Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 2212–2221

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

# Machine learning for coronavirus covid-19 detection from chest x-rays

Luca Brunese[a], Fabio Martinelli[b], Francesco Mercaldo[a,b,*], Antonella Santone[c]

[a] *Department of Medicine and Health Sciences "Vincenzo Tiberio", University of Molise, Campobasso, Italy*
[b] *Institute for Informatics and Telematics, National Research Council of Italy (CNR), Pisa, Italy*
[c] *Department of Biosciences and Territory, University of Molise, Pesche (IS), Italy*

**Abstract**

At the end of 2019, a new form of Coronavirus, called *COVID-19*, has widely spread in the world. To quickly screen patients with the aim to detect this new form of pulmonary disease, in this paper we propose a method aimed to automatically detect the *COVID-19* disease by analysing medical images. We exploit supervised machine learning techniques building a model considering a data-set freely available for research purposes of 85 chest X-rays. The experiment shows the effectiveness of the proposed method in the discrimination between the *COVID-19* disease and other pulmonary diseases.

*Keywords:* Coronavirus, COVID-19, machine learning, artificial intelligence, medical images, x-ray

## 1. Introduction and Related Work

Coronaviruses represent an extended family of respiratory viruses that can cause mild to moderate diseases, from the common cold to respiratory syndromes such as MERS (Middle East respiratory syndrome) and SARS (Severe acute respiratory syndrome) [18]. They are so called because of the crown-shaped tips that are present on their surface [17].

These kind of viruses are common in many animal species (such as camels and bats) but in some cases, though rarely, they can evolve and infect humans and then spread to the population [1]. A new coronavirus strain that has never previously been identified in humans is the one appeared at the end of 2019 i.e., the 2019 novel coronavirus (*COVID-19*, acronym of *COronaVIrus Disease 19*) [19, 3, 28]. The first cases were found during the *COVID-19* pandemic of 2019-2020 [22], which probably started around the end of December 2019 in the city of Wuhan [28], the capital of the Chinese province of Hubei, and subsequently spread to various countries of the world.

[*] Francesco Mercaldo
*E-mail address:* francesco.mercaldo@unimol.it

10.1016/j.procs.2020.09.258

In fact, as of January 28 2020, there were more than 4600 confirmed cases of contagion in many countries of the world and 106 deaths while on February 15 these data had already risen to 49053 cases and 1381 deaths[1]. As of January 23 2020, Wuhan has been quarantined with the suspension of all public transport into and out of the city, which measures were extended the following day to the neighboring cities of Huanggang, Ezhou, Chibi, Jingzhou and Zhijiang. Further limitations and controls have been adopted in many areas of the world, also in Europe where several cases have also been recorded. The country most affected in Europe is Italy, where the authorities have struggled to contain an outbreak that has infected at least 400 people, most of them in northern Italy, near Milan. As of March 2, there have been over 1800 confirmed coronavirus cases and 30 deaths in Italy, with the third highest number of infections per country in the world, after China and South Korea.

The *COVID-19* infection caused clusters of fatal pneumonia with clinical presentation greatly resembling SARS-CoV. In fact, patients experience flu-like symptoms such as fever, dry cough, tiredness, difficulty breathing. In more severe cases, often found in subjects already burdened by previous pathologies, pneumonia develops, acute renal failure, up to even death [22], but this new coronavirus presents also several unique features [28, 19]. While the diagnosis is confirmed using polymerase chain reaction (PCR), infected patients with pneumonia may present on chest X-ray and computed tomography (CT) images a pattern that is only moderately characteristic for the human eye as demonstrated by researchers in [26]. The rate of transmission of *COVID-19* depends on the capacity to reliably identify infected patients with a low rate of false negatives. In addition, a low rate of false positives is required to avoid further increasing the burden on the healthcare system by unnecessarily exposing patients to quarantine if that is not required. Along with proper infection control, it is evident that timely detection of the disease would enable the implementation of all the supportive care required by patients affected by *COVID-19*.

In late January 2020, Chinese researchers discussed the clinical and paraclinical features *COVID-19* specific [19]. They reported that patients present abnormalities in chest CT images with most having bilateral involvement [19]. Bilateral multiple lobular and subsegmental areas of consolidation constitute the typical findings in chest CT images of intensive care unit (ICU) patients on admission [19]. In comparison, non-ICU patients show bilateral ground-glass opacity and subsegmental areas of consolidation in their chest CT images [19]. In these patients, later chest CT images display bilateral ground-glass opacity with resolved consolidation [19].

As stated in [13, 2] *COVID-19* is possibly better diagnosed using radiological imaging, for this reason in this paper we evaluate the possibility to detect the *COVID-19* disease directly from x-ray images. For this reason, in this paper supervised machine learning is exploited to build a model starting from a set of patients *COVID-19* diagnosed and patients with other respiratory diseases exhibiting symptoms similar to *COVID-19*.

The remaining of the paper proceeds as follow: Section 2 presents the proposed method from *COVID-19* detection from x-rays, Section 3 describes the performance results in the evaluation of real-world chest X-rays and, in the last section, conclusion and future works are drawn.

## 2. The Method

In this section the proposed method for *COVID-19* detection from x-rays is discussed.

The proposed method relies in supervised machine learning [24, 11, 8] and it is composed by two main phases: training, depicted in Figure 1 and testing, shown in Figure 2.

The training phase is aimed to build a model for discriminating between x-rays images *COVID-19* related and images related to *other* pathologies.

As shows from Figure 1, from the Picture archiving and communication system (PACS) chest X-rays are obtained. As required by supervised classification, we need a label (i.e., a diagnosis): for this reason the proposed method requires the domain experts (i.e., radiologists) to assign to each training chest X-rays a label (i.e, COVID-19 or other). In particular in the *other* category following pulmonary diseases are considered: *Streptococcus*, *SARS*, *ARDS* (Acute respiratory distress syndrome) and *Pneumocystis* [14].

To obtain numerical values from medical images, we consider a set of color layout descriptor (CLD) features. CDL features are designed to capture the spatial distribution of color in an image [20]. The feature extraction process

[1] https://www.worldometers.info/coronavirus/

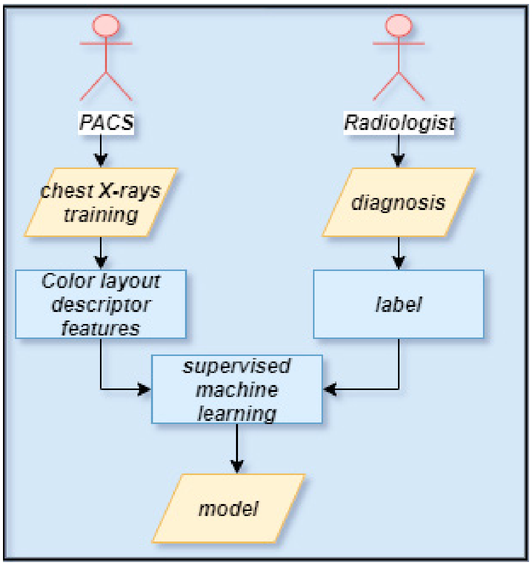

Fig. 1: The training phase.

consists of two parts: grid based representative color selection and discrete cosine transform with quantization [9]. Clearly, color is the most basic quality of the visual contents, therefore it is possible to use colors to describe and represent an image. The MPEG-7 standard has tested the most efficient procedure to describe the color and has selected those that have provided more satisfactory results [23]. This standard proposes different methods to obtain these descriptors, and one method defined to describe the color is the CLD, that permit to describe the color relation between sequences or group of images [12], for this reason this standard is chosen in this paper to extract meaningful numeric features from x-ray images.

MPEG-7 visual descriptors include the color, texture and shape descriptor for efficient content-based image retrieval [21].

Basically, an x-ray image is divided into 64 equal blocks and we compute the average color for each block, and then features are calculated from the averages [23]: this is resulting in 64 different features (one feature for each block in which we divide the x-ray image). We apply a feature selection algorithm i.e., the CfsSubsetEval [16] one: the final feature set is composed from a total of 33 CLD features that are included in this study.

The feature set is obtained from each chest X-ray and, with the associated label, it represents the input for the supervised machine learning algorithm, that will output the model.

Once generated the model, in the testing phase we evaluate the performance of the model built in the training phase.

As shows from Figure 2, in this phase we obtain the numerical features from a set of chest X-ray not considered in the previous phase: this represents the input for the model that will generate the prediction i.e., whether the input chest X-ray is related to the *COVID-19* or to the *other* category. Subsequently, the output of the model is validated by the radiologist.

## 3. Experimental Analysis

The effectiveness of the proposed feature set in discriminating between *COVID-19* and *other* disease is organised in descriptive statistics i.e., boxplot analysis and the evaluation of the model obtained as output from the machine learning classifier.

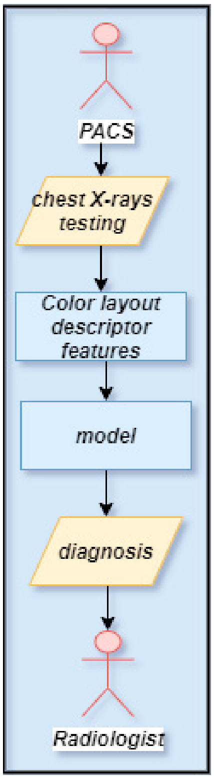

Fig. 2: The testing phase.

### 3.1. The Data-set

Real world x-ray images are considered in this work and are available for research purposes [2].
In details, the data-set consists in x-ray images belonging to the pathologies shown in Table 1

Table 1: The considered pathologies.

| *COVID-19* | *Streptococcus* | *SARS* | *ARDS* | *Pneumocystis* |
|---|---|---|---|---|
| 63 | 6 | 11 | 4 | 2 |

Figure 3 shows the detail about the patient age with regard to the *COVID-19* patients.

As shown from Figure 3, the majority of patients are in the 51-55 years range. Moreover, the age classes with a number of patients equal to 7 are following: 0-39, 39-43, 55-59 and 63-67.

Figure 6 shows an example of chest X-rays belonging to the analysed data-set.

The medical images are obtained from different institutions, as shown in Table 2:

As shows from Table 2, the medical images are obtaining from different institutions: from China (where *COVID-19* disease started to manifest) from Italy (the European area in which it was found to be most widespread), but also from Australia and USA.

Figure 4 shows the sex distribution in the X-rays.

Unfortunately, we do not have the detail about the sex for all the patients involved in the experiment. Anyway, the majority of X-ray is related to men with regard to the *COVID-19* disease, while for the *other* pathologies the majority is female.

[2] https://github.com/ieee8023/covid-chestxray-dataset

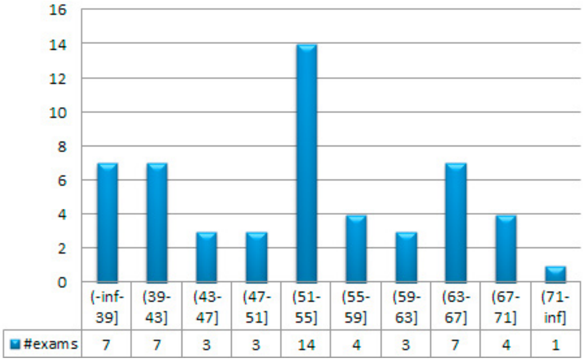

Fig. 3: Patient age: on the *x* axis the number of exams, while on the *y* axis the number of patients belonging to the age clusters we defined.

Table 2: Institutions involved the study.

| *Hospital* | *Country* |
|---|---|
| Ospedale Santo Spirito, Rome | Italy |
| Riccione | Italy |
| Myongji Hospital, Goyang | Korea |
| Jönköping | Sweden |
| Melbourne | Australia |
| Royal Brisbane and Women's Hospital, Brisbane | Australia |
| Sichuan Provincial People's Hospital, Chengdu | China |
| Tongji Medical College, Wuhan, Hubei Province | China |
| Taoyuan General Hospital, Taoyuan | Taiwan |
| Snohomish County, Washington | USA |

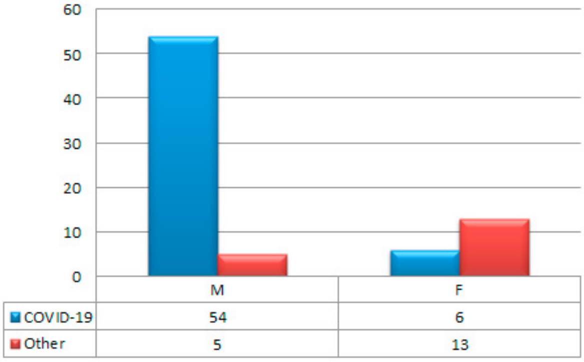

Fig. 4: Patient sex: on the *x* axis the number of patients in the data-set afflicted by *COVID-19* and other diseases, while on the *y* axis with *M* we indicate the *male* patients, while with *F* we indicate the *female* ones.

Figure 5 shows the details about the status (dead or alive) or the patient involved in the experiment. This data is updated to March 2020.

With regard to *COVID-19* disease all the patients for whom we have this label are alive. Relating to *other* pathologies, 2 patients resulting alive, while 6 are dead.

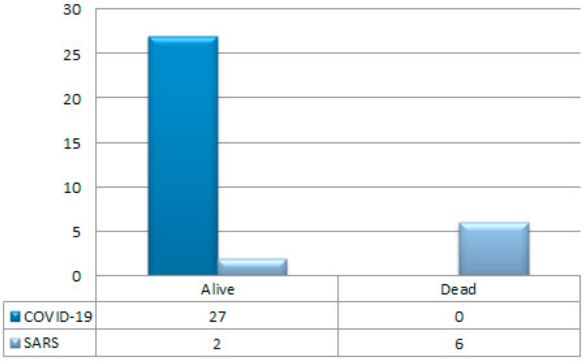

Fig. 5: Alive and dead patients: on the *x* axis the number of patients in the data-set with by *COVID-19* and other diseases, while on the *y* axis with *Alive* we indicate the *alive* patients, while with *dead* we indicate the *dead* ones.

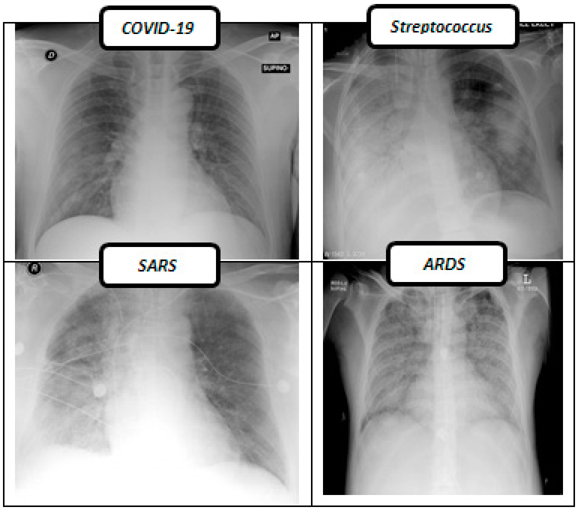

Fig. 6: Example of chest X-rays.

### 3.2. Descriptive statistics

In this section, we present the boxplot analysis. For reason space, we present plots related only to four color layout descriptor features but similar consideration can be made also for the remaining features.

Figure 7 shows the boxplot for the third feature.

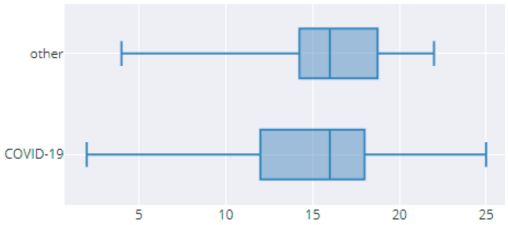

Fig. 7: Color layout descriptor feature 3 boxplot.

As emerges from the boxplot, the COVID-19 distribution for the color layout descriptor feature 3 is partially overlapped with the distribution of the other pulmonary pathologies. Symptomatic that this feature can exhibit a medium discrimination effectiveness between *other* respiratory diseases and the *COVID-19* disease.

Figure 8 shows the boxplots for the color layout descriptor feature 6.

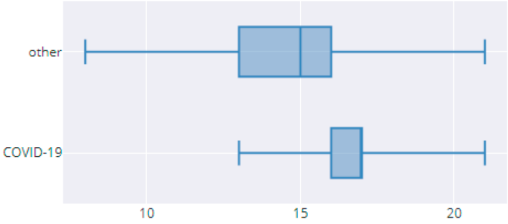

Fig. 8: Color layout descriptor feature 6 boxplot.

In this case, the two distributions range in different numeric values: in fact, there is just a slightly overlapping between the two distributions. For this reason it is reasonable that this feature can exhibit a more discriminative power if compared to feature 3.

The next boxplot in Figure 9 we discuss is the one related to the color layout descriptor feature 10 features.

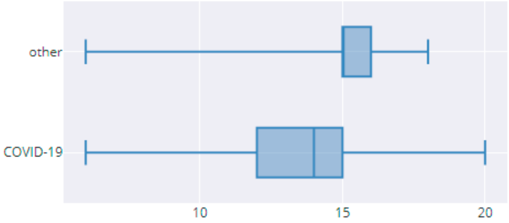

Fig. 9: Color layout descriptor feature 10 boxplot.

The distributions, similarly to the previous color layout descriptor feature we analysed, does not exhibits overlapped areas. Coherently with the previous features, also in this boxplot, the numerical values for the *COVID-19* distribution are greater if compared to the ones related to the other numeric values.

The last boxplot we discuss is the one depicted in Figure 10, related to the color layout descriptor feature 15.

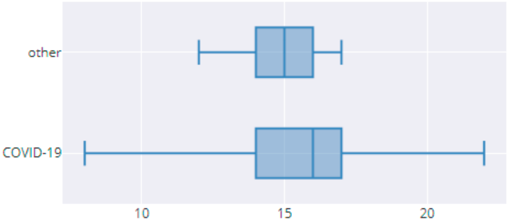

Fig. 10: Color layout descriptor feature 15 boxplot.

This boxplot seems to exhibit a less discriminative power if compared to the previous ones. In fact, the area overlapped between the two boxplots is extended, and there is also a little area belonging to the *COVID-19* distribution that is not overlapped.

### *3.3. Experimental Analysis*

In the follow we explain the settings of the experiment to generate the *COVID-19* detection model.

With regard to the model building, for training the model, we defined $T_{detection}$ as a set of labeled messages $\{(M_{detection}, l_{detection})\}$, where each $M_{detection}$ is the label associated to a $l_{detection} \in \{$ *COVID-19*, *other*$\}$. For each $M_{detection}$ we built a feature vector $F \in R_y$, where $y$ is the number of the features used in training phase ($y = 33$).

For the learning phase, we consider a $k$-fold cross-validation: the data-set is randomly partitioned into $k$ subsets. A single subset is retained as the validation data-set for testing the model, while the remaining $k - 1$ subsets of the original data-set are used as training data. We repeated the process for $k = 10$ times; each one of the $k$ subsets has been used once as the validation data-set. To obtain a single estimate, we computed the average of the $k$ results from the folds.

We evaluated the effectiveness of the built model with the following procedure:

1. build a training set $T \subset D$;
2. build a testing set $T' = D \div \mathrm{T}$;
3. run the training phase on $T$;
4. apply the learned classifier to each element of $T'$.

Each classification was performed using 90% of the data-set as training data-set and 10% as testing data-set employing the full feature set exploiting the *K-nearest neighbours classifier* (k-NN) classification algorithm implementation available in the Waikato Environment for Knowledge Analysis[3] (Weka) software, a suite for machine learning experiments. In the k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. We experimented with $k$=4.

Five metrics are considered to evaluate the performance of the classifiers: FP-Rate, Precision, Recall, F-Measure and Roc Area.

The FP-Rate (i.e., false positive rate) is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events:

$$FP\ Rate = \frac{fp}{fp+tn}$$

where *fp* indicates the number of false positives and *tn* indicates the number of true negatives.

The precision has been computed as the proportion of the examples that truly belong to class X among all those which were assigned to the class. It is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved:

$$Precision = \frac{tp}{tp+fp}$$

where *tp* indicates the number of true positives and *fp* indicates the number of false positives.

The recall has been computed as the proportion of examples that were assigned to class X, among all the examples that truly belong to the class, i.e., how much part of the class was captured. It is the ratio of the number of relevant records retrieved to the total number of relevant records:

$$Recall = \frac{tp}{tp+fn}$$

where *tp* indicates the number of true positives and *fn* indicates the number of false negatives.

The F-Measure is a measure of a test's accuracy. This score can be interpreted as a weighted average of the precision and recall:

$$F\text{-}Measure = 2 * \frac{Precision*Recall}{Precision+Recall}$$

The ROC area is created by considering the true positive rate against the false positive rate.

[3] https://www.cs.waikato.ac.nz/ml/weka/

Table 3 shows the classification results.

Table 3: Classification results.

| **Class** | *FP Rate* | *Precision* | *Recall* | *F-Measure* | *ROC Area* |
|---|---|---|---|---|---|
| *COVID-19* | 0.087 | 0.968 | 0.984 | 0.976 | 0.989 |
| *other* | 0.016 | 0.955 | 0.913 | 0.933 | 0.989 |
| *average* | 0.068 | 0.965 | 0.965 | 0.964 | 0,989 |

As shows by classification results shows in Table 3, an average precision and an average recall equal to 0.965 are obtained. In particular, for the *COVID-19* detection a precision equal to 0.968 and a recall of 0.964 are reached, demonstrating the ability of supervised machine learning for the discrimination between *COVID-19* and *other* pulmonary pathologies.

To better understand the ability prediction of the proposed model, Table 4 shows the confusion matrix.

Table 4: Confusion matrix.

| **a** | *b* | ←*classified as* |
|---|---|---|
| 61 | 1 | a = *COVID-19* |
| 2 | 21 | b = *other* |

Confusion matrix in Table 4 shows that only 3 X-rays in total are misclassified, in particular, 1 *COVID-19* x-ray image is erroneously classified as belonging to the *other* pulmonary disease category, while 2 x-ray images are classified as belonging to the *COVID-19* category, while radiologists marked these medical images with the *other* pulmonary disease category.

## 4. Conclusion and Future Work

Considering the rate of spread of *COVID-19*, automatic techniques are needed for the detection of this disease. In this paper, a machine learning method is proposed for the detection of *COVID-19* disease. The evaluation demonstrated the effectiveness of the proposed method, by obtaining an average precision and recall equal to 0.965 in the discrimination between the *COVID-19* and *other* pulmonary diseases with similar symptoms.

As future work, we plan to validate the proposed method considering also a set of healthy chest X-rays. Moreover, we will investigate whether deep learning [5, 25] and formal verification techniques [15, 27] can obtain better performances in the *COVID-19* as demonstrated in similar contexts, from cancer detection to malware detection [7, 4, 10, 6].

## ACKNOWLEDGEMENTS

This work has been partially supported by MIUR - SecureOpenNets and EU SPARTA contract 830892, CyberSANE projects, and the EU project CyberSure 734815.

## References

[1] Abroug, F., Slim, A., Ouanes-Besbes, L., Kacem, M.A.H., Dachraoui, F., Ouanes, I., Lu, X., Tao, Y., Paden, C., Caidi, H., et al., 2014. Family cluster of middle east respiratory syndrome coronavirus infections, tunisia, 2013. Emerging infectious diseases 20, 1527.

[2] Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., Xia, L., 2020. Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. Radiology , 200642.

[3] Brooks, W.A., 2020. Bacterial pneumonia, in: Hunter's Tropical Medicine and Emerging Infectious Diseases. Elsevier, pp. 446–453.

[4] Brunese, L., Mercaldo, F., Reginelli, A., Santone, A., 2019a. Formal methods for prostate cancer gleason score and treatment prediction using radiomic biomarkers. Magnetic resonance imaging .

[5] Brunese, L., Mercaldo, F., Reginelli, A., Santone, A., 2019b. Neural networks for lung cancer detection through radiomic features, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–10.

[6] Brunese, L., Mercaldo, F., Reginelli, A., Santone, A., 2019c. Prostate gleason score detection and cancer treatment through real-time formal verification. IEEE Access 7, 186236–186246.

[7] Brunese, L., Mercaldo, F., Reginelli, A., Santone, A., 2020. An ensemble learning approach for brain cancer detection exploiting radiomic features. Computer methods and programs in biomedicine 185, 105134.

[8] Carfora, M.F., Martinelli, F., Mercaldo, F., Nardone, V., Orlando, A., Santone, A., Vaglini, G., 2019. A "pay-how-you-drive" car insurance approach through cluster analysis. Soft Computing 23, 2863–2875.

[9] Cieplinski, L., 2001. Mpeg-7 color descriptors and their applications, in: International Conference on Computer Analysis of Images and Patterns, Springer. pp. 11–20.

[10] Cimino, M.G., De Francesco, N., Mercaldo, F., Santone, A., Vaglini, G., 2020. Model checking for malicious family detection and phylogenetic analysis in mobile environment. Computers & Security 90, 101691.

[11] Cimitile, A., Martinelli, F., Mercaldo, F., 2017. Machine learning meets ios malware: Identifying malicious applications on apple environment., in: ICISSP, pp. 487–492.

[12] Eidenberger, H., 2003. How good are the visual mpeg-7 features?, in: Visual Communications and Image Processing 2003, International Society for Optics and Photonics. pp. 476–488.

[13] Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., Ji, W., 2020. Sensitivity of chest ct for covid-19: comparison to rt-pcr. Radiology , 200432.

[14] Glass, W.G., Subbarao, K., Murphy, B., Murphy, P.M., 2004. Mechanisms of host defense following severe acute respiratory syndrome-coronavirus (sars-cov) pulmonary infection of mice. The Journal of Immunology 173, 4030–4039.

[15] Gradara, S., Santone, A., Villani, M., 2006. Delfin+: An efficient deadlock detection tool for ccs processes. Journal of Computer and System Sciences 72, 1397–1412. doi:10.1016/j.jcss.2006.03.003.

[16] Hall, M.A., 1998. Correlation-based feature subset selection for machine learning. Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato .

[17] van der Hoek, L., Pyrc, K., Jebbink, M.F., Vermeulen-Oost, W., Berkhout, R.J., Wolthers, K.C., Wertheim-van Dillen, P.M., Kaandorp, J., Spaargaren, J., Berkhout, B., 2004. Identification of a new human coronavirus. Nature medicine 10, 368–373.

[18] Holmes, K.V., 2003. Sars-associated coronavirus. New England Journal of Medicine 348, 1948–1951.

[19] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al., 2020. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. The Lancet 395, 497–506.

[20] Kasutani, E., Yamada, A., 2001. The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval, in: Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), IEEE. pp. 674–677.

[21] Kim, S.M., Park, S.J., Won, C.S., 2007. Image retrieval via query-by-layout using mpeg-7 visual descriptors. ETRI journal 29, 246–248.

[22] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y., et al., 2020. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. New England Journal of Medicine .

[23] Manjunath, B.S., Ohm, J.R., Vasudevan, V.V., Yamada, A., 2001. Color and texture descriptors. IEEE Transactions on circuits and systems for video technology 11, 703–715.

[24] Mercaldo, F., Nardone, V., Santone, A., 2017. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Computer Science 112, 2519–2528.

[25] Mercaldo, F., Santone, A., . Deep learning for image-based mobile malware detection. Journal of Computer Virology and Hacking Techniques , 1–15.

[26] Ng, M.Y., Lee, E.Y., Yang, J., Yang, F., Li, X., Wang, H., Lui, M.M.s., Lo, C.S.Y., Leung, B., Khong, P.L., et al., 2020. Imaging profile of the covid-19 infection: radiologic findings and literature review. Radiology: Cardiothoracic Imaging 2, e200034.

[27] Santone, A., Vaglini, G., Villani, M., 2013. Incremental construction of systems: An efficient characterization of the lacking sub-system. Science of Computer Programming 78, 1346–1367.

[28] Wang, C., Horby, P.W., Hayden, F.G., Gao, G.F., 2020. A novel coronavirus outbreak of global health concern. The Lancet 395, 470–473.