Submitted: 18 February, 2022 Accepted: 07 May, 2022 Published: 08 May, 2023 DOI:10.22514/sv.2022.070

Open Access
Signa Vitae

ORIGINAL RESEARCH

A real-world intelligent system for the diagnosis and triage of COVID-19 in the emergency department

Miguel Lastra Leidinger[1], Francisco Aragón Royón[2], Oier Etxeberria[2], Luis Balderas[2], Antonio Jesús Láinez Ramos-Bossini[3,4,*], Genaro López Milena[3,4], Liz Alfonso[5], Rosario Moreno[5], Antonio Arauzo[6], José M. Benítez[2]

[1]Department of Software Engineering, DiCITS Lab, iMUDS, DaSci, University of Granada, 18071 Granada, Spain
[2]Department of Computer Science and Artificial Intelligence, DiCITS Lab, iMUDS, DaSci, University of Granada, 18071 Granada, Spain
[3]Department of Information and Communication Technologies, Andalusian Health Service, 41001 Andalucia, Spain
[4]Department of Radiology, Hospital Universitario Virgen de las Nieves, 18014 Granada, Spain
[5]Biosanitary Institute of Granada (ibs.GRANADA), 18014 Granada, Spain
[6]Rural Engineering Department, DiCITS Lab, University of Cordoba, 14005 Cordoba, Spain

***Correspondence**
ajbossini@ugr.es
(Antonio Jesús Láinez Ramos-Bossini)

Abstract

The Coronavirus Disease 2019 (COVID-19) pandemic has had an unprecedented impact on healthcare systems, prompting the need to improve the triaging of patients in the Emergency Department (ED). This could be achieved by automatic analysis of chest X-rays (CXR) using Artificial Intelligence (AI). We conducted a research project to generate and thoroughly document the development process of an intelligent system for COVID-19 diagnosis. This work aims at explaining the problem formulation, data collection and pre-processing, use of base convolutional neural networks to approach our diagnostic problem, the process of network building and how our model was validated to reach the final diagnostic system. Using publicly available datasets and a locally obtained dataset with more than 100,000 potentially eligible CXR images, we developed an intelligent diagnostic system that achieves an average performance of 93% success. Then, we implemented a web-based interface that will allow its use in real-world medical practice, with an average response time of less than 1 second. There were some limitations in the application of the diagnostic system to our local dataset which precluded obtaining high diagnostic performance. Although not all these limitations are straightforward, the most relevant ones are discussed, along with potential solutions. Further research is warranted to overcome the limitations of state-of-the-art AI systems used for the imaging diagnosis of COVID-19 in the ED.

Keywords

COVID-19; Diagnosis artificial intelligence; Emergency care; Epidemiology; SARS-CoV-2

1. Introduction

The Coronavirus Disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has had an unprecedented impact on healthcare systems, with terrible consequences for the health and welfare of people worldwide. To date, nearly 200 million deaths due to COVID-19 and significant morbidity associated with the so-called 'long-COVID syndrome' have been reported [1]. No health care system was prepared for such an outbreak, and urgent measures and procedures were needed to face such a critical situation, especially in the early stages of the pandemic, when the lack of access to protection measures such as personal protective equipment [2], effective treatments [3], and triage protocols [4] was evident, particularly in low-income countries [5]. In this context, one of the most critical deficiencies in hospitals and health centers, which were overwhelmed by patients suffering from a number of symptoms requiring rapid diagnosis, was access to efficient diagnostic tests. As a consequence, many patients were managed on the basis of symptoms and imaging findings by computed tomography (CT) or chest X-ray (CXR).

Although the gold standard technique for a reliable diagnosis of SARS-CoV-2 infection is the reverse transcriptase-polymerase chain reaction (RT-PCR) test [6], the shortage of test kits in low-resource settings along with the need to wait several hours to obtain an accurate diagnosis remain factors limiting patient triage in the emergency department (ED). These limitations have led researchers to seek alternative screening solutions.

The rest of this section is organized as follows. First, how artificial intelligence (AI) can be applied on CXRs to detect COVID-19 infections is addressed followed by a review on the use of AI techniques for medical diagnosis.

1.1 The role of radiography and artificial intelligence in the diagnosis of COVID-19

One of the main effects of SARS-CoV-2 in the human body is a particular form of pneumonia. Thus, proper visualization of the lungs using medical imaging techniques (*e.g.*, CXR or CT)

This is an open access article under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

should reveal COVID-19 infection when there is pulmonary involvement. Accordingly, an automatic procedure to identify the particular pattern of this disease on a CXR seems an efficient way for patient triage in the ED. The advantages of CXRs include its speed, wide availability, accessibility, and portability. Therefore, automatic analysis of CXR images seeking specific imaging patterns would suffice to obtain a rapid diagnosis. The automation of this specific diagnosis can be approached by using AI systems, the so-called 'intelligent systems'.

One of the most outstanding fields in which AI has been applied in the last decade is computer vision, and the diagnosis of COVID-19 from CXR images seems a proper task to be tackled using this approach. In fact, many groups around the world have explored this idea, and several articles have been published on the subject to date [7–10]. Unfortunately, the quality of reported research is quite diverse and most proposals cannot be reproduced due to several issues, including lack of data availability, missing details in the description of processes or parameters, or incomplete descriptions of the methods followed. Hence, it is very difficult to build an intelligent system for this endeavor solely based on the information published in the scientific literature.

1.2 Artificial intelligence for medical diagnosis

The use of AI in medical diagnosis began with the use of expert systems. Instead of designing an 'all-in-one' algorithm, expert systems combine knowledge with a general algorithmic problem-solving method to reach a required solution, *e.g.*, a diagnosis in medical problems. By using this architecture, expert systems provide transparency (explain their solutions), flexibility (knowledge can be improved), user-friendliness (no programming skills are needed), and competence (high capability of solving problems) [11]. Expert systems raise the work of a medical expert to a new qualitative level, reducing the risk of misdiagnosis [12].

Machine learning (ML) techniques focus on increasing knowledge through automated application of statistics. The underlying idea is to make the computer learn from data. Since large amounts of medical data exist, new knowledge can be acquired and applied in medicine. For instance, in prognostics, an ML model can learn the patterns of health trajectories of a vast number of patients and this can help physicians to anticipate the course of a given condition [13]. Regarding disease detection and diagnosis, the utilization of several ML artifacts, like Bayesian networks, K-means clustering or Support vector machines, has proved to be very successful [14].

Artificial neural networks (ANN) are part of ML techniques. These networks try to mimic the structure of the brain and are formed by a set of layers of neurons. Each neuron is connected to some of the neurons of the previous layer (in fact, to all of them if it is fully connected) and produces an output using a weighted sum of its inputs (which are the outputs of neurons in the previous layer). These weights are the model parameters to be adjusted during the training phase. The larger the number of neurons, the more parameters have to be learned, especially in the case of fully connected networks, where each neuron is connected to all the neurons of the previous layer.

With the recent advances in parallel computing, the capabilities of ANNs have improved significantly, and larger networks with many layers of artificial neurons are built and trained. Deep Learning (DL) studies these networks along with all the algorithms that work with them [15]. There are many applications of DL in medicine, but the most common ones concern medical image analysis [16]. Traditional ML techniques can achieve results in the range of small-sized ANNs, although it is difficult to reach the performance of state-of-art DL networks in the field of image analysis [17]. Such performance levels are obtained at the cost of computationally intensive model creation processes.

All these ML techniques address different problems. One of them is classification, *i.e.*, the process of assigning a class to each object under consideration. These objects are usually described in data terms by the value of some of their features (*e.g.*, size, weight, age, color). However, sometimes the available data about the object is an image, described by a matrix of the color value of each pixel (points of the image). This is called 'image classification', which should not be confused with object detection or image labeling, where more than one label is assigned to a given image. Medical image classification aims to differentiate different types of medical images according to a certain criterion, such as clinical conditions [18].

Many methods have been proposed to classify images. Image classification systems usually require extensive preprocessing to prepare images for analysis, including edge detection, color feature extraction or normalization [19]. With the advent of DL, some of this preprocessing is addressed directly within the network layers [20]. Auto-encoders and convolutional ANNs address this preprocessing, being able to improve image classification and obtain striking results. This automatic annotation process has been used for a wide variety of medical image classification analyses, including histopathological classification of colon cancer [18] or skin cancer [21].

Despite a very high number of papers have been published presenting different attempts to create automatic systems for image detection of COVID-19 [10, 18, 22], unfortunately, most of them suffer from several weak points and usually fail to provide enough information to allow for their reproduction [23]. In this article, we strive to provide not only a comprehensive description of the process followed to build a robust and effective system for that purpose, but also to make it easy to understand and fully reproducible. This article reports the work performed, as well as the results obtained by our research team, and provides all the necessary information to allow interested researchers to reproduce all the work carried out. It also discusses the issues we have run through and provides hints to avoid or mitigate them.

The diagnosis of COVID-19 can be approached as a normal data science problem: it can be formulated as a classification problem and different models can be built by applying any of a diverse set of possible ML techniques. However, a rigorous process involving data pre-processing and many technical details need to be undertaken to eventually achieve an effective system. Such process can be structured in the

following steps: problem formulation, data collection and preprocessing, model definition and construction, validation, and, finally, deployment.

2. Methods

This section presents detailed formulation of the problem being tackled along with the computational tools that were applied to develop a solution. Moreover, the image preprocessing techniques that were used and assembled into a preprocessing pipeline and the neural networks used to build the diagnosis system are explained.

2.1 Problem formulation

The goal of the present work was to develop an automatic diagnostic system for COVID-19 based on CXRs. The gold standard diagnostic test for COVID-19 is the RT-PCR test, but several factors may hinder its application, including the shortage of test kits or the time required to obtain the result. Since one of the main effects of COVID-19 is lung involvement, a faster diagnostic approach is possible by analyzing CXRs. If an automatic procedure without the intervention of a radiologist could be developed, the approach would take shorter times and demand fewer resources.

The resulting system would take a CXR as the only input and feed it to a computer system which, in turn, would provide an answer in terms of whether the patient is affected by COVID-19 or not. The diagnostic challenge boils down to an image classification problem. However, the task is not that simple and several issues need to be addressed:

• Enough data to train the AI system are required, which entails having a large amount of correctly labeled CXRs.

• The system should properly distinguish COVID-19 lung involvement from other types of effects produced by other diseases.

• Real-world CXRs are not perfect. They may suffer from several technical deficiencies that must be addressed before images are processed.

The starting point for the model we need relies on recent ML-based solutions for problem classification. These systems have been built out of the last generation of ANNs, an extensive set of models collectively known as DL models [24].

DL generically refers to the whole area, including models, architectures, training algorithms and roughly anything related to them. The most relevant early models for images were built by Google researchers on the ImageNet dataset [25]. ImageNet is a research project that has developed a large database of images with diverse annotations. The dataset contains over 14 million images (out of 1 million annotated) for 21,000 classes. This dataset also served as the basis for an image classification challenge from which several related but different models have been developed using the dataset as a starting point. Some of the most notable models are VGG16-VGG19 [26], MobileNet [27], ResNet [28], Inception [29], DenseNet [30] or Xception [31].

Creating a brand new model for image classification imposes a high tax in terms of necessary data and resources. These requirements are usually hard to satisfy. However, the knowledge extracted by the networks along their training is embedded within their structure (both topology and weight values). Most of their general knowledge can be useful for other classification problems. Based on this premise, a neural network able to address other classification problems can be built. This requires making some adaptation to the network topology, particularly, input and output layers, and a relevant dataset.

2.2 Data collection and preprocessing

The availability of a high-quality dataset is a key ingredient in any data-driven solution. Research groups all over the world have been gathering their own datasets as the COVID-19 pandemic evolved. However, only a few have been published for general use. To face the challenge posed by the present research work, two datasets have been considered: a publicly available one and a locally gathered one.

During the pandemic time, a few repositories with CXRs (or other kinds of radiological images) have been published on the Internet. However, the quality of the data or the representativeness of the sample hinder their final effectivity. They can be used for exploring purposes, but not to develop a robust real-world system. Among all the public repositories, the one that has probably received the most attention is COVIDx from the COVID-Net Open Source Initiative [32]. COVIDx is a dataset of CXR images that comprises 15,190 samples. It has been generated by combining five different publicly available data repositories. The images in the dataset are labeled and therefore divided into three groups: “Normal”, “Pneumonia” and “COVID-19”. For the purpose of our problem, we aggregated the first two classes into a single one (“Non-COVID”). The distribution for each group are: Non-COVID-19 CXR images (91.5%) and COVID-19 CXR images (8.5%). This is the original version of the dataset, as published in [32]. We will refer to it as COVIDx 1. The dataset, nevertheless, has been updated in successive versions. The last one was published in November 2021. It is now composed of ~30,000 images divided into two categories (Non-COVID-19 and COVID-19), extracted from several health institutions all over the world. The percentages of the two classes are 14,192 COVID-19-negative CXR images and 16,690 COVID-19-positive CXR images.

The second dataset to consider was collected locally in the province of Granada, Andalusia, southern Spain. It is composed of 114,251 anonymized Digital Imaging and Communication in Medicine (DICOM) files from patients that attended several public hospitals in the province of Granada. More specifically, these hospitals were: Hospital Universitario Virgen de las Nieves of Granada, Hospital Comarcal Santa Ana de Motril and Hospital Comarcal de Baza. These DICOM files contain, among other information, high resolution (4000 $\times$ 3000 pixels) CXR images. We will refer to this dataset as GranaCov.

Since CXR images are not in perfect condition, a thorough cleaning and preprocessing step was necessary, and a subsequent filtering process had to be applied. The images can be both lateral and frontal radiographs showing the lung area. During the extraction of images from the DICOM files,

only the frontal images were selected, discarding the lateral ones, resulting in a total of 68,360 frontal images. Moreover, because labels (frontal, lateral) assigned to the images were not completely consistent, a separate image classification neural network was trained to distinguish between frontal CXR images from the rest. This not only allowed us to discard images that showed a lateral view (*i.e.*, mislabeled), but also some other images not related to this project which were erroneously included in the dataset. The accuracy level obtained on this specific task was 98.6%. This model automatically filtered out images of children, mislabeled images, and images that do not correspond to the lung area, such as arms, hips or undefined objects. After this process, the total number of available frontal images was 64,790.

From the filtered frontal images, due to the poor quality of some CXRs, a process of image quality improvement was necessary. In this regard, two types of problems were identified: some images with a white background presented inverted color tones and some were excessively saturated. The images with the white background were improved by inverting the colors, while the saturated ones underwent a process of histogram equalization to increase contrast. Both types of problems are automatically detected by our preprocessing pipeline and do not require human intervention. Fig. 1 shows examples of these cases.

The labels for the CXR images were obtained from an anonymized hospital database. In this database, each patient can have more than one associated CXR and the COVID-19 infection identification criterion is based on PCR test results. It contains 40,751 records of PCR, Immunoglobulin G (IgG) and Immunoglobulin M (IgM) antibody tests (antibody tests were not used in this study). On this subject, we encountered several problems as we learned that the database also contained CXRs associated with previous, non-COVID-19 related, episodes. This required another filtering process to select only images acquired during the same episode in which the PCR test outcome was obtained. Once these problems were solved, the database for the study was limited to 6569 COVID-19 negative and 1267 COVID-19 positive images, yielding a final dataset with 7836 instances (Table 1).

TABLE 1. Class distribution of GranaCov instances.

COVID-19 Negative	COVID-19 Positive
6569	1267

COVID-19, coronavirus disease 2019.

2.3 Base Convolutional Neural Networks

A review of the scientific literature reveals that Convolutional Neural Networks (CNN) are generally used to address image classification problems. These networks are a special type of neural network where each layer performs a convolutional operation (or set of operations) on the input image or the result of the previous layer. In contrast to neurons of regular fully connected networks, convolutional neurons perform a convolutional operation (whose associated weights have to be learned during the training phase) only on a region of the input image or the result produced by the previous layer. This reduces the number of parameters to be estimated in comparison to a fully connected network applied to the same type of data. It has to be noted that the dimensionality of inputs like images is very high even for low-resolution input samples as each pixel is an input element. If no additional classification model is used, the last layers of CNNs are fully connected ones that are only geared to produce the output, *i.e.*, a classification label. In summary, the layers of a CNN are trained to learn a set of convolutional filters which should be helpful to extract relevant features for the classification task.

CNNs are not usually built from scratch but rather pre-trained networks are employed. To this end, Transfer Learning approaches are used. Transfer Learning provides the possibility to re-use the previous parameter values acquired by training one network and adapt it to another—similar—problem. The knowledge acquired by the network is, thus, re-used to solve a different problem. The reasoning behind this procedure is that this type of neural network, during the training phase, learns a set of filters that are useful for extracting features from any image and that these filters can be re-used, although a fine-tuning step is always required (network re-training step). In other words, a neural network that can classify regular images can be re-engineered to classify CXR images. Therefore, it is a good starting point and helps to save a great amount of computing time.

After some initial tests with different models, MobileNetV2 [33] and VGG16 [26] were the state-of-art neural network models chosen as the base to build our diagnosis system. Trained versions of these models are available to download from the official TensorFlow site[1,2] (version core 2.8.0 at the time of writing). VGG16 is a representative of models with a very large number of parameters (~138 million). The network is composed of five convolutional blocks. The first two blocks consist of two convolutional layers and a Pooling layer ($224 \times 224 \times 64$ and $112 \times 112 \times 128$), while the others consist of three convolutional layers ($56 \times 56 \times 256$, $28 \times 28 \times 512$, $14 \times 14 \times 512$). Three Fully-Connected layers of different depths (4096, 4096, and 1000) follow the stacked convolutional blocks. The final layer is a soft-max layer.

MobileNetV2, with ~3.5 million parameters is a network of smaller size. The architecture contains an initial convolutional layer ($224 \times 224 \times 3$), followed by seven bottleneck blocks ($112 \times 112 \times 32$, $112 \times 112 \times 16$, $56 \times 56 \times 24$, $28 \times 28 \times 32$, $14 \times 14 \times 64$, $12 \times 12 \times 96$, $7 \times 7 \times 160$), to finish again with two convolutional layers ($7 \times 7 \times 320$, $1 \times 1 \times 1280$). The same parameters have been used as in the previous model. Both models had already been pre-trained using the ImageNet dataset with images of size 224×224.

[1]https://www.tensorflow.org/api_docs/python/tf/keras/applications/mobilenet_v2/MobileNetV2 (Accessed February 2022)

[2]https://www.tensorflow.org/api_docs/python/tf/keras/applications/vgg16/VGG16 (Accessed February 2022)

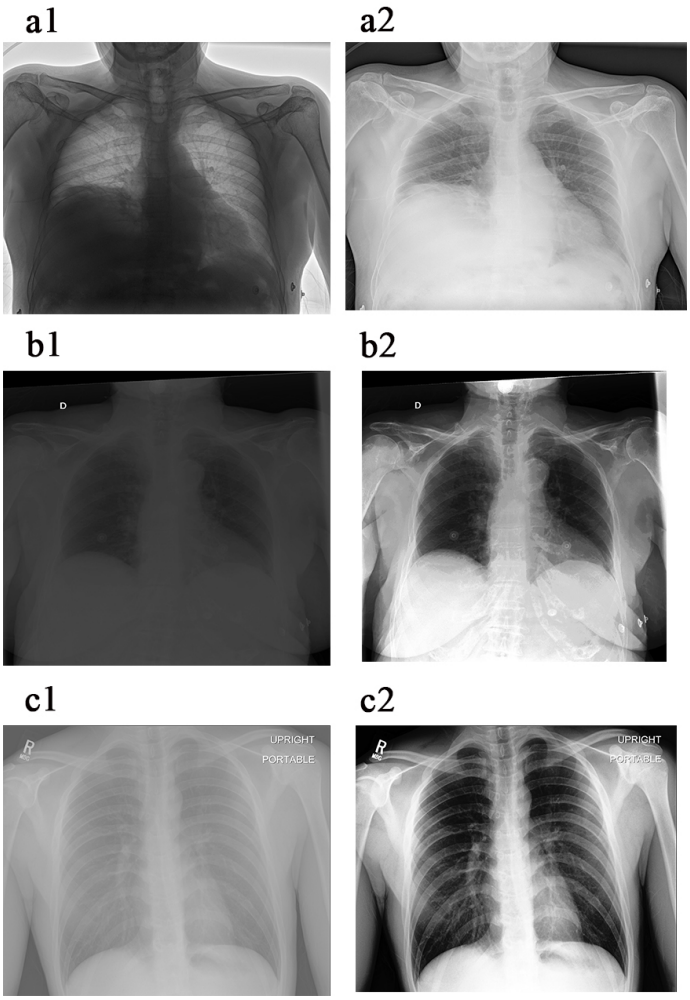

FIGURE 1. Image preprocessing examples of chest X-rays (CXR). a. Color inversion. a1. CXR with white background (*i.e.*, negative). a2. Result from color inversion. b and c. histogram equalization. b1. Black saturated CXR (*i.e.*, overexposure). b2. Histogram equalization. c1. White saturated CXR (*i.e.*, underexposure). c2. Histogram equalization.

2.4 Building of the network

As mentioned above, the networks used are not trained from scratch, but rather a Transfer Learning approach is used. Our starting points were ImageNet pre-trained networks. From the different choices, VGG16 and MobileNetV2 were the models finally selected. The original formulations of the model had to be adapted to the task of classifying CXR images to detect COVID-19 lung involvement. Accordingly, the last two layers were removed and some new trainable layers were added on top of the remaining ones: a dense layer with 1024 units, a dropout layer with 0.5 rate of units to drop and, finally, a new dense layer, as it is completely dependent on the number of classes. In the original problem, the number of classes was over 1000, but in our case, it is just two.

The detailed architecture of the VGG16 network is provided (Fig. 2). The first two layers are convolutional, with 64 3 $\times$ 3 filters (per layer). The stride value used is 1. The convolutional filter size and the stride value are maintained constant for the whole network. The third layer is a max-pooling layer of size 2 $\times$ 2 with a 2 $\times$ 2 stride. This layer down-samples input data computing the maximum over a 2 $\times$ 2 window. The fourth and the fifth layers are also convolutional, with 128 filters (per layer). The sixth layer is another max-pooling layer of the same type as the preceding one. The next three layers are convolutional, with 256 3 $\times$ 3 filters. Max-pooling is applied by the tenth layer. The previous scheme (three convolutional layers followed by a max-pooling one) is repeated twice with 512 filters per convolutional layer. Finally, the data is flattened and passed to two fully connected layers of size 4096, each followed by a dropout layer. Dropout layers randomly set some of the inputs to 0 to prevent overfitting.

As the last layer, a fully connected one is used to produce the classification result. Its output size depends on the number of classes of the classification problem. This layer uses the softmax function to produce a result in the (0, 1) range associated with each possible class. The sum of all outputs of a layer of this type is equal to 1.

The architecture of the VGG16 net can be summarized as follows: a series of convolutional layers followed by a max-pooling one. The dimension of the data gets progressively reduced while the number of filters per convolutional layer is progressively increased before data reaches a pair of fully connected layers and the final softmax one.

Regarding domain adaptation, since our datasets were not large enough for these networks to learn, we used various openly available CXR image datasets. The first dataset used

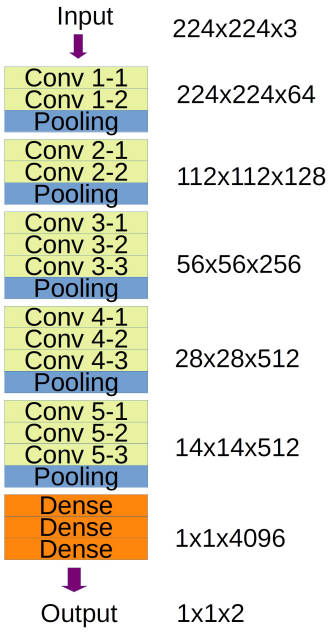

FIGURE 2. Schematic layer representation of a VGG16 neural network used for classifying CXR images of COVID-19. The type and size of each layer is included. Conv, convolutional.

was the one made available to the scientific community by the National Institute of Health (NIH) Clinical Center [34]. The dataset contains more than 130,000 anonymized CXR images with 14 categories of lung diseases. Remarkably, it does not contain any COVID-19 cases. Nevertheless, it serves the purpose of letting the network discover features specific to CXR images. The networks trained on this CXRs of non-COVID-19 images were then further trained to build the actual models that would diagnose suspected COVID-19 patients.

The tasks performed for model adaptation include the following:

• Network architecture tuning: light adjustments to the number of layers/blocks to learn how the network size was related to the results.

• The pre-trained layers were frozen to allow only the initial and last layers to be modified during the re-training phase to keep intact most of the pre-learned features. This test was carried out to account for the possibility that the number of CXR images was not large enough to re-train networks of this size.

• Different parameters related to the re-training phase were tested, *e.g.*, optimizer method or learning rate.

• Data augmentation: the dataset was artificially modified by adding modified versions of the original data. These modifications included horizontal and vertical flips, zoom and up to 20 degrees of rotation.

Then the networks were trained on the COVIDx 1 dataset since this was the only one available at the early stages of the research. By following the established empirical methodology, the dataset was randomly split into two independent parts: training set (70%) and test set (30%). Only the training set was used in the model building tasks, including hyperparameter selection identification, reserving the test set for the exclusive task of model testing. The COVIDx 1 dataset showed a clear class imbalance problem as the number of COVID-19 negative samples was clearly higher than the positive ones. This required the use of the well-known majority class (*i.e.*, COVID-19 negative) under-sampling technique. By having a similar number of samples of both classes, the learning process does not get biased towards mostly generating majority class predictions because it is sufficient to achieve global accuracy levels.

The employed training procedure used the following parameters: as an optimizing procedure, both Stochastic Gradient Descent (SGD) and Adam were tested. SGD always yielded higher accuracy values. After some initial test-and-error experiments, the remaining parameters were set as learning rate at 0.001, decay at 0.000001, and momentum at 0.9. The network performance is measured based on its accuracy, defined as the rate between right answers and the total number of instances. A deeper insight is offered by the confusion matrix: each row represents the instances of each actual class while each column represents how the test samples of each class were labeled by the classification system.

3. Results

This section presents the experimental results as well as the latter validation procedure. Finally, a description of the application developed to deploy the diagnosis model is included.

3.1 Experimental results

While data augmentation was applied at different intensification levels, no improvement in the results was observed. Thus, the process was eventually not applied, and only the original set was used for training.

The results obtained in terms of classification accuracy for the VGG-16 and MobileNetV2 were 92% and 91%, respectively (both on the COVIDx 1 test set). The confusion matrices are shown in Table 2 and Table 3.

TABLE 2. Confusion matrix obtained using VGG16 on the COVIDx 1 dataset.

Actual \ Predicted	COVID-19 negative	COVID-19 positive
COVID-19 negative	90	10
COVID-19 positive	7	93

F1-scores for both networks were 0.9163 and 0.91282, respectively, pointing to a lightly better performance of VGG16 over MobileNetV2.

COVIDx dataset has been updated several times, the last one in November 2021. This version is composed of 30,000 CXR images with 16,690 positive cases, the rest being negative. We named this updated version as COVIDx 2. We reproduced the same procedure applied to COVIDx 1 with this enlarged version of the dataset. On this occasion, VGG16 achieved a

TABLE 3. Confusion matrix obtained using MobileNetV2 on the COVIDx 1 dataset.

Actual \ Predicted	COVID-19 negative	COVID-19 positive
COVID-19 negative	94	6
COVID-19 positive	11	89

TABLE 4. Confusion matrix obtained using VGG16 on the COVIDx 2 dataset.

Actual \ Predicted	COVID-19 negative	COVID-19 positive
COVID-19 negative	194	6
COVID-19 positive	29	171

TABLE 5. Confusion matrix obtained using MobileNetV2 on the COVIDx 2 dataset.

Actual \ Predicted	COVID-19 negative	COVID-19 positive
COVID-19 negative	190	10
COVID-19 positive	9	191

91% accuracy rate in the test, while MobileNetV2 showed 95% accuracy. The confusion matrices are shown in Table 4 and Table 5.

F1-scores for both networks were 0.9072 and 0.9526, respectively, resulting in a better result for the simpler architecture of MobileNetV2.

The experiments have been replicated by using different random partitions of the dataset into training and test sets. The accuracy values and confusion matrices obtained are quite similar to those reported above, with a standard deviation of 2.0 units of accuracy level.

Our final goal was to have a system tailored to the patients undergoing a CXR with suspicion of COVID-19 in the province of Granada. Thus, a local CXR image dataset was needed. The gathering of the local dataset was severely delayed due to administrative and technical reasons, but after one year the dataset, GranaCov, described in section 3.2, was assembled.

Once the dataset was available, we made a straight evaluation of the COVIDx trained network on the GranaCov dataset. Since the two datasets (*i.e.*, COVIDx and GranaCov) are assumed to represent the same problem and to be samples of the same phenomenon, similar results, in terms of accuracy were expected. Surprisingly, the results on the GranaCov dataset were quite different to the ones obtained on COVIDx. This unexpected issue triggered the search for explanations. The following causes were identified:

• COVID-19 positive cases with no lung involvement. In this case, no information can be obtained from the CXR image. A review process conducted on our database has made this issue surface.

• Doubtful or interim cases in which even trained radiologists find it difficult to reach a diagnosis with no additional information (*i.e.*, only using CXR information).

• Mislabeled images due to errors produced when the information was added to the database. The unprecedented admission rates, especially during the first months of the pandemic, increased the chance of human errors of this kind.

To address the first of the above points, a model capable of discerning between lung-involved and non-lung-involved images was generated, so that it could be used to filter out CXR images of COVID-19 patients with no lung involvement, that is, images that may be hindering the development of a model. The COVIDx dataset was used for this purpose. Instead of using the COVID/Non-COVID labeling of the data, we opted to label CXRs as normal-pneumonia/COVID, *i.e.*, joining the last two classes into a single one. The generated model achieved a 91% test performance for the task of classifying between pulmonary versus non-pulmonary-involved CXR images. The application of this model to the local dataset produces some interesting outcomes: it establishes 3557 images labeled as Non-COVID as being pulmonary-involved cases, and 444 COVID-19 positive images as unaffected cases.

Analogously, a segmentation process was applied to optimize the training process as non-relevant parts of the images are eliminated and substituted with a solid black background color. For this task, U-net, which is a pre-existing CNN designed for biomedical data segmentation was used [35]. This model was trained using pre-segmented CXR images and achieved 98% accuracy on our images.

The output of the U-net is a binary mask of the same size as the input image. The application of this mask to the original image results in any pixels not associated with the lung area being transformed into black pixels (background). Therefore, U-net is only used during the pre-processing step to remove pixel data not belonging to lung areas and it is not used to obtain a COVID-19 diagnosis. Fig. 3 shows a sample of the image obtained from the DICOM file and the resulting segmented image using U-net.

FIGURE 3. Example of CXR image segmentation using U-net.

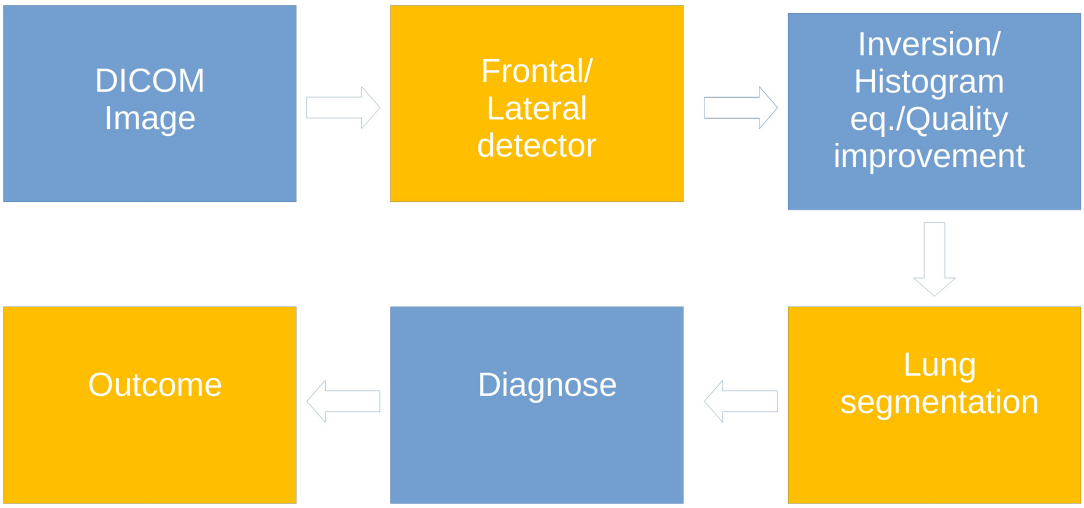

FIGURE 4. Intelligent diagnosis system blocks. DICOM, Digital Imaging and Communication in Medicine.

The segmentation process avoids any learning effort to be wasted outside the area of interest and does not require any human intervention, except a small percentage of very low-quality images that cannot be repaired using the image preprocessing pipeline explained in this section.

All the adjustments performed on the processing pipeline were geared towards achieving competitive performance on the GranaCov dataset too. The steps composing the currently deployed processing pipeline are depicted in Fig. 4.

3.2 Validation

The validation process of all the steps explained in previous sections is based on the use of independent training and test sets. The error or accuracy achieved in the training set is useful for understanding how the model is learning, but it is not a realistic estimate of how the model behaves in the case of new observations. To achieve a more accurate estimate, it is necessary to use a previously 'unseen' dataset, the test set. The model validation strategy adopted in this work consisted of dividing the dataset into training and test subsets, and additionally into a validation subset. The validation set was used to choose the best training model.

All the accuracy values included in this work refer to the behavior of the model on the test set, not on the training or validation set. The training and validation sets used during the learning process contain 70% and 10% of the images, respectively. The remaining ones are used in the test set to classify using the trained model and evaluate the model performance.

The accuracy of every trained network described above on the training set was over 99%, meaning perfect learning on the instances used for the network weights adaptation. However, the most relevant figure is the accuracy on unknown instances, totally independent of those used during training, namely, the test set. As previously indicated, the average performance of VGG16 models on both COVIDx 1 and COVIDx 2 was 91%, and the performance of the MobileNetV2 was 92% and 95%, respectively. These percentages are average values over several independent partitions of the dataset, so they are a valuable approximation of the real values. On the other hand, the performance on the GranaCov dataset in training was again over 98%. However, the performance observed on the test set was rather low. Thus, further research is necessary before an intelligent model reliable enough is available.

Once the models are trained and validated, they are tested for some time in medical practice. The intention of this is to validate the models in a real-world environment (Emergency Department) by medical professionals. The workflow and a web application have been developed to allow its implementation but have not been implemented in clinical practice yet.

3.3 Deployment of the final system

The AI-based diagnosis system was fully implemented using the Python programming language using Tensorflow with Keras for all the DL related tasks as stated previously.

In order to make the use of the developed tool easier for physicians, a user-friendly web application has been designed and implemented. The architecture of the tool is quite straightforward, including the image database, the intelligent system model and the user interface. This application has been developed using the JavaScript and PHP programming languages. The image database is stored in a Redundant Array of Independent Disks (RAID) storage device, with an ext4 file system. The database is managed by PostgreSQL. All the components of the application are deployed with containers—through docker services—scattered along with several servers with Linux based operating systems. Therefore, only open-source software is used.

The workflow supported by the application is quite straightforward: as new patients arrive at the Emergency Department, their data are captured, a CXR is obtained and, once it is available, the physician retrieves the patient's data and requests a diagnosis from the tool. Once the image is analyzed and the result is available (it takes less than one second), the latter is displayed on the web page. This way, the application can be used from any device with an Internet connection, namely universally accessible. This workflow is illustrated in Fig. 5. As the initial motivation for the usage of these intelligent

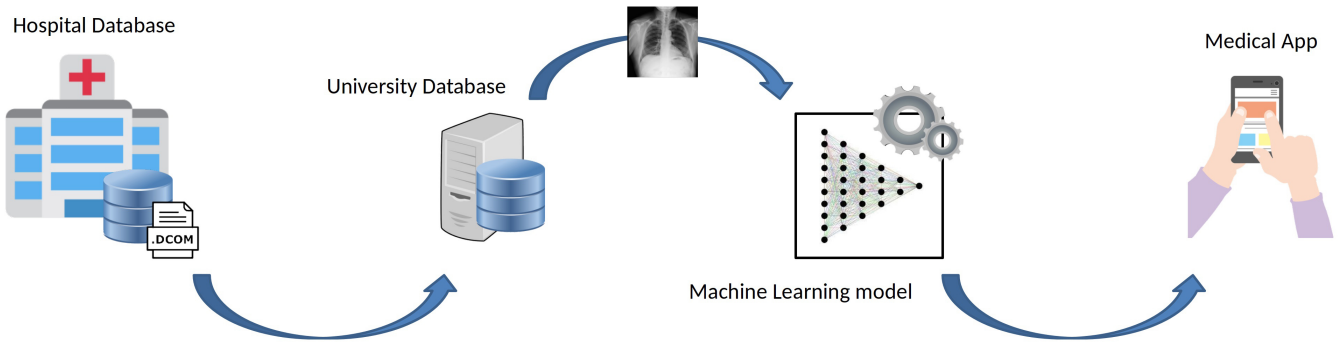

FIGURE 5. Medical diagnosis application. DICOM, Digital Imaging and Communication in Medicine. App, application.

systems was a fast triaging system, measurements of running times are necessary. The most computationally intensive task is the image analysis step, which takes an average running time of 0.75 seconds per image. This is not the limiting time for the whole process, but the acquisition of the CXR, which is the longest step. A screenshot of the current version of the application as used during physician evaluation is shown in Fig. 6.

4. Discussion

The work described in this paper includes both a scientific and engineering process. From the scientific perspective, we have addressed the problem of building a computer-based system to identify patients possibly affected by COVID-19 from their CXR images. The supporting hypothesis is that if a patient has been infected by COVID-19, then his or her lungs suffer from a specific type of pneumonia and that it can be identified from the image of the lungs. A program developed with ML techniques can be tuned to extract and identify relevant signs of that infection and, thus, provide a reliable radiological diagnosis.

The idea seemed promising, and a number of research groups scattered around the globe have put it into practice. Unfortunately, it is not completely straightforward, and several technical issues need to be thoroughly addressed. The available literature on research projects conducted on this topic is extensive. However, most of the published articles suffer from several weak points. A recent systematic review of the literature performed by Wang *et al.* [36] (2021) identified common pitfalls in the research reported in over 2200 papers.

Due to the limitations of previous papers on the topic, we intend to build a reliable and robust intelligent system for COVID-19 screening. While the idea is easy to catch on, it is not completely straightforward. We have gone through a rigorous process that has been thoroughly described in this paper and have been able to build a fully operational system that assists physicians in COVID-19-suspected patient triage. While other publications with a similar objective usually omit details—both major and minor—, we provide detailed information allowing for reliable reproduction of the scientific and engineering procedures followed in this project. This contribution is a major highlight of this paper.

Another important fact, which is barely mentioned—let alone recognized—in the literature is that, while the effects of SARS-CoV-2 are expected to be equal all over the world, CXR image datasets assembled from different populations might not be from the same—statistical/epidemiological—population. That is, a screening system performing well on a given dataset may not perform so well on a different dataset. Thus, an effective data engineering procedure is required to allow the creation of adapted versions of the models. This preprocessing step is held responsible for over 60 to 70% of the time in data science tasks. Detailing the actual problem faced during this stage is of the highest relevance for scientists and engineers when their goal is to develop a real-world intelligent system.

5. Conclusions, limitations and future research

As the COVID-19 pandemic has stormed all over the world with devastating effects on the health of the population, fast tools for screening COVID-19-suspected patients are needed. One promising approach is to develop a system based on CXR image analysis to assess whether the lung involvement pattern is compatible with that caused by SARS-CoV-2. This hypothesis has been analyzed by a host of researchers reporting varying degrees of success. Accordingly, a project to build a robust system is on execution by our research team. While it is not over yet, interesting results have been obtained and are described in this article.

The first relevant conclusion is that indeed it is feasible to develop an effective intelligent system based on ML and computer vision techniques to automatically diagnose COVID-19 from CXR images. Starting from well-known image classification neural networks, we have developed a system that achieves a performance in line with state-of-the-art published results. With an average performance of 93% success, the system is rather reliable on CXR of the same population. This has been reached using a publicly available CXR dataset.

A reliable and fully operative system has been developed, endowed with a friendly user interface—a web application. The average running time of the image analysis is 0.75 s, so it greatly reaches the target of becoming a fast screening system.

Next, it is not straightforward to adapt a particular system to a different population. A full understanding of the causes is not yet available, but different patient populations may

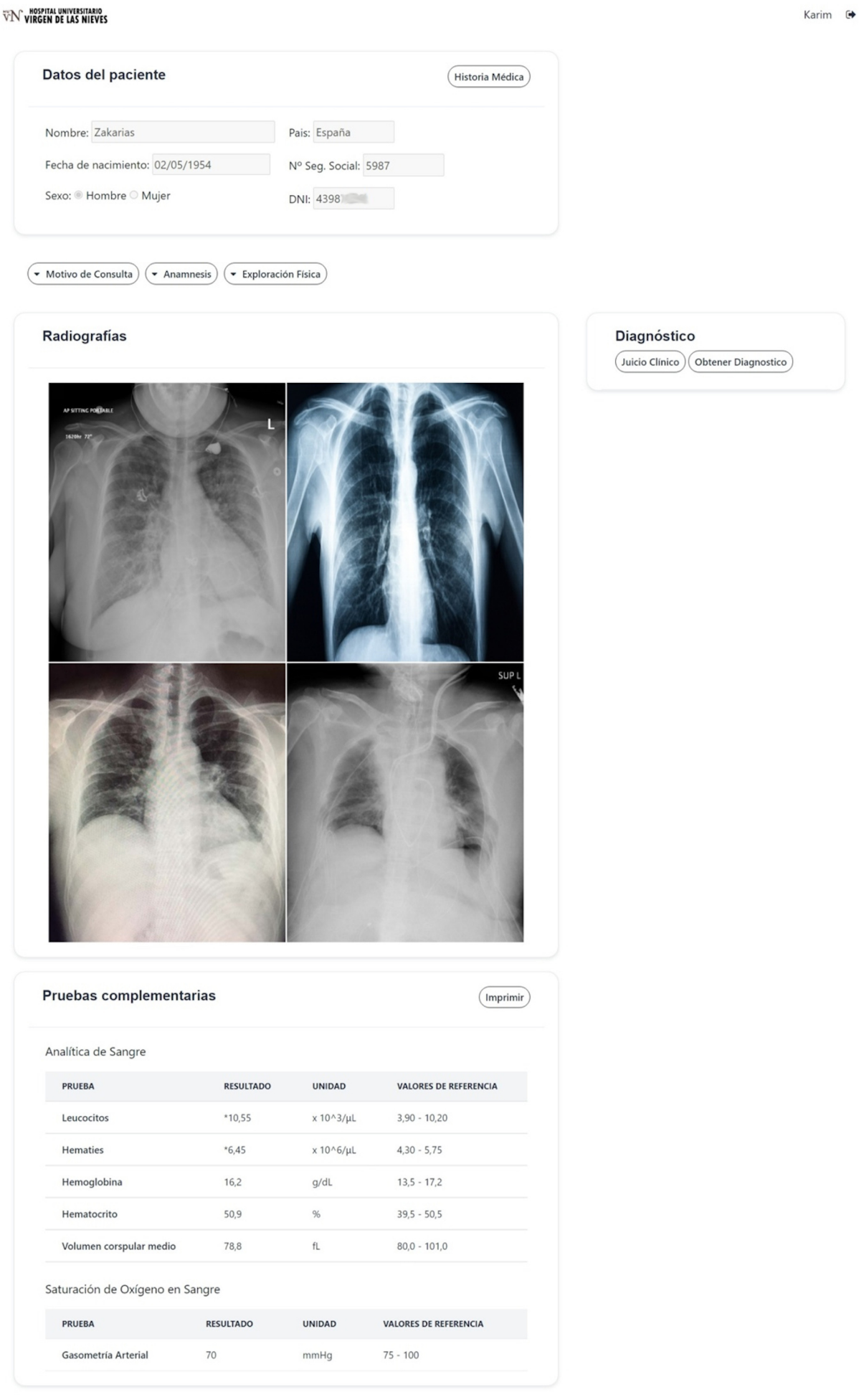

FIGURE 6. Diagnosis application screenshot.

produce very different CXR datasets. This leads to a data drift situation and trained networks on a specific dataset might not be effective on different datasets for this problem.

When seeking to adapt the system to a different dataset, a number of pitfalls arise that require careful analysis. The analysis of the dataset has surfaced different properties of the data so images require careful preprocessing. Several filtering and transformation steps have been developed—*e.g.*, segmentation, histogram equalization, inversion—which has led to enhanced quality data. Unfortunately, a completely satisfactory solution has not been found yet and further research is needed. The future research will therefore focus on increasing the performance of our system on the GranaCov dataset to match the results obtained on COVIDx.

AVAILABILITY OF DATA AND MATERIALS

The data presented in this study are available on reasonable request from the corresponding author.

AUTHOR CONTRIBUTIONS

MLL, JMB and AA—designed the research study. MLL, FAR, OE and LB —performed the research. GLM, LA, AJLRB and RM —obtained and reviewed the data. MLL, FAR, OE, LB, AA and JMB—analyzed the data. MLL, AJLRB and FAR—wrote the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study complies with the principles of the Declaration of Helsinki. A completely anonymized database was used for the analyses. The need for informed consent has been waived owing to its retrospective nature. The study was approved by the Provincial Research Ethics Committee of Granada (code IACovid).

ACKNOWLEDGMENT

Not applicable.

FUNDING

This work was supported by the 'Artificial Intelligence for the diagnosis and prognosis of COVID-19' project (CV20-29480), funded by the Consejería de Transformación Económica, Industria, Conocimiento y Universidades, Junta de Andalucía, and the FEDER funds.

CONFLICT OF INTEREST

The authors declare no conflict of interest. Antonio Jesús Láinez Ramos-Bossini is serving as one of the Guest editors of this journal. We declare that Antonio Jesús Láinez Ramos-Bossini had no involvement in the peer review of this article and has no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to VL.

REFERENCES

[1] Rivera-Izquierdo M, Láinez-Ramos-Bossini AJ, de Alba IG, Ortiz-González-Serna R, Serrano-Ortiz Á, Fernández-Martínez NF, *et al*. Long COVID 12 months after discharge: persistent symptoms in patients hospitalised due to COVID-19 and patients hospitalised due to other causes—a multicentre cohort study. BMC Medicine. 2022; 20: 92.

[2] Hakim M, Khattak FA, Muhammad S, Ismail M, Ullah N, Atiq Orakzai M, *et al*. Access and use experience of personal protective equipment among frontline healthcare workers in Pakistan during the COVID-19 emergency: a cross-sectional study. Health Security. 2021; 19: 140–149.

[3] Rivera-Izquierdo M, Valero-Ubierna MDC, R-delAmo JL, Fernández-García MÁ, Martínez-Diz S, Tahery-Mahmoud A, *et al*. Therapeutic agents tested in 238 COVID-19 hospitalized patients and their relationship with mortality. Medicina Clinica. 2020; 155: 375–381.

[4] Rivera López E, Abal F, Rekers R, Holzer F, Melamed I, Salmún D, *et al*. Proposal for the elaboration of a triage guideline in the context of the COVID-19 pandemic. Revista de bioetica y derecho & perspectivas bioeticas. 2020; 2020: 37–61.

[5] Alhalaseh YN, Elshabrawy HA, Erashdi M, Shahait M, Abu-Humdan AM, Al-Hussaini M. Allocation of the 'already' limited medical resources amid the COVID-19 pandemic, an iterative ethical encounter including suggested solutions from a real life encounter. Frontiers in Medicine. 2021; 7: 616277.

[6] Rivera-Olivero IA, Henríquez-Trujillo AR, Kyriakidis NC, Ortiz-Prado E, Laglaguano JC, Vallejo-Janeta AP, *et al*. Diagnostic performance of seven commercial COVID-19 serology tests available in South America. Frontiers in Cellular and Infection Microbiology. 2022; 12: 787987.

[7] Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, *et al*. Can AI help in screening viral and COVID-19 pneumonia? IEEE Access. 2020; 8: 132665–132676.

[8] Kim CK, Choi JW, Jiao Z, Wang D, Wu J, Yi TY, *et al*. An automated COVID-19 triage pipeline using artificial intelligence based on chest radiographs and clinical data. Npj Digital Medicine. 2022; 5: 5.

[9] Mei X, Lee H, Diao K, Huang M, Lin B, Liu C, *et al*. Artificial intelligence—enabled rapid diagnosis of patients with COVID-19. Nature Medicine. 2020; 26: 1224–1228.

[10] Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Computers in Biology and Medicine. 2020; 121: 103792.

[11] Puppe F. Systematic introduction to expert systems: knowledge representations and problem-solving methods. 1st. Springer-Verlag: Berlin. 1993.

[12] Kiseleva TV, Toropchina EV. Expert systems in medical diagnostics: analysis of functioning. Automation and Remote Control. 2014; 75: 1316–1322.

[13] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019; 380: 1347–1358.

[14] Sajda P. Machine learning for detection and diagnosis of disease. Annual Review of Biomedical Engineering. 2006; 8: 537–565.

[15] Schulz H, Behnke S. Deep learning. German Journal of Artificial Intelligence. 2012; 26: 357–363.

[16] Bakator M, Radosav D. Deep learning and medical diagnosis: a review of literature. Multimodal Technologies and Interaction. 2018; 2: 47.

[17] Bustos N, Tello M, Droppelmann G, García N, Feijoo F, Leiva V. Machine learning techniques as an efficient alternative diagnostic tool for COVID-19 cases. Signa Vitae. 2022; 18: 23–33.

[18] Zhang J, Xie Y, Wu Q, Xia Y. Medical image classification using synergic deep learning. Medical Image Analysis. 2019; 54: 10–19.

[19] Lu D, Weng Q. A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing. 2007; 28: 823–870.

[20] Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, *et al*. AI applications to medical images: from machine learning to deep learning. European Journal of Medical Physics. 2021; 83: 9–24.

[21] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et*

al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542: 115–118.

[22] Tartaglione E, Barbano CA, Berzovini C, Calandri M, Grangetto M. Unveiling COVID-19 from chest X-Ray with deep learning: a hurdles race with small data. International Journal of Environmental Research and Public Health. 2020; 17: 6933.

[23] Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, *et al*. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence. 2021; 3: 199–217.

[24] Goodfellow I, Bengio Y, Courville A. Deep learning. Edition. The MIT Press: Cambrige. 2016.

[25] Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009; 248–255.

[26] Simonyan K, Zisserman A. 'Very deep convolutional networks for large-scale image recognition', The 3rd International Conference on Learning Representations. San Diego, California, United States. 2015.

[27] Howard A, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, *et al*. MobileNets: efficient convolutional neural networks for mobile vision applications. 2017. Available at: https://arxiv.org/abs/1704.04861 (Accessed: Day 13 February 2022).

[28] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 770–778.

[29] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 2818–2826.

[30] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017; 4700–4708.

[31] Chollet F. Xception: deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017; 1251–1258.

[32] Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Scientific Reports. 2020; 10: 19549.

[33] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018; 4510–4520.

[34] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 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 (CVPR). 2017; 2097–2106.

[35] Ronneberger O, Fischer P, Brox T. 'U-net: convolutional networks for biomedical image segmentation', International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany, 2015. Springer International Publishing: Switzerland. 2015.

[36] Wang L, Zhang Y, Whang D, Tong X, Liu T, Zhang S, *et al*. Artificial intelligence for COVID-19: a systematic review. Frontiers in Medicine. 2021; 8: 704256.

How to cite this article: Miguel Lastra Leidinger, Francisco Aragón Royón, Oier Etxeberria, Luis Balderas, Antonio Jesús Láinez Ramos-Bossini, Genaro López Milena, *et al.* A real-world intelligent system for the diagnosis and triage of COVID-19 in the emergency department. Signa Vitae. 2023; 19(3): 91-102. doi: 10.22514/sv.2022.070.