Influence of Preprocessing and Segmentation on the Complexity of the Learning Machines in Medical Imaging

Juan Guillermo Paniagua[3], David Restrepo Rivera[1], Leandro Ariza-Jiménez[1],
Jose Julian Garces Echeverri[1], Christian Andrés Diaz León[2], Diana Lucia Serna-Higuita[1],
Wiston Arrazola[1], Sebastian Arango[1], Ramiro Velez-Koeppel[1], Miguel Angel Mejia[1],
Wayner Barrios[4], Jesus Francisco Vargas-Bonilla[5] and O. L. Quintero[1]

[1]Department of Mathematical Sciences
EAFIT University
cra 49 n 7 sur -50, Medellín Colombia
drestre30@eafit.edu.co,larizaj@eafit.edu.co,sarangom3@eafit.edu.co,
mmejiam@eafit.edu.co,oquinte1@eafit.edu.co,jgarcesec@gmail.com,
wistonarrazolalara@hotmail.com, dserna@gmx.net,rvelezk@eafit.edu.co

[2]Department of Social Communication
EAFIT University
cra 49 n 7 sur -50, Medellín Colombia
cdiazleo@eafit.edu.co

[3]Faculty of Engineering
Instituto Tecnologico Metropolitano
Calle 73 No. 76A-354, Medellín Colombia
juanpaniagua@itm.edu.co

[4]Wiqonn Technologies
Barranquilla Colombia
wbarriosq@wiqonn.com

[5]Department of Electronic Engineering-SISTEMIC
Universidad de Antioquia
cra 49 n 7 sur -50, Medellín Colombia
jesus.vargas@udea.edu.co

ABSTRACT

Medical applications of learning machines offer a challenge for both medical doctors and mathematicians/engineers. The main question arises when you decide to teach the model the same way a human being understands slight similarities and differences or improve the human (medical) capabilities for interpretation. Mathematically speaking, this translates on the artifact removal, preprocessing and contrast for the spatial frequencies on the images and their relationship with the complexity of the learning machine looking for the generalization and robustness of the solution for proper artificial intelligence based decision making.

This paper addresses these challenges and provides insights on their solution in real life scenario. Particularly, we offer a discussion on the knowledge database built from medical expertise and take care of the mathematical boundaries to be satisfied by the models. Also, we consider during our approach several restrictions from the authorities for the medical tool deployment using them as part of the indexes of performance.

Keywords: medical knowledge, learning machines, complexity, spatial frequencies, medical decision making, medical imaging, preprocessing and segmentation.

2010 Mathematics Subject Classification: 62H30, 62H35, 62M20, 68T05, 68T10, 68T99.
2012 ACM Computing Classification System: 10010336, 10010447, 10010083, 10010383.

1 Introduction

Recently, medical imaging became a fundamental problem for the development of learning machines in the artificial intelligence field. The problem of identifying several features in high dimensional spaces become more complex when technology advances allowing medical doctors to achieve high resolution on the human body images. Trying to build a learning machine or a computational solution is not exclusively for mathematicians or engineers; medical expertise is always needed for the proper assessment not only of the needs but also, the knowledge database construction.
"There is no ambiguity that a machine will never replace an MD expert, but machine intelligence will benefit and its aimed for human decision making. Finally, developing solutions is not just training a preconceived convolutional/deep network, neither machine learning is just a matter of a toolbox. It requires interdisciplinary work and mathematical background to imagine and formalize a proper feature extractor closest to the human perception of the world." (Quintero and Paniagua, 2020).
Deep learning is a non-deterministic polynomial-time hard (NP-hard) problem. Some people call it ”badly NP hard” because of the architecture of many layers of hidden variables connected by a non linear operations. One imagines that NP-hardness is not a barrier to provable algorithms in Machine Learning because the inputs to the learner are drawn from some simple distribution P and are not worst-case. If you want to really contribute to the medical applications, it is needed to extend the concepts already established for complexity in machine learning to deep networks. In 2013, (Arora et al., 2013) defined a theoretical approach for provable bounds in learning for deep learners. Aroras algorithm learnt almost all networks in this class with polynomial running time. Sample complexity is quadratic or cubic depending upon the details of the model, assuming a generative model $\mathcal{M}$ with many layers of denoising autoencoders. It allows a different learning methodology rather than backpropagation: layer-wise learning and unsupervised. This assumption is promising for theoretical development, but there is no known mathematical condition that describes neural networks that are or are not denoising autoencoders. Later these ideas were applied in the GoogLeNet design. Yes, even powerful practitioners use theory to support their achievements. Nowadays, Deep learners can be trained on images from scratch. Pre-trained CNNs can function as additional feature extractors that can be combined with existing handcrafted image features and the outputs

of pre-trained CNNs can be fine-tuned by another simpler classifier, on novel target images (Krizhevsky et al., 2012; Huang et al., 2016; Wang et al., 2017).

Yet, another class of approaches combines different CNN architectures to exploit the strengths and offset the weaknesses of a given architecture. It seems that the practical key will be the feature extraction automation via image processing tools and the inception of more layers (Quintero and Paniagua, 2020).

Nevertheless, for real life scenarios the mathematical boundaries hold and take serious influence on the model construction. One particular key problem is the definition of the information to be learned by the model in terms of the following: high resolution images, compressed images, high/low frequencies artifact removal, standardized solutions and human body proportionality. Consequently, all the former conditions became fundamental part of the technical solution. Generally speaking, the pre-processing and resizing of a large (and properly defined size database) take place during the model development and also, the enhancement techniques of artificial vision field constitutes a hot topic when the training of the learning machines is on duty. To train a learning complex machine based on the human expertise appreciation of the enhanced image (i.e large contrast) versus a raw image from the sensing machines is not a straightforward task.

The papers of Nino-Ruiz et al. (Nino-Ruiz et al., 2018), Precup et al. (Precup et al., 2016), Nino-Ruiz and Yang (Nino-Ruiz and Yang, 2019), Precup et al. (Precup and Tomescu, 2015), present several approaches for solving inverse and implicit optimization problems via local search, evolutionary, fuzzy algorithms and tabu search. Industrial applications of not only intelligent systems but also heuristic methods are widely applied and pottentially adressed on images.

Several papers have been published during the COVID-19 pandemic. Some of them focused on natural language processing of scientific literature, aiming to summarize the main insights to fight against the virus. Others, using artificial intelligence or machine learning for classification of lung pathology from Chest x-ray and Computerized Tomography (Wang and Wong, 2020). Many scientists presented their learning machines trained with previous X-ray data-sets and some new samples (compared to the needed to really build a model that can generalize) but do not provide the model, to be used on a rapid solution for real applications (Xie et al., 2016; Zhou et al., 2015). Several works address the architectures and the codes for training along with data-sets(Rajpurkar et al., 2017; Jaeger et al., 2014; Kermany et al., 2018) . But, this arises a question: how to develop a tool that *really* can be used by a medical doctor?

We aim to address several topics on this paper: first to highlight on the need for a large enough database previously analyzed by the medical doctors to the construction of a learning machine. Secondly, its impact on the possible modelling solution because of the mathematical boundaries for generalization, overfitting and complexity. Third, influence of the human understanding when pre-processing and artifact removal techniques are used to enhance the image to be used for the learning machine. Fourth the interpretability for the physicians and finally, their influence on the real life scenario of a technological solution when the health agencies set conditions such as the Food and Drugs Administration (FDA) in the U.S or Instituto Nacional de Vigilancia de Alimentos (INVIMA) in Colombia. These conditions and recommendations established by regulatory institutions allow the clinical needs and requirements of these detection algorithms and models to be aligned with the technical and functional specifications of the solution. In this way, three types of recommendations emerge aimed not only at establishing the conditions for obtaining use permits in the clinical setting, but also at feeding back the design and implementation of the medical device: (i) Design, implementation and evaluation of the software, (ii) Design and development of the algorithm and model and (iii) evaluation and validation of the model.

In this paper we propose a different concept of applied mathematical problems for learning machines in medical applications based on transformations that do depend on the parameters of spatial frequencies and artifacts removal; a new notion of learning laws in real life scenarios for decision making in COVID-19 treatment; and the corresponding calculation of several indexes supporting our findings.

2 Performance indexes for learning machines in medical applications

According to (Sendak et al., 2020) *”there has not been a systematic effort to ensure that front-line clinicians actually know how, when, how not, and when not to incorporate model output into clinical decisions. Nor is there an expectation that those who develop and promote models are responsible for providing instruction of model use and for the consequences of inappropriate use”*.

2.1 Complexity in learning machines

We will assume a space X full of data points to be used for learning (typically a classification but it may be an interpolation of data with aim to extrapolation and prediction). In our space X lies a lot of real valued features of the original and raw data. It allows us to use one set of data S for training (or teaching), in order to perform well on new data that have not been seen yet by the model $\mathcal{M}$ (or the learner). We are also looking for the simplest possible model built from a *representative* set of data. Consequently, we need to define what a *simple model* means and how the training set S should be in order to make it representative of the entire space X. To begin, we should assume there is a probability distribution P over the space X. The construction of our training set S seems very simple by randomly sampling points from X. We can be successful if we can predict well with our model $\mathcal{M}$ any other possible new points from X drawn by P. In that sense, we must assume that our training data are representative of the future data. There is a very tight relationship between the model $\mathcal{M}$ we construct and the sample set S. But, we must find a way to judge the former in terms of the latter:

Definition 2.1. The set C_i as the *target concept*, then, it will be paired with a subset of X according with this class.

Consequently, we produce a set $h \subset X$ called *Hypothesis* close to C_i with respect to the distribution P.

Definition 2.2. The true error of our hypothesis h is defined by

$$err_P(h) = prob(h\triangle C_i) \tag{2.1}$$

With $\triangle$ defining the symmetric difference with respect to the probability mass P. $\triangle$ is also known as the disjunctive union, it is the set of elements which are in either of the sets and not in their intersection. True error of h is the probability of misclassification of a data point drawn from X via P. We can judge our model through the hypothesis set! Focusing on our learner $\mathcal{M}$, what we really want is to have a very low *true error*, but we cannot control the fact we are not using the entire space X.

Definition 2.3. Define the training error as:

$$err_S(h) = \frac{\mid S \cap (h\triangle C_i) \mid}{\mid S \mid} \tag{2.2}$$

which is the fraction of points in our sampling space S on which the hypothesis class h and the target concept C_i disagree.

The most important thing about this intuition is that we can have a training set S with an almost perfect match of the hypothesis h with the target and yet still having a very high true error. This phenomenon is called *overfitting*. If you do not take it into account, your models are not going to generalize, even though you may still succeed in training.

Transfer learning is a common technique used in Deep Learning (DL) to re-purpose a trained model for a different task (Tan et al., 2018). Let $x\& \in \mathbb{R}^{HxWx3}$ denote the image of an x-ray, where H is height, W is width and there are three color channels for an image in the RGB color space. Let $y \in \mathbb{R}^n$ be the probabilities of the predicted labels. Given an image classifier $C \to \{1, 2, ..., n\}$, e.g., for an ImageNet (Krizhevsky et al., 2012) dataset n = 1000. It is possible to re-purpose a classifier $C1 \to \{1, 2, ..., n1\}$ and remap the classes of a deep neural network (DNN) such that $C1 \to \{1, 2, ..., n2\}$ where $n2$ is equal to the number of classes of the new dataset. But, in order to transfer the learning you need an available, generalizable, non overfitted model. This is not the case for our application.

The main idea is to specify that most data sets are incompressible. Suppose our goal is to encode a binary sequence of length n. Then, no matter what description length method we use, only a fraction of at most 2^-k sequences can be compressed by more than k bits. This, if data are generated by fair coin tosses, then, no matter what code we use, the probability that we can compress a sequence by more than k bits is at most 2^-k.

The notion when you use an arbitrary distribution translates into the following: the probability that a code not corresponding to the distribution P in the space S compresses the data more than the code that does not correspond to P is negligible. Consequently, having a short description length for the data is equivalent to identifying the data as belonging to a tiny, very special subset out of all priori possible data sequences.

2.2 Common performance indexes

Since our mathematical model will be close related to the sample data, in practice, the entire space needs to be randomly sub sampled in train, validation and test sets to perform proper validations. The confusion matrix, also known as an error matrix, is a specific layout for simple visualization of the performance of a supervised learning algorithm and is commonly used as a base for evaluate and assert the individual skills of complex classifiers providing different metrics for quantification of the capabilities of the model. Some of the most common used metrics includes the accuracy which gives information about the quality of the detection and the recall (or sensibility) which tell us the fraction of the total amount of relevant instances that were actually retrieved. The f1 score, is also widely used since it giving us general information about the model.

Therefore, these simple approaches could easily mislead to optimistic results since we are not performing a proper validation in the complete problem space. As generalization and performance becomes more important for Machine Learning applications, current researchers have shifted away from simply presenting accuracy results when performing an empirical validation to give a set of full descriptors of the model test performance.

- Sensitivity: It is also known as recall metric and it evaluates the model's ability to predict true positives of each category, i.e, the proportion of actual positive cases that got predicted as positive or true positives.

- Specificity: It is also known as false negative rate and corresponds to the metric that evaluates a model's ability to predict true negatives of each category, in other words, this metric measures the proportion of negatives that are correctly identified as negative.

- F1 score: it is defined as the harmonic mean of precision and recall and often interpreted as a weighted average of both metrics.

- ROC: The Receiver Operator Characteristic (ROC) curves are practical plots commonly used to explore the real capabilities of a binary classifier. The graph is usually conformed for the sensitivity or probability of detection and the false-positive rate interpreted as the probability of false alarm and arranged in a 2D plot for different threshold settings [47]. Typical positive results exhibit a ROC curve moving towards the upper-left corner of the plot, this event can be interpreted as high level of quality and probability of detection for a given classifier.

- AUC: Area under the ROC Curve (AUC) presents an aggregate metric of performance across all possible classification thresholds. One interpretation meaning of AUC is as the probability that the model ranks a random positive sample more highly than a random negative sample. Besides, higher AUC means better performance of the model at distinguishing between the positive and negative cases.

- Training FLOPS: FLOPS is acronym for floating point operations per seconds. This benchmark is commonly used to measure the computing device performances. It does define how many instructions a processor can perform per second. This measure also allow to anyone to determinate code portability, for example, recent ML techniques help to get more efficiently neuronal network models improving the trade-off between accuracy and on-device latency running in low performance GPU/CPU and mobile devices (Jacob et al., 2018).

2.3 Medical interpretability and regulatory restrictions

2.3.1 Utility

Developing medical devices for the diagnosis of different pathologies from a set of medical images, it is important to consider not only the requirements and technical specifications associated with detection, but also include the context and the clinical requirements of operation and use of the algorithm.

In the path that means creating a diagnostic detection system, from its design and development to clinical practice and use, it is important to include from the beginning, during the design and implementation exercise, the constraints that the clinical context and the regulatory system can impose on the design.

As stated in the Biodesign (Yock et al., 2015) methodology, consider the aspects of clinical use of the device and the requirements that it must meet at the regulatory level, not only feedback the design of the solution, but also accelerate the technology implementation processes in the clinical context.

2.3.2 Regulatory Guides Recommendations and Influence

The FDA has paid special attention to defining the regulatory guidelines to consider in medical diagnostic detection devices that use intelligent algorithms, for this reason it has approved several medical devices of this nature in recent years. Regarding the efforts carried out by the FDA in this case, the Guidance for Industry and Food and Drug Administration Staff Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data was published (Food and Administration, 2012). Recommendations regarding the design, development, reporting, and validation of the algorithms, databases, and models used in these types of diagnostic devices are established in this guide.

In this way, three types of recommendations emerge aimed not only at establishing the conditions for obtaining use permissions in the clinical setting, but also at feeding back the design and implementation of the medical device: (i) Design, implementation and evaluation of the software, (ii) Design and development of the algorithm and model and (iii) evaluation and validation of the model. This section focuses on describing and analyzing how the recommendations raised by the FDA in the last two previous points define requirements at the level of design and validation of the model and algorithm.

3 On the knowledge database and models for COVID-19

In order to properly address the influence of pre-processing and segmentation on the complexity of the learning machines in medical imaging we tackle the problem as follows:

- Build a database that satisfies minimally the requirements for learning as in section 2.1
- Use a preprocessing scheme that allows to adapt, improve and standardize the images fed to the learning machines such as resizing, contrast enhancement, among others.
- Mimic the utility of the medical expertise on the knowledge database construction via Contrast Limited Adaptive Histogram Equalization (CLAHE)
- Evaluate the results of the pre-processing + CLAHE or its absence on a learning machine complexity and its performance indexes including interpretability of medical results

Consequently, the influence of the human understanding when pre-processing and artifact removal techniques on learning machine will be presented in next section.

3.1 Database

Looking for a quick answer to the needs of Colombian rural areas when no high-level hospitals exist, we aim to develop a robust, generalizable, not overfitted model to assist on the X-ray patients classification. On these hospitals, a radiologist is not always available and medical doctors are not necessarily proficient of x-rays assessment and must be completely overwhelmed with the treatment of COVID-19 patients. Consequently, we put together a team of Radiologists Medical Doctors, Intensive care Medical Doctors, Medical Doctors, Mathematicians and engineers to develop a medical validated model. In section 2.1 we stated the main considerations for the proper learning task and of course the real life scenario takes place when the size of the entire space for sampling is large and the hypothesis class is as complex as a radiology result.

For the creation of the database, images from the following Datasets were selected: CheXNet dataset (Rajpurkar et al., 2017; Wang et al., 2017), Chest X-Ray Images (Kermany et al., 2018), Montgomery County chest X-ray set (Jaeger et al., 2014), RSNA Pneumonia Detection Challenge, ChestX-ray14, The open source COVIDx dataset, Chest Xray Pneumonia, Chest CT Scans with COVID-19 Related Findings, Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements and Prognosis of COVID-19 Pneumonia Using Computed Tomography; which are available on the internet and are freely accessible, in which the identity of the patients is respected, and which also have images of patients diagnosed with COVID-19, and we also have images obtained from local hospitals (Kermany et al., 2018; Jaeger et al., 2014; Wang and Wong, 2020).

The images were classified as follows: ”HEALTHY” and ”UNHEALTHY”, and this second one is subdivided into ”Bacterial Pneumonia”, ”Viral Pneumonia - Possible COVID” and ”Others”. This last category includes other pathologies that affect the lung parenchyma such as nodules, lung mass, pleural effusion, pneumothorax, among others. The importance from a medical point of view is to develop a clinical decision support system that allows health personnel to take early measures regarding the treatment of patients suspected of having COVID-19 infection, given that the result of the PCR for COVID-19 takes several days to process in our environment, and the health systems collapsed during the peak of the pandemic. We currently have the following images:

Class	Chest X Ray	CT Thorax
Healthy	8.184	1452
Others	8.684	598
Bacterial Pneumonia	1.957	797
Viral Pneumonia - Possible COVID	519	1.620
Total	19.344	4.467

Table 1: A description of the database built for the preliminary study

3.2 Pre-processing scheme

Visual interpretation and analysis of digital images depends on how well the scene being imaged, e.g., the structures of the human body, are captured by the acquisition device. Once acquisition is done, it is customary to pre-process images with the purpose of adapting or improving them before feeding them into more complex processing stages. Common tasks performed in the pre-processing stage, when deep machine learning is used to process images, include the modification of their original dimensions and the enhancement of their contrast.

The dimensions of input images are usually reduced to an appropriate size before feeding the learning machines. The premise of size reduction is essentially to decrease the computing times of the learning machines, at the expense of the image quality.

Since the selection of an output size depends on the choice made on the learning machines, there is no consensus regarding the image final size after reduction, except that the reduced image is usually square (Ronneberger et al., 2015; Zhang et al., 2020). In our case, input images will be resized to 256×256 pixels.

Before addressing the problem of image contrast enhancement, we will provide the following definitions. Let $X(i,j)$ be an input grayscale image, where $i = 0, 1, \ldots, M-1$ and $j = 0, 1, \ldots, N-1$, with L discrete intensity levels in the range $[0, L-1]$.

The actual range of intensity levels found in an image $X(i,j)$ is called the *dynamic range* of the image. This particular ranges omits evident outliers. Ideally, the dynamic range should match with all the L usable intensity levels. In such case, the image is said to exhibit *high contrast*. On the contrary, an image has *low contrast* when the dynamic range is low, i.e, a small range $[X_{\text{low}}, X_{\text{high}}]$ of the available intensity levels are used, where $0 < X_{\text{low}}$ and $X_{\text{high}} < L-1$ (Dougherty, 2009).

Conventional pre-processing techniques of contrast enhancement focuses then in improving how images use the available intensity scale. For instance, a low contrast image $X(i,j)$ can be automatically adjusted to obtain a second image $Y(i,j)$, in which the available range of intensities is fully covered, based on the following operation (Burger and Burge, 2009):

$$Y(i,j) = \frac{L-1}{X_{\text{high}} - X_{\text{low}}}(X(i,j) - X_{\text{low}}).$$

Digital images can be characterized based on their corresponding histogram. Histograms are a simple and efficient tool that allows us to identify problems that originate during the image acquisition stage. The amount of contrast and the dynamic range of an image, for instance, can be inferred based on histograms.

In particular, the histogram of $X(i,j)$ is a discrete function $h(k) = n_k$, where $k \in [0, L-1]$ and n_k is the number of pixels having an intensity level equal to k (Gonzalez and Woods, 2006).

In the following section, we present a more complex contrast enhancement method that focuses in redistributing the histogram of an input image with the purpose of improve the differentiation of anatomical structures in medical images.

3.2.1 Contrast Limited Adaptive Histogram Equalization (CLAHE)

In different medical images, such as x-rays and CT scans, the distribution of grayscale intensity levels is extremely localized. Thus, it may affect contrast, image quality and cause sensitivity issues in deep learning machine classification.

The Contrast Limited Adaptive Histogram Equalization (CLAHE) transforms the grayscale image for contrast enhancement. The aim is to obtain an image with intensity levels uniformly distributed throughout the intensity scale. It calculates several histograms, each corresponding to a different section of the image, and uses them to redistribute the brightness values of the image. Therefore, it is suitable for improving local contrast and improving edge definitions in each region of an image. CLAHE performs local adjustments of image contrast with low noise amplification. The contrast adjustments are interpolated between the neighboring rectilinear image patches called kernels and the spatial adaptivity in CLAHE is achieved through selection of the kernel size. The intensity range of the kernel histogram (or local histogram) is set by a clip limit that restrains the noise amplification in the outcome (Stimper et al., 2019).

Histogram equalization is one of the well-known methods used to improve the contrast of an image. The histogram equalization distributes pixel values uniformly and results in an enhanced image $Y(i,j)$ with linear cumulative histogram, given by (Wadi and Zainal, 2012)

$$Y(i,j) = X_{\text{low}} + (X_{\text{high}} + X_{\text{low}})c(X_k) \tag{3.1}$$

where, X_{low} and X_{high} are the minimum and maximum gray levels values, respectively; $c(X_k)$ is cumulative distribution probability given by

$$c(X_k) = \sum_{j=0}^{k} P(X_j) \tag{3.2}$$

where

$$P(X_k) = \frac{n_k}{MN} \tag{3.3}$$

where, $k = 0, 1, ..., L-1$ is the intensity level; n_k is the number of pixels with intensity level k, M and N are the dimensions of the image, and X_k value of intensity level k.

Histogram equalization can cause unsatisfactory results, as the histogram of the final image becomes almost flat. This does not adapt to local contrast conditions, ignoring contrast values where the gray scales are relatively small.

When using CLAHE, two parameters are needed: the block size and the clip limit, which are used to control the quality of the image. These parameters allow the image to be divided into sub-images or blocks, thus, the equalization of the histogram is made for each sub-image. Artifacts between neighboring blocks are minimized through filters or bi-linear interpolation. The clip limit allows reduction of the noise problem (Min et al., 2013).

4 Influence of pre-processing on the learning machines for medical decision making

As well known, the learning of a mathematical model is not the same as training any mathematical structure. So many optimization problems can be solved by means of several methods and always a set of optimal parameters that full-fill a cost function can be found. Learning a class or multiple classes involves the process of determining the proper sample size and the selection of a mathematical model with a complexity that satisfies the bounds for learning. In pattern recognition in low dimensional spaces this guarantee is satisfied in most of the cases, but for pattern recognition in high dimensional spaces such images, most models now are focused on deep learning machines.

This is a machine learning model that seeks not only to classify healthy or unhealthy patients, but also to robustly generalize the results for rural care areas where access to high-resolution tomography is impossible. An extension of this tool will be taken to the classification of healthy patients, patients with bacterial pneumonia and finally patients with pneumonia of viral origin (multi-class learning machine). Our contribution also incorporate a Model facts label (Sendak et al., 2020) prepared by the interdisciplinary team in order to find a common language between physicians and engineers.

In order to reply to the medical doctors' potential needs, we preferred to perform transfer learning on a known model (Rajpurkar et al., 2017), but the main problem was the lack of the real supposed to be available one. Then, the entire training process had to be carried out in order to teach the model how to represent at least two classes. Related to COVID-19 the pathology useful to be detected on X-rays is viral-pneumonia. Then the training set for this task is a real quest. After a dataset curation done by medical doctor (Kermany et al., 2018; Jaeger et al., 2014; Wang and Wong, 2020) and several attempts of modelling, we provide a useful model to help on the improvement of the health services, assisting to medical doctors with no access to computerized tomography on the early decision-making process. Deep learning-based models are, without a question, powerful tools for processing digital images. When it comes to implement solutions for digital image processing applications, it is customary to train models end-to-end in order to handle by itselft the feature extraction and classification steps. Therefore, models had to learn to identify patterns from complete images rather than from previously segmented regions of interest. To overcome this, an alternative is using models for different and specialized tasks. A first model can be trained to segment regions of inter-

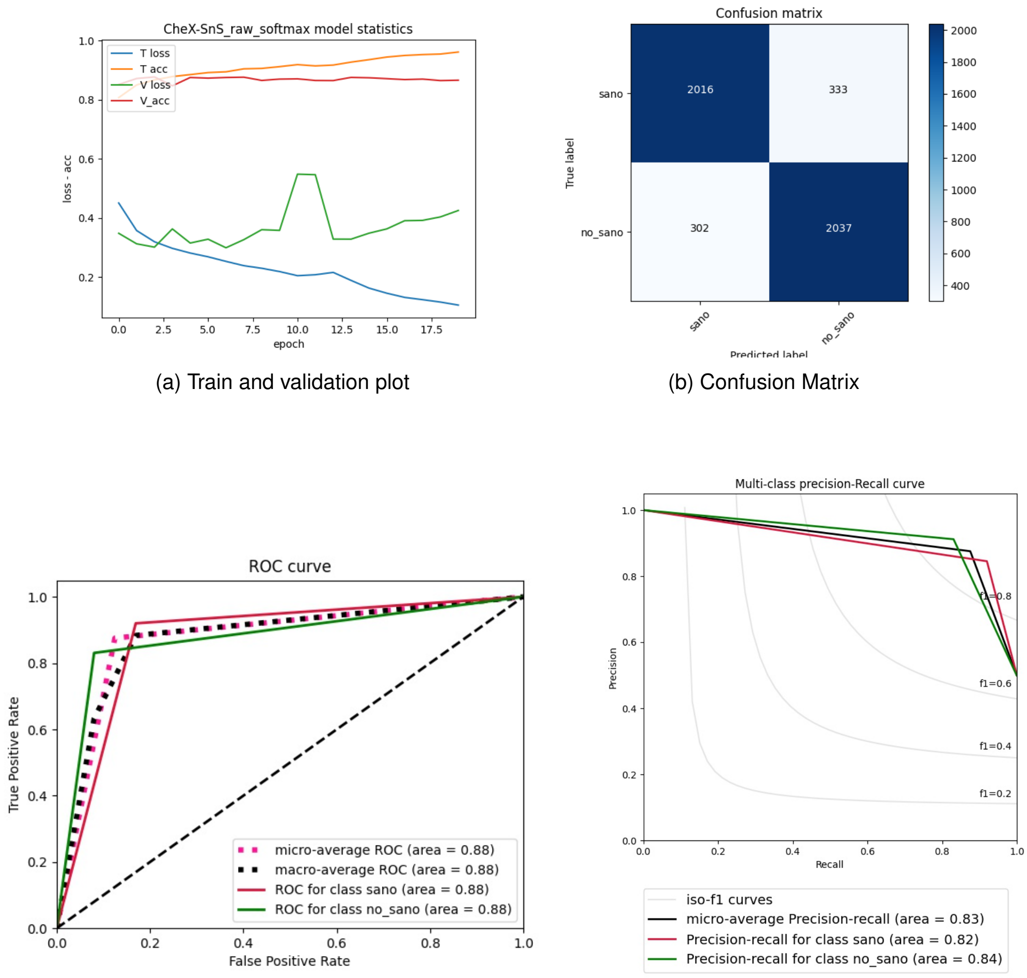
est based on datasets containing samples with their corresponding ground-truth segmentation mask. Then, a second model can be trained to learn features and a subsequent classification from the above regions of interest. A motivation behind this alternative workflow is that by limiting the scope of the DL-based models to specific and refined regions of input images, a better performance on classifying these regions can be achieved.
U-Net is a complete deep convolutional network based on auto encoders focused on biomedical image segmentation. One of its most notorious features is that it can be trained with few example images and still leads to precise segmentations. To achieve this, a U-Net model is re trained on manual lung segmentation masks applying a strategy based on the intensive use of the concept of ”data augmentation”, which allows increasing the amount of training images from an initial set to which elastic deformations are applied (Ronneberger et al., 2015).
We proposed a DenseNet-121 (Huang et al., 2016) model pretrained on the ImageNet (Krizhevsky et al., 2012) dataset as feature extractor with a binary classifier with two softmax activated neurons. The training was made to optimize a binary cross-entropy loss with an Adam optimizer using batches of data of 16 images during 20 epochs of total training saving only the models with best validation accuracy.
In terms of computational complexity, our model presents $6.96 * 10^6$ number of parameters and 5.76 GFLOPs. The results obtained indicate that our architecture performs on par with the state of the art models such as ResNet architecture (He et al., 2016) on image classification problem baseline. It also requires significantly fewer parameters and computation to achieve suitable performance in compared to more deeper models such as ResNet-152 in the same image classification baseline.

4.1 Learning machine raw input images

Best accuracy is reached after 3 epochs of training from where the model start overfitting the data as show in Figure 1 a), the validation loss make big jumps while the training accuracy continues increasing suggesting that the model is starting to learn features related to the training set but not generalized to the entire problem space. The best trained model presents a sensitivity of 83% with accuracy and precision of 88% and 91% respectively which is a good performance for a model and can also be evidenced in the PRC and ROC curves with AUC of 88% calling for good generalization properties.

4.2 Learning machine with segmented lungs and NO contrast on the input images.

For the case of training with segmented lung images, the model reach its best performance after 9 epochs of training furthermore, the validation and training curve shows small sights of overfitting after epoch 3. For the best trained model, some metrics decreased in relation with the no segmented model, this could be due to some information loss related to pathology presented outside the lung region which are presented in our database at small rate. Nevertheless, the model displays good performance as presented in Figure 2 c), d).



(a) Train and validation plot

(b) Confusion Matrix

(c) ROC Curve

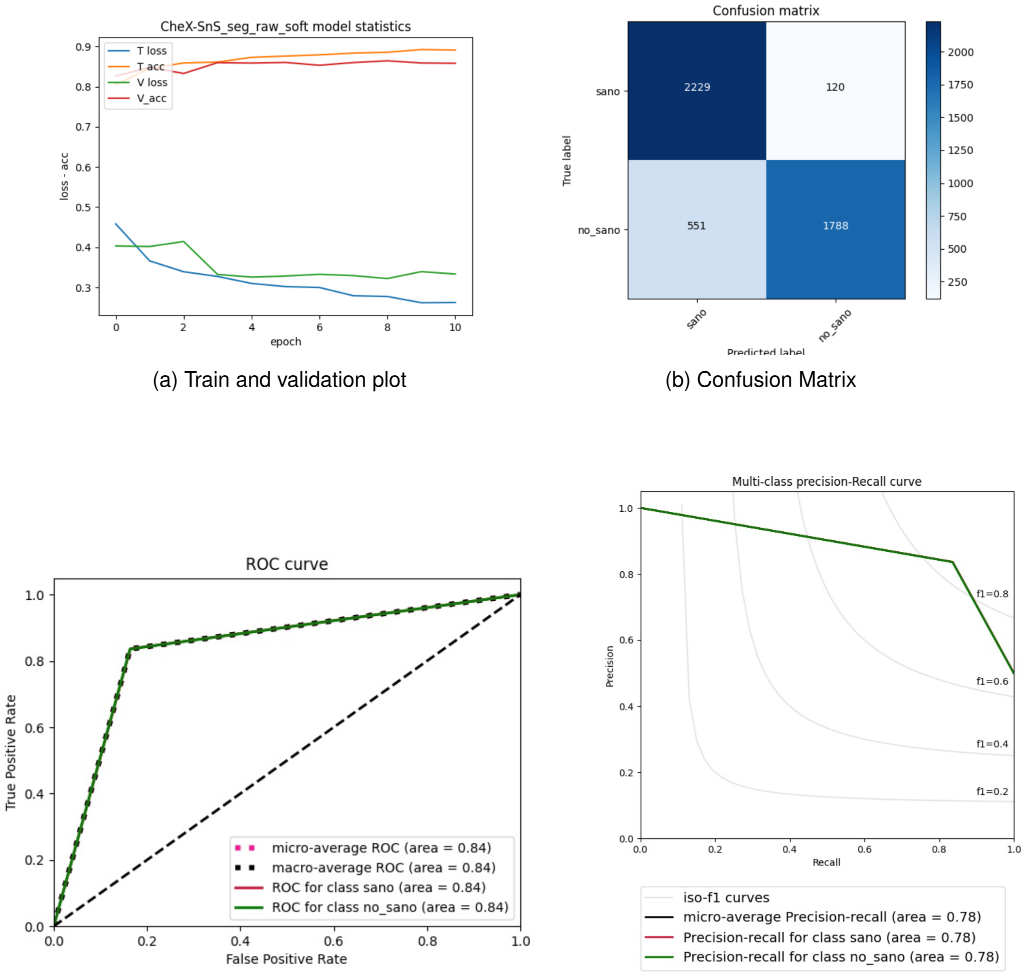
(d) PRC Curve

Figure 1: CheXNet model with raw input image, best validation accuracy at 3 epochs of training.

4.3 Learning machine with strong contrast on the input images

Figure 3a and 3c show the comparison between the original Rx image and the Rx image processed with CLAHE with a clip limit of 0.02. We can see the the changes in the contrast of the image. Figure 3b and 3d show the comparison between the original histogram and the CLAHE histogram. We can note that the histogram is redistributed in the intensity values.
The model trained with the proposed technique reach its best accuracy at epoch 7, from Table 2 we can evidence that the procedure improves the sensitivity of the model impacting the false negative rate which is a important result in medical applications.
The main recommendations raised by the FDA regarding the validation process are categorized into three groups:

- Recommendations related to the study population: At this point, the regulatory entities propose that the evaluation of the algorithm should be applied standalone, without the medical interpretation included, and considering the classification made, as well as the score associated with the classification. Another important recommendation is that a different dataset should be used for the validation process and training process for each



(a) Train and validation plot

(b) Confusion Matrix

(c) ROC Curve

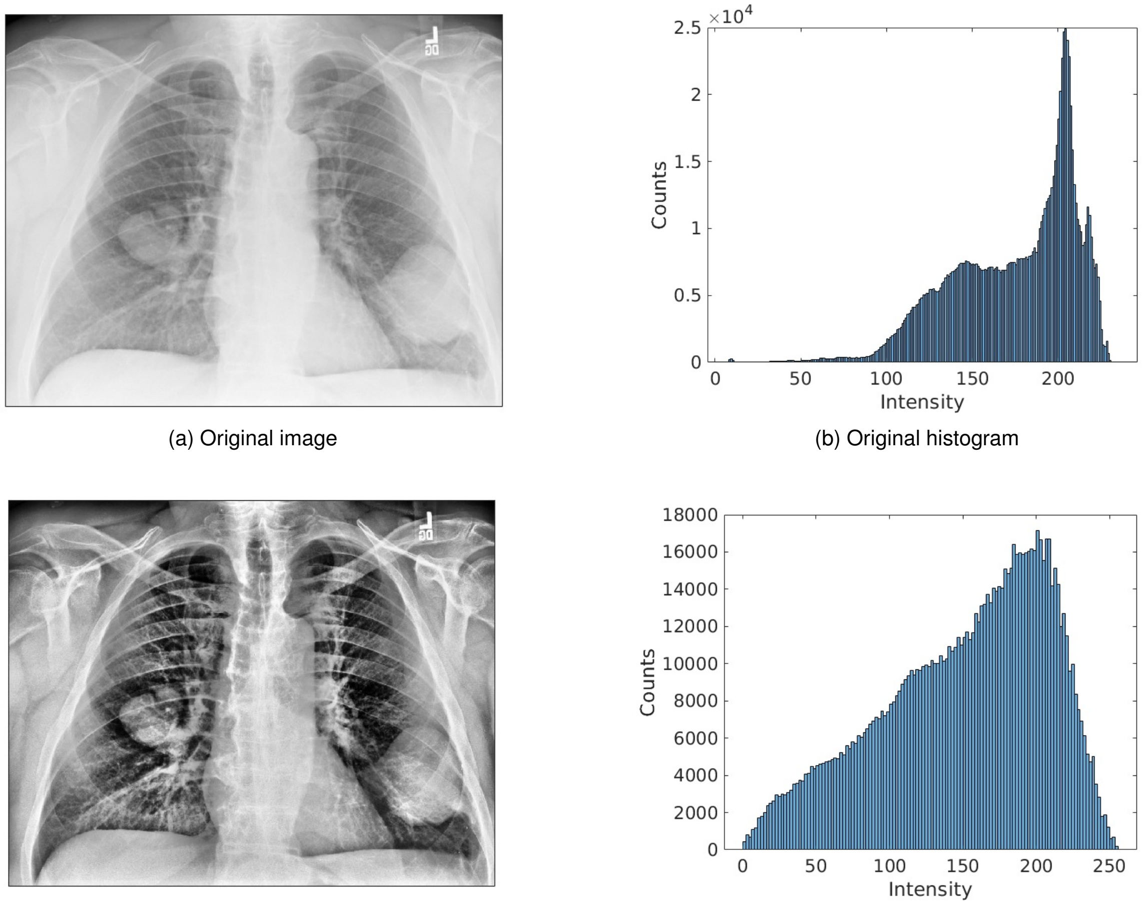
(d) PRC Curve

Figure 2: CheXNet model with segmented lungs as input image (no clahe), best validation accuracy at 9 epochs of training.

of the algorithm phases. The testing database should be representative of the target population and the target disease, condition, or abnormality for which your device is intended. Additionally, the sample size of the study should be large enough to provide adequate power of statistical significance in detection.

- Recommendations on data reuse: In general, the FDA recommends the non reuse of data for evaluation and validation, since the interpretation of the results obtained by the device could be problematic. In case, data reuse is used for the evaluation, it must be demonstrated that reusing parts of the test data does not introduce bias in the algorithm estimates and that the integrity of the data is maintained.

- Recommendations related to detection accuracy: The recommendations made by regulatory entities focus on the performance of the algorithm itself and the variability factors that can affect the precision of the device in this item. Considering the standalone performance of the algorithm as mentioned, the classification and the score or probability in the classification made by the device should be evaluated, without considering the inter-



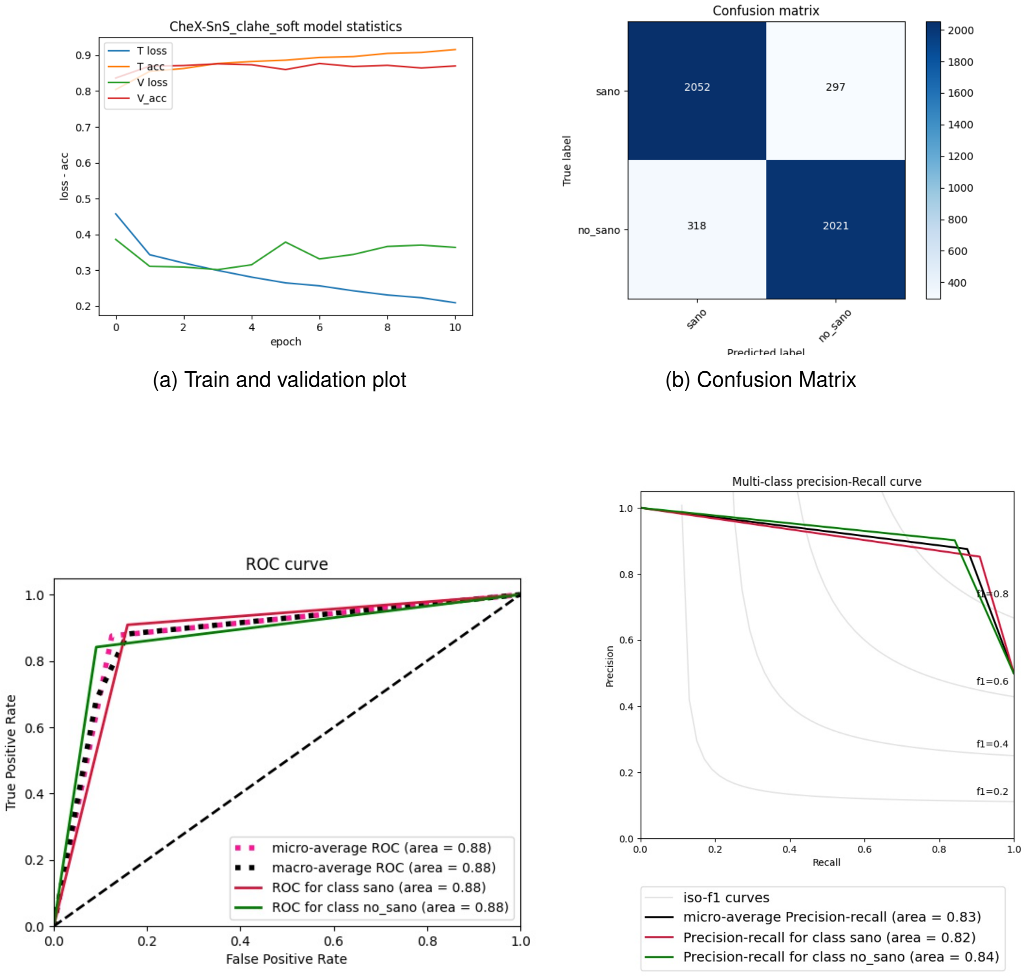
(a) Original image

(b) Original histogram

(c) Rx with CLAHE

(d) CLAHE histogram

Figure 3: Contrast enhancement with CLAHE 3a Original Rx image, 3b histogram original Rx, 3c Rx image with CLAHE, 3d histogram Rx with CLAHE



(a) Train and validation plot

(b) Confusion Matrix

(c) ROC Curve

(d) PRC Curve

Figure 4: CheXNet model with input image processed with clahe, best validation accuracy at 7 epochs of training.

pretation of a doctor. Furthermore, the definition of a true positive, true negative, false positive, and false negative CADe mark should be consistent with the intended use of the device. Considering the variability factors, a grouping of these into two categories is proposed, those related to the manual classification and labeling of the images that are part of the database. One method of accounting for variability in the reference standard is to resample the expert truthing panel (Miller et al., 2012). The other category corresponds to all those factors that affect the nature of the image or the abnormality that is to be detected, such as, for example, the type of injury, the size of the injury, the location of the injury, the stage of the disease, the imaging protocols, and the characteristics of the capture devices. Finally, it is recommended to analyze during the evaluation process the different thresholds or algorithms that can manipulate or affect the performance of the algorithm alone. In this article, for example, it is evaluated how a preprocessing and segmentation stage can affect the performance of the models and the classification made by it. At this point it is important to consider whether thresholds used or scoring mechanisms may vary or modify the performance of the device.

- Recommendation related to generizability testing: Finally, the latest recommendations made by regulatory entities are related to the generalization of the model and the algorithm when it is going to be applied in a less controlled environment than the one evaluated and validated. The most important recommendation in this category refers to justifying the extent to which the results of the studies carried out can be generalized, considering for example the device and image acquisition protocols. For this reason, it is recommended to evaluate the impact that the imaging hardware, imaging and scanning protocol and image or data characteristics has on the performance of the algorithm itself.

All these recommendations and suggestions raised by the clinical context and the regulatory system become requirements and design specifications that must be taken into account during the conceptualization, design and validation of the detection device. Failure to consider them during early stages of development may mean reprocessing in later stages of the project. Likewise, the construction of the detection device and the models should be aimed at facilitating and contributing to the interpretation that health professionals made to the classification and the scoring method provided, considering the way in which they are calculated and obtained and also how they are presented and presented to the medical professional. Our contribution fulfills all the regulatory considerations by filling all mathematical requirements.

During the COVID-19 pandemic, several Artificial Intelligence tools have arise. Nevertheless, the authors are far of being credible by medical doctors, due to the fact that rigorous clinical studies must be carried out. Also, they claim to have the ultimate solution but neither publish their model nor satisfy the learning bounds for generalization and robustness of the learning machine. These are bad news if you want to have a quick solution for an underdeveloped country. The Colombian health system and that of several Latin American countries may collapse due to the potential massive assistance to the emergency services. In rural areas the scenario is not very optimistic due to the lack of high level hospitals, imaging devices and expertise in medical personnel. Looking for a rapid, robust, and usable solution to medical doctors in rural areas we propose to have a mobile device application. This application has the purpose to provide a support system for medical decision based on the learning machines trained by our team on a database reliable and curated by medical doctors, and radiologist experts on the field.

The medical responsibility in real life is higher than the scientists can expect and we are not finished yet, but our solution can be used because the learning task can guarantee the learning of at least two classes in our models. We used several databases curated by medical doctors and we are carefully satisfying the learning theorems for the training, validation and test of our models. We do not want to over-fit our models, but to provide a reliable tool. Robustness is necessary and we also want to provide our model results in an understandable way for medical doctors. That is why we present our solution in terms of the medical labels proposed by (Sendak et al., 2020).

Metric	Raw input	CLAHE	Lung segmentation
Epochs to convergence	3	7	9
Sensitivity	0.83	0.84	0.84
Specificity	0.92	0.91	0.84
Precision	0.91	0.90	0.84
Negative Predicted Value	0.85	0.85	0.84
False Positive Rate	0.08	0.09	0.16
False Discovery Rate	0.09	0.10	0.16
False Negative rate	0.17	0.16	0.16
Accuracy	0.88	0.88	0.84
F1 Score	0.87	0.87	0.84
Matthews Correlation Coeff	0.75	0.75	0.67

Table 2: According to results must be provided in such a way the medical doctors will rely on the tool for decision making

5 Concluding remarks

Systems for support of clinical decision-making in the field of medicine and whose operation is based on deep learning and artificial intelligence, are becoming increasingly important in helping doctors make the best decision regarding the management and treatment of patients, increasing safety in medical care.

Building a medical assistance tool based on Artificial intelligence is not as easy as train a deep learner with a data-set. Medical doctors from several specialties have been developing protocols for assistance based not only in epidemiological information, but also comorbidities and radiology information of the patients.

Medical studies and protocols are complex and their relevance the highest possible. Lives depend upon their decisions. If they base their decisions on a learning machine to diagnose or define the treatment for a patient, even legal consequences may be carried out.

The main objective of image pre-processing is to improve the features of the image by eliminating artifacts, distortions, unwanted noise and / or to improve some relevant details in such a way that the learning machines can benefit and improve their performance.

The images used were resized to establish a standard size (256×256) to feed our artificial intelligence algorithm. The aim is to ensure that the images have the same size and aspect ratio and obtain less network learning time.

Comparing the data obtained with raw input images and with input images preprocessed with CLAHE, before being resized and fed, although the accuracy remains constant, it improves the sensitivity of the algorithm by improving the classification of positive cases. Likewise, it reduces the false negative rate, an unwanted indicator at the time of classification, and improves the false positive rate, an indicator of preferable classification.

In future work, we will design and define other preprocessing modules to emphasize the specific features of the images to improve the accuracy and sensitivity of the learning machines. By developing a customized preprocessing module for x-ray and CT images, we hope to achieve better results. The results of the learning machines previously using the segmentation network with and without the subsequent application of CLAHE were not entirely satisfactory. Therefore, we will continue to analyze another possible methodologies in the preprocessing stages, such as a combination of spatial filtering techniques for sharpenning the input images and intensity transformation functions (e.g, power-law or gamma transformations (Gonzalez and Woods, 2006)).

Acknowledgment

This work was supported by the Funding of Ministerio de Ciencia y Tecnologia Colombia.

References

Arora, S., Bhaskara, A., Ge, R. and Ma, T. 2013. Provable bounds for learning some deep representations, *in* E. P. Xing and T. Jebara (eds), *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32 of *Proceedings of Machine Learning Research*, PMLR, Bejing, China, pp. 584–592.
URL: *http://proceedings.mlr.press/v32/arora14.html*

Burger, W. and Burge, M. J. 2009. *Principles of Digital Image Processing: Fundamental Techniques*, Undergraduate Topics in Computer Science, Springer London, London.

Dougherty, G. 2009. *Digital Image Processing for Medical Applications*, Cambridge University Press.
URL: *https://www.cambridge.org/core/product/identifier/9780511609657/type/book*

Food and Administration, D. 2012. *Guidance for Industry and Food and Drug Administration Staff Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions*, FDA.

Gonzalez, R. C. and Woods, R. E. 2006. *Digital Image Processing (3rd Edition)*, Prentice-Hall, Inc., USA.

He, K., Zhang, X., Ren, S. and Sun, J. 2016. Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* .
URL: *http://dx.doi.org/10.1109/cvpr.2016.90*

Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K. Q. 2016. Densely connected convolutional networks.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H. and Kalenichenko, D. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X. and Thomas, G. 2014. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases, *Quant Imaging Med Surg* **4**(6): 475–477.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F. et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning, *Cell* **172**(5): 1122–1131.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, pp. 1097–1105.

Miller, D., O'Shaughnessy, K., Wood, S. and Castellino, R. 2012. Gold standards and expert panels: A pulmonary nodule case study with challenges and solutions, *Proc. of the SPIE, Medical Imaging*, Vol. 5372, pp. 173–184.

Min, B. S., Lim, B. S., Kim, S. J. and Lee, S. J. 2013. A novel method of determining parameters of clahe based on image entropy, *International Journal of Software Engineering and Its Applications* **7**(5): 113–120.

Nino-Ruiz, E. D., Ardila, C. and Capacho, R. 2018. Local search methods for the solution of implicit inverse problems, *Soft Computing* **22**(14): 4819 – 4832.

Nino-Ruiz, E. D. and Yang, X. S. 2019. Improved tabu search and simulated annealing methods for nonlinear data assimilation, *Applied Soft Computing* **83**(105624).

Precup, R.-E., David, R.-C., Petriu, E. M., Szedlak-Stinean, A.-I. and Bojan-Dragos, C.-A. 2016. Grey wolf optimizer-based approach to the tuning of pi-fuzzy controllers with a reduced process parametric sensitivity, *IFAC-PapersOnLine* **49**(5): 55 – 60.

Precup, R.-E. and Tomescu, M. L. 2015. Stable fuzzy logic control of a general class of chaotic systems, *Neural Computing and Applications* **26**(3): 541 – 550.

Quintero, O. and Paniagua, J. G. 2020. *From Artificial Intelligence to Deep learning in Biomedical Applications in Deep Learners and Deep Learner Descriptors for Medical Applications*, Springer International Publishing, pp. 253–284.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D. Y., Bagul, A., Langlotz, C., Shpanskaya, K. S., Lungren, M. P. and Ng, A. Y. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, *arXiv preprint arXiv:1711.05225* .

Ronneberger, O., Fischer, P. and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation, *in* N. Navab, J. Hornegger, W. M. Wells and A. F. Frangi (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Vol. 9351, Springer International Publishing, Cham, pp. 234–241.

Sendak, M. P., Gao, M., Brajer, N. and Balu, S. 2020. Presenting machine learning model information to clinical end users with model facts labels, *npj Digital Medicine* **3**(1): 1–4.

Stimper, V., Bauer, S., Ernstorfer, R., Scholkopf, B. and Patrick Xian, R. 2019. Multidimensional contrast limited adaptive histogram equalization, *arXiv preprint arXiv:1906.11355v3* .

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. and Liu, C. 2018. A survey on deep transfer learning.

Wadi, S. M. and Zainal, N. 2012. Contrast enhancement methods based on histogram equalization technique: Survey, *International Conference on Engineering and Built Environment (ICEBE)* .

Wang, L. and Wong, A. 2020. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106.

Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K. 2016. Aggregated residual transformations for deep neural networks.

Yock, P., Zenios, S., Makower, J., Brinton, T., Kumar, U., Watkins, F., Denend, L., Krummel, T. and Kurihara, C. 2015. *Biodesign: The Process of Innovating Medical Technologies*, Cambridge University Press.

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K., Ye, L., Gao, M., Zhou, Z., Li, L., Wang, J., Yang, Z., Cai, H., Xu, J., Yang, L., Cai, W., Xu, W., Wu, S., Zhang, W., Jiang, S., Zheng, L., Zhang, X., Wang, L., Lu, L., Li, J., Yin, H., Wang, W., Li, O., Zhang, C., Liang, L., Wu, T., Deng, R., Wei, K., Zhou, Y., Chen, T., Lau, J. Y.-N., Fok, M., He, J., Lin, T., Li, W. and Wang, G. 2020. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography, *Cell* **181**(6): 1423 – 1433.e11.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. 2015. Learning deep features for discriminative localization.