional investigation that can help to validate the result obtained is to run the same experiment in a different CNN specialized in the automatic detection of COVID-19 infection from chest X-ray images and share a similar accuracy than CoroNet. It is important to identify and apply the same hyperparameters in the new CNN to make sure that the only factors that

will vary in the experiment are the data augmentation techniques. The results in the new neural network are expected to be like those obtained using CoroNet as one of the principles of this study is to demonstrate the effectiveness of the use of specific Data Augmentation techniques, regardless of other factors such as the CNN layout.

Finally, a complementary work for this investigation is to optimize the values for the Data Augmentation techniques found as significant in the experiment. For this purpose, optimization experimental designs such as composite core designs can be used. this design is defined as *2^k* factorial treatments with *2k* additional combinations called axial points and *nc* center points [30]. This technique will allow to find the best value that allows to maximize the accuracy of the neural network when using Data Augmentation in the generation of artificial images in the training set.

REFERENCIAS BIBLIOGRÁFICAS

[1] A. Narin, C. Kaya and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," Pattern Anal Applic, 2021.

[2] A. Elgendi, N. M., T. M. U., S. Q., G. D., B. J. P. and S. Nicolaou, "The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective.," Frontiers in Medicine, 2021.

[3] T. Ji, Z. Liu, G. Wang, X. Guo, C. Lai, H. Chen and Q. Zhou, "Detection of COVID-19: A review of the current literature and future perspectives," Biosensors and Bioelectronics, 2020.

[4] A. K. C. &. P. Z. Narin, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," Pattern Analysis and Applications, 2021.

[5] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna and J. J. Rodrigues, "Identifying pneumonia in chest X-rays: a deep learning approach," Measurement, 2019.

[6] B. Van Ginneken, C. M. Schaefer-Prokop and M. Prokop, "Computer-aided diagnosis: how to move from the laboratory to the clinic," Radiology, 2011.

[7] A. V. Vasilakos, Y. Tang, Y. Yao and others, "Neural networks for computer-aided diagnosis in medicine: A review," *Neurocomputing,* vol. 216, pp. 700-708, 2016.

[8] S. Albawi, T. A. Mohammed and S. & Al-Zawi, "Understanding of a convolutional neural network," International Conference on Engineering and Technology (ICET), 2017.

[9] A. Shoeibi, M. Khodatars, R. Alizadehsani, N. Ghassemi, M. Jafari, P. Moridian and P. Shi, "Automated detection and forecasting of covid-19 using deep learning techniques: A review," arXiv preprint arXiv, 2020.

[10] L. Wang, Z. Q. Lin and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," Scientific Reports, 2020.

[11] M. M. A. Monshi, J. Poon, V. Chung and F. M. Monshi, "CovidXrayNet: Optimizing data augmentation and CNN hyperparameters for improved COVID-19 detection from CXR," Computers in biology and medicine, 2021.

[12] A. D. Algarni, W. El-Shafai, G. M. El Banby, A. El-Samie and N. F. Soliman, "An efficient CNN-based hybrid classification and segmentation approach for COVID-19 detection.," Computers, Materials and Continua, 2021.

[13] M. W. Cohen, O. Gilo and L. David, "A Computer Aided Medical Classification System of COVID-19 CT Lung Scans using Convolution Neural Networks," Computer-Aided Design and Applications, 2021.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 1251-1258, 2017.

[15] A. I. Khan, J. L. Shah and M. M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Computer Methods and Programs in Biomedicine,* vol. 196, p. 105581, 2020.

[16] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway and A. Haworth, "A review of medical image data augmentation techniques for deep learning applications," *Journal of Medical Imaging and Radiation Oncology,* 2021.

[17] M. Elgendi, M. U. Nasir, Q. Tang, D. Smith, J.-P. Grenier, C. Batte, B. Spieler, W. D. Leslie, C. Menon, R. R. Fletcher and others, "The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective," *Frontiers in Medicine,* vol. 8, 2021.

[18] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data,* vol. 6, no. 1-48, 2019.

[19] L. Perez and W. Jason, "The effectiveness of data augmentation in image classification using deep learning," arXiv preprint arXiv:1712.04621, 2017.

[20] D. C. Montgomery, Design and Analysis of Experiments, New York: Wiley, 2013.

[21] M. Lujan and A. Gustavo, "Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study," Expert Systems with Applications 109, 2018.

[22] J. D. Kechagias, "A comparative investigation of Taguchi and full factorial design for machinability prediction in turning of a titanium alloy.," Measurement 151, 2020.

[23] C. Staelin, "Parameter selection for support vector machines.," 2003, Hewlett-Packard Company, Tech. Rep. HPL-2002-354R1 1 .

[24] F. Chou, Y. Tsai, Y. Chen, J. Tsai and C. Kuo, "Optimizing Parameters of Multi-Layer Convolutional Neural Network by Modeling and Optimization Method," IEEE Access, vol. 7, 2019.

[25] K.-T. Fang and D. K. Lin, "Uniform experimental designs and their applications in industry," in *Handbook of Statistics*, Elsevier, 2003, pp. 131-170.

[26] J. P. Cohen, P. Morrison and L. Dao, "COVID-19 image data collection," 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset.

[27] P. Mooney, "Chest X-Ray Images (Pneumonia)," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

[28] A. Z. Khuzani, M. Heidari and S. A. Shariati, "COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images," *Scientific Reports,* vol. 11, pp. 1-6, 2021.

[29] Google, "TensorFlow," Google LLC, 2021. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator.

[30] G. F. Preciado, "Optimización de una superficie de respuesta utilizando JMP IN," Mosaicos Matemáticos, 2003.

[31] "The effectiveness of data augmentation in image classification using deep learning".