IJCI V9-02(2022) 88–95

International Journal of Computers and Information (IJCI)

Available online at *https://ijci.journals.ekb.eg/*

Improving COVID 19 Detection Based on a Hybrid Data Mining Approach

Dina Abdelftah Gouda, Nader Mahmoud

Faculty of Computers and Information, Menofia University, Shebin Elkom 32511, Egypt.
dinagouda1@gmail.com, nader.mahmoud@ci.menofia.edu.eg

Abstract

The worldwide spread of coronavirus disease (COVID-19) has become a threatening risk for global public health. Currently, doctors resort to PCR analysis, however, it suffers from low accuracy problems. On the other hand, Convolutional neural network (CNN) and despite its high accuracy incorrect classification, it takes a long time to train data, in addition it requires large training dataset. In this paper, we propose a hybrid approach for COVID-19 detection and diagnosis. Our contribution consists of two phases to provide high detection accuracy. In the first phase, we propose a hybrid features-fusion phase that works by fusing four common features extracted from medical image, Row pixel intensity, Colour histogram, Harlick texture and Threshold. Each single classifier is fed with these four features and yielded a 4 different predictions for each feature. A well-known voting technique is then applied to provide final predication result for each classifier. Secondly, the ensemble stacking technique is employed to fuse predication of each classifier, which significantly improves final detection accuracy. The proposed approach has been quantitatively evaluated on a public dataset of 5000 CT- images. The proposed approach yields accuracy of 99.3% and overcome traditional approaches such as KNN (K-nearest neighbours) that yields 92%, and SVM (Support vector machines) that yields 92% comparable computational time that is approximately 4.9 minutes.

Keywords: COVID-19, Computerized Tomography; Chest X- ray; CNN; Deep Learning.;

1. Introduction

COVID-19 is the most deadly pandemic that faced the world in last two years. The COVID-19 pandemic is among the most disruptive events to global health in living memory. The COVID-19 continues to generate a constant pandemic threat with new mutations of the viral agent (SARS-CoV-2) that create socioeconomic issues. One of the fundamental problems is the evaluation of the preparedness of countries to cope with COVID-19 pandemic crisis to detect and support factors associated with the reduction of mortality and the growth of vaccinations in society. Thus, it is important to diagnose the disease timely and correctly for the treatment and epidemic control. Viral nucleic acid testing is the main method for diagnosing COVID-19, and PCR is the most widely traditional way for diagnosing COVID-19. However, these traditional ways suffer from low accuracy in addition to, it is time consuming [1].

On the other hand, medical images (e.g. CT-image and X-ray) play an important role in the COVID-19 detection. These images are being widely used for different classifications and diagnosis problems. Currently, the use of data mining classification techniques has been researched for COVID-19 detection [2]. It provides a promising detection result, with lower computational time.

In this paper, we propose a hybrid approach for COVID-19 detection using CT-images. The proposed approach consists of two main phases. In the initial phase, we fed four individual classifiers with four different features separately. The voting technique is then employed between these initial 4 predictions for each classifier to obtain an initial prediction. Secondly, we fed the ensemble stacking classifier with highly voted class from each classifier to produce final classification result.

The rest of the paper is organized as follows. Sec. 2 presents a literature review of relevant works. And Sec. 3 presents Proposed Work. Experimental results are discussed in Sec. 4. Conclusion and future work are presented in Sec. 5

2. Related Work

The COVID-19 outbreak was first reported in Wuhan, China by the end of 2019, it spread quickly around the World in a matter of months. The evidence points to an exponential growth in the number of cases, as of right now there are more than 298 million confirmed cases worldwide (World Health Organization. [3]).In some cases, infected people may be asymptomatic. Due to the contiguous nature of the virus, the early detection of COVID-19 is required. Nowadays, the infection of Coronavirus is tested through Reverse Transcription-Polymerase Chain Reaction (RT-PCR). However, the RT - PCR test is time-consuming and suffers from high false-negative [4].

To resolve the above-mentioned problems, radiographic imaging techniques such as CT and X-ray are used [5]. These techniques are used to quantify the disease severity in the lung infection. CT provides better sensitivity than X-ray and RT-PCR [6]. Due to the rapid growth of essential data and the need to analyse the relationships and hierarchy between data, machine learning and data mining techniques are researched in the health system to provide efficiency in prevention, diagnosis and treatment [7].

In [8], the authors perform an early comparison between a series of technologies and study their impact on public health systems. They conclude that the field of Artificial Intelligence and data mining are the most prominent solution to be used for public healthcare diseases detection and diagnosis. Several researchers studied the use of deep learning, due to its remarkable classification performance in other application, as an efficient solution for COVID-19 detection.

In [9], they have researched the use of three different deep neural network models (i.e. ResNet50, Inception-V3, and InceptionResNetV2) with both CT and X-ray images for COVID- 19 detection. They used a dataset consists of 100 images: 50 COVID-19 infected patients gathered from the open source GitHub repository shared by [10], and 50 images of healthy people obtained from Kaggle repository [11]. The results were obtained using fivefold cross validation, and yielded a detection accuracy of: 98% with ResNet50 model, 97% with the Inception- V3 model, and 87% with InceptionResNetV2. Despite, its high detection accuracy, it requires a long time to classify a few CT or X-ray images, in addition to large training data.

Another research work has researched the use of traditional data mining classifiers such as [12]. They have investigated the use of these classifiers with a small dataset consists of 250 X-ray image. They obtained a detection of 79.27% with KNN and 80.49% with NB (Naivbayes). In [13], they employed more sets of mining classifiers on a slightly larger dataset consists of 5644 CT-images and yielded a minor improvements in detection accuracy of: 82.68% with NB, 86. 59% with SVM, 80. 44% with KNN and 87. 15% with RF (Random Forest). However, such data mining approaches provide good computation time, it still suffers from low detection accuracy in contrast with CNN approaches, thus requires further improvements. Similarly, in [14] they have researched the use of gold-standard classifiers (SVM, KNN, RF, NB, and LR (Logistic regression)) with a dataset of CT-images consists of 1473 images. They yielded a detection accuracy of 91.63% with SVM, 90.28% with KNN and 91.32% with RF.

On the other hand other research works have studied the effect of medical images (e.g.: CT-images) features, which is being extracted from infected and healthy patients, on final classification [15]. They yielded a detection accuracy of 86.27% with SVM, 86.35% with RF and 94.13% with JRIP (Joint Reserve Intelligence Program).

To summarize, CNN approaches prove to be a reliable alternative to PCR for virus detection with high detection accuracy results. However, these approaches suffer from long processing time to train the data, in addition to large training data that is required for fine-tuning algorithm main parameters. On the other hand, data mining traditional approaches do not require such huge training data and time as CNN approaches. However, these approaches suffer from low detection accuracy in contract with CNN approaches. Thus, the proposed approach consists of two- folds:

1- A features-fusion based approach, in which we feed a single classifier with 4 different features. For each feature, single classifier produced a prediction, thus voting technique is then employed to select highly voted class between four predication of each classifier. This initial estimate is then fed to the next stage.

2- An efficient ensemble stacking based on a set of carefully chosen classifiers to provide a final estimate. We utilize an LR Meta model to combine initial estimates from each classifier.

3. Proposed Work

Fig.1 shows the main phases of the proposed phases. It starts with a traditional preprocessing phase, followed by feature extraction, and finally a classification phase. Preprocessing phase is firstly introduced in Sec. A. and Sec. B and Sec. C, discusses the features fusion step and ensemble stacking step, respectively.

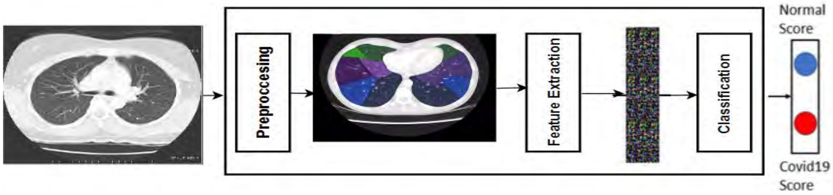


Fig. 1: Overview of the proposed approach.

A. Data Preprocessing

Typically, medical image data tend to be not fully complete, noisy, conflicting, and irregular. Thus, data pre-processing is necessary in order to reduce errors and noise. The worldwide health organization published several large datasets for COVID-19 cases. These datasets consist of CT-images of a patient's lung, in addition to a dataset of CT-images of healthy people. Fig. 2, shows CT-images, samples of infected and non-infected patients. Due to the intense similarity between lesions and normal tissues in CT-images, the precise detection and segmentation of the infected area is certainly important as pre-processing task. Image segmentation is a complex and challenging area of the biomedical engineering task that is affected by numerous aspects, including illumination, low contrast, noise, and irregularity of the lesions[16-18].In this paper, we relied on region-based segmentation [19,20], which separates lesions into different regions based on a Thresholding value. For the Threshold we used a Threshold equal 120 on the gray image. Fig. 2, shows samples of segmented images of patient's lung that represent our area of interest.

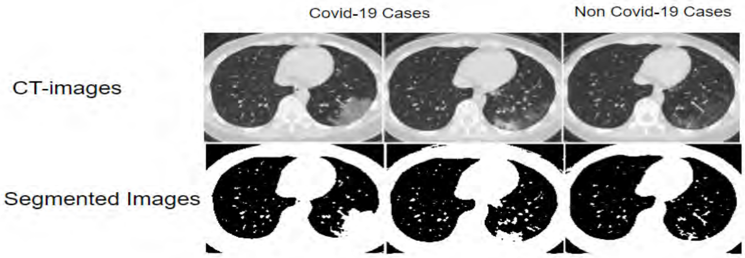


Fig. 2: image segmentation of CT-image

B. Feature fusion phase

This phase exploits four extracted features: Row pixel intensity, Colour histogram, Harlick texture and Threshold, to improve the accuracy of the single classifiers. It works as follows, the single classifier is applied with a single feature on all images in the dataset. Thus we obtain an evaluation matrix of 0's (represent Non-COVID) and 1's (represent COVID) for all images in the dataset, as shown in Fig. 3, which is an example of using Rowpixel feature with SVM classifier. This process is repeated with the other 3 features (Colour histogram, Harlick texture and Threshold), thus yielding 4 equal-sized matrices contain prediction results of all images in the dataset using 4 different features. A simple voting- based mode function is to be applied in the four matrices to get one final prediction matrix for all images in the dataset, as shown in Fig. 3. Moreover, this process is repeated for other classifiers (KNN, NB and RF), each of them is applied with 4 features as described before.

A simple voting-based mode function is to be applied to get one initial prediction matrix for all images in the dataset. KNN, SVM, and RF yielded a highest prediction accuracy due to their good initial prediction estimates for each feature. The prediction accuracy of both NB and LR is improved significantly despite their low initial prediction estimates for each feature.

C. Ensemble stacking phase

At this stage we use Ensemble-Stacking approach [24, 25] with our four classifiers. Our Ensemble-Stacking approach is divided into two main steps. Firstly, our four base classifier is applied as discussed in Sec. C. Then, the prediction of features fusion approach is fed to a Meta classifier to produce final prediction results as shown in Fig. 4. We choose Logistic Regression as our Meta model classifier for the Ensemble- Stacking.

4. **Results and Discussion**

This section presents a quantitative evaluation of the proposed approach. Firstly, the datasets are introduced in Sec. A. We provide computational time of each step of the proposed approach in Sec. B. Sec. C provides evaluation of individual classifiers. Sec. D provides evaluation of features fusion, Sec. E provides evaluation of ensemble stacking, and Sec. F provides results comparison.

A. Experimental Setup and dataset

Our data set has been collected from several Kaggle datasets [26], [27] and [28]. Kaggle is an online community which helps researchers to explore and publish datasets for testing their models. However, the major drawback with Kaggle datasets is their small sizes, which lead to over-fitting issues. It consists of 5000 CT-images: 2500 for COVID-19 infected patients and 2500 for healthy people. We split the dataset into 70% training and 30% testing sets. Fig. 5 shows an image samples from our dataset. All the experiments have been conducted on a computer with AMD CPU with 10 Cores 4C+6G 1.8 GHz, and 16.0 GB RAM. The accuracy has been measured as follows by adding the number of correctly classified samples and dividing it by the total number of reference sample with computational complexity equal to O (N^3).

B. Computational time

Table I, shows the average computational time for every step in the proposed approach. We measured computational time for the main steps. In the first step we run individual classifiers with each feature. This step takes 3.23 minutes. It is the summation of computational time for individual classifiers, as shown in Table II. We also measured the computational time for the voting step used for fusing classifier results from set of features, this step takes 0.837 minutes. In the last step, we measured the computational time for ensemble stacking phase, which takes 0.9 minutes.

C. Quantitative evaluation of individual classifiers

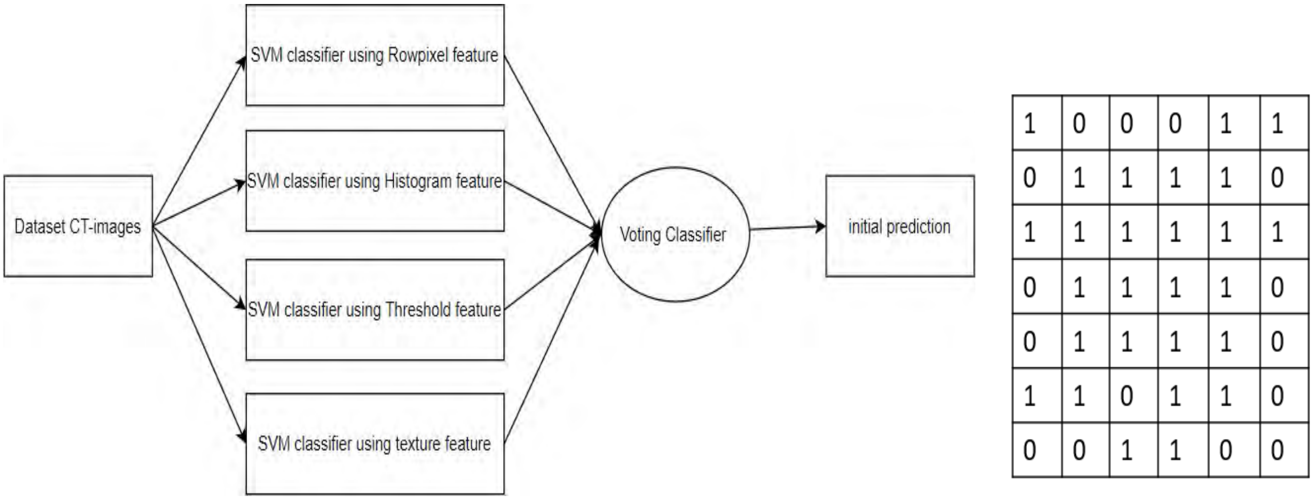
Initially, we have evaluated each individual classifier on the dataset. We have used Python implementation of SVM, KNN, RF, NB, and LR, with default tuning for different parameters. Table III, shows the yielded prediction accuracy of each class servers on four features: Row pixel intensity, Colour histogram, Harlick texture and Threshold. KNN, RF, and LR provide high prediction accuracy and yields 92%, 91%, and 84%, respectively when applied with Colour histogram. SVM and NB provide similar prediction accuracy with all features.

D. Quantitative evaluation of features-based fusion

In this section, we quantitatively evaluated the fusion of all features for individual classifiers. The last row in Table III shows the superior yielded prediction accuracy for each classifier. After using single feature. It can be clearly seen that all classifiers yielded a significant accuracy improvements when applied by fusion of four features: Row pixel intensity, Colour histogram, Harlick texture, and Threshold. It works as follows, the single classifier is applied with a single feature on all images in the dataset. Thus, we obtain an evaluation matrix of 0's and 1's for all images in the dataset. This step is repeated for each classifier with all features, yielding 4 equal-size matrices for prediction results of all images in the dataset. A simple voting- based mode function is to be applied to get one final prediction matrix for all the images in the dataset. KNN, SVM, and RF yielded a highest prediction accuracy due to their good initial prediction estimates for each feature. The prediction accuracy of both NB and LR is improved significantly despite their low initial prediction estimates for each feature.

E. Quantitative evaluation of ensemble stacking

In this section, we quantitatively evaluated the ensemble stacking step. As shown in Table III. Firstly, the traditional ensemble stacking is applied using KNN, SVM, RF, and NB classifiers with Row pixel intensity feature, as shown in row 1 in Table III. Secondly, the same process is repeated with other features and accuracy results are reported in the last column in Table III. The last row in Table III, shows the overall accuracy of the proposed approach. It reports the feature-based fusion that has been used for each class, which significantly improve prediction accuracy for each classifier. Then, these estimates are fed for the ensemble stacking, thus yield 99.3% prediction accuracy and overcome the traditional ensemble stacking approach with different features.


Dataset CT-images
SVM classifier using Rowpixel feature
SVM classifier using Histogram feature
SVM classifier using Threshold feature
SVM classifier using texture feature
Voting Classifier
initial prediction

1	0	0	0	1	1
0	1	1	1	1	0
1	1	1	1	1	1
0	1	1	1	1	0
0	1	1	1	1	0
1	1	0	1	1	0
0	0	1	1	0	0

Fig. 3: feature fusion step using SVM classifier.

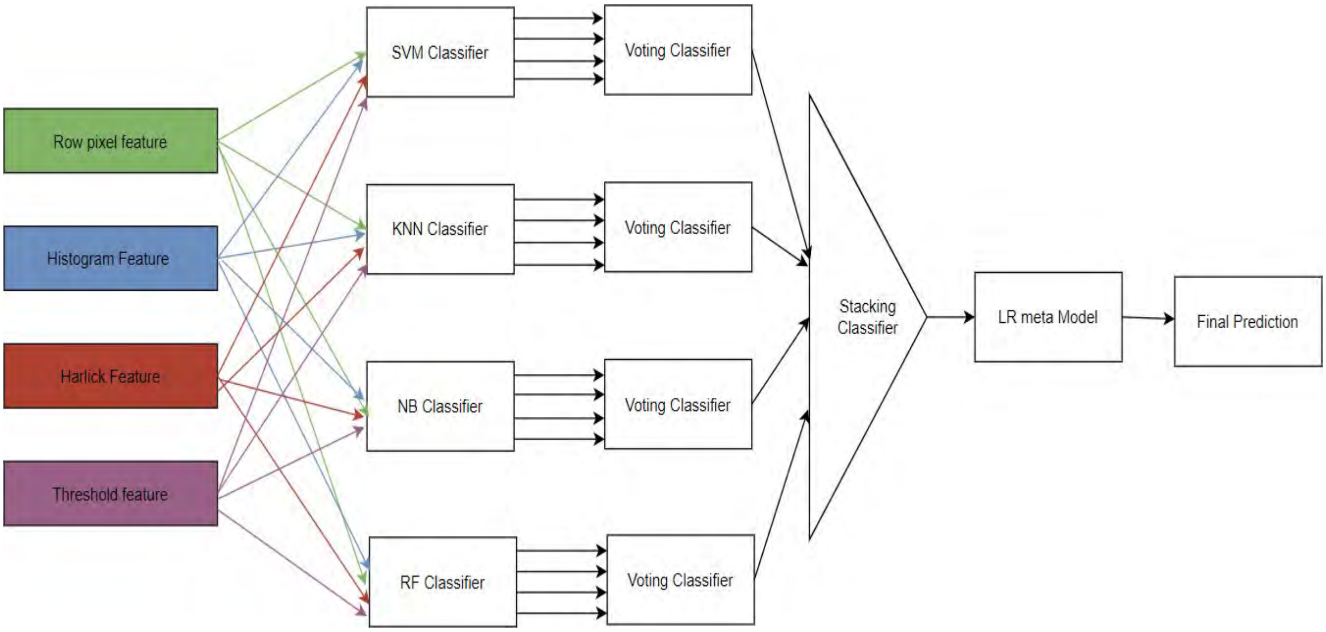


Fig. 4: Ensemble stacking

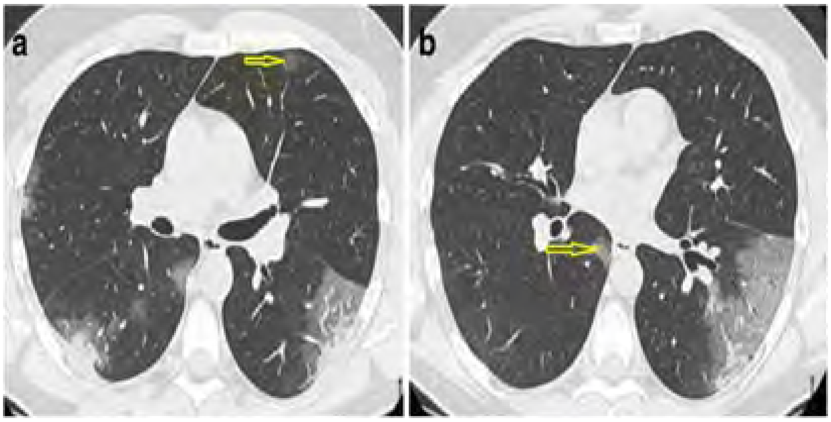


Fig. 5: Samples of COVID CT dataset [29]

F. Results comparison

In this section we provide a quantitative evaluation comparison with the most relevant research works. These researches work differ in their nature, where some of them rely on single classifiers, and others use multiple classifiers. We report in Table IV, the quantitative comparison between the proposed approach and relevant similar research work. It can be seen the proposed approach is more superior to other similar research works in terms of accuracy.

5. Conclusion

In this paper, an ensemble-stacking based classifiers model and features fusion model were proposed for COVID-19 classification in chest CT scan images. The two models utilized the SVM, KNN, RF, NB and LR as single classifier and Row pixel intensity, Colour Histogram, Harlick texture and Threshold as features. The two models are able to handle the sensitivity issue that is associated with RT-PCR. The two models have been tested on a large chest CT dataset and compared with other models. Experimental results reveal that those models achieve a high accuracy in less time. After applying ensemble- stacking based classifiers model and features fusion we aim to apply different fusion approach on features or classification algorithms to achieve high accuracy.

Tables

TABLE I: Computational time of every step of the proposed approach

Step	Individual Classifiers	Voting Algorithm	Staking Classifier
Time in minutes	3.23	0.837	0.90

TABLE II: Computational time of individual classifiers.

Classifiers	KNN	SVM	NB	RF
Time in minutes	0.81	0.84	0.72	0.85

TABLE III: Evaluation of Ensemble stacking approach

classifier	KNN	SVM	RF	NB	Classifiers ensemble-stacking
Row pixel intensity	85%	90%	87%	69%	92.8%
Color Histogram	92%	91%	91%	50%	94.8%
Harlick texture	88%	92%	89%	65%	94.8%
Threshold	85%	89%	85%	63%	90.8%
Feature-based fusion	98.9%	98.6%	98.7%	93%	**99.3%**

TABLE IV: Quantitative evaluation comparison with similar research works.

Sample	Hussein, L Nguyen et al. [12]	A.Akhtar et. al. [13]	Patel, Rajneesh Kumar et al [14]	Farid, Ahmed et al [15]	Proposed approach
Accuracy	80.49%	87.15%	91.63%	94.13%	99.3%

References

[1] Śpibida, M., Krawczyk, B., Olszewski et al. “Modified DNA polymerases for PCR troubleshooting” J Appl Genetics , (2017): 133-142

[2] Michael Steinbach. (2005) Introduction to Data Mining. Pang Ning

[3] WHO. coronavirus disease (COVID-19) pandemic [online] From: https: //www.who.int/

[4] Shi, Heshui, Xiaoyu Han, Nanchuan Jiang, Yukun Cao, Osamah Alwalid, Jin Gu, Yanqing Fan, and Chuansheng Zheng. "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study." The Lancet infectious diseases 20, no. 4 (2020): 425-434.

[5] Mesut DEM˙IRKO¨ SE, Tu¨lay U¨ NVER ULUSOY et al. “The role of chest tomography in the diagnosis of COVID-19” (COVID -19) in china: Journal of Medicine and Palliative Care. (2022): 1-6

[6] Shi, Heshui, Xiaoyu Han, Yukun Cao, Osamah Alwalid, and Chuansheng Zheng. "CT screening for early diagnosis of SARS-CoV-2 infection–Authors' reply." The Lancet Infectious Diseases 20, no. 9 (2020): 1011.

[7] Tang, Ruifan, Lorenzo De Donato, Nikola Besinović, Francesco Flammini, Rob MP Goverde, Zhiyuan Lin, Ronghui Liu, Tianli Tang, Valeria Vittorini, and Ziyulong Wang. "A literature review of Artificial Intelligence applications in railway systems." Transportation Research Part C: Emerging Technologies 140 (2022): 103679.

[8] Gunasekeran, Dinesh Visva, Rachel Marjorie Wei Wen Tseng, Yih-Chung Tham, and Tien Yin Wong. "Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies." NPJ digital medicine 4, no. 1 (2021): 1-6.

[9] Narin, Ali, Ceren Kaya, and Ziynet Pamuk. "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks." Pattern Analysis and Applications 24, no. 3 (2021): 1207-1220.

[10] Cohen, Joseph Paul, Paul Morrison, and Lan Dao. "COVID-19 image data collection." arXiv preprint arXiv:2003.11597 (2020).

[11] Kaggle. Healthy Lung Classification Spectrogram Fast.ai [online] Available at:https://www.kaggle.com/dienhoa/healthy-lung-classification-spectrogram-fast-ai

[12] Hussain, Lal, Tony Nguyen, Haifang Li, Adeel A. Abbasi, Kashif J. Lone, Zirun Zhao, Mahnoor Zaib, Anne Chen, and Tim Q. Duong. "Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection." BioMedical Engineering OnLine 19, no. 1 (2020): 1-18.

[13] Akhtar, Asma, Samia Akhtar, Birra Bakhtawar, Ashfaq Ali Kashif, Nauman Aziz, and Muhammad Sheraz Javeid. "COVID-19 detection from CBC using machine learning techniques." International Journal of Technology, Innovation and Management (IJTIM) 1, no. 2 (2021): 65-78.

[14] Patel, Rajneesh Kumar, and Manish Kashyap. "Automated diagnosis of COVID stages from lung CT images using statistical features in 2-dimensional flexible analytic wavelet transform." Biocybernetics and Biomedical Engineering 42, no. 3 (2022): 829-841.

[15] Farid, Ahmed Abdullah, Gamal Ibrahim Selim, and Hatem Awad A. Khater. "A novel approach of CT images feature analysis and prediction to screen for corona virus disease (COVID-19)." (2020).

[16 Haralick, Robert M., and Linda G. Shapiro. "Image segmentation techniques." Computer vision, graphics, and image processing 29, no. 1 (1985): 100-132.

[17] Yunfei Ge, Qing Zhang, Yuantao Sun et al.” Grayscale medical image segmentation method based on 2D3D object detection with deep learn- ing”, BMC Med Imaging , (2022): 1471-2342

[18] Hanane Allioui, Mazin Abed Mohammed et al. “A Multi-Agent Deep Reinforcement Learning Approach for Enhancement of COVID-19 CT Image Segmentation”, Jorge Luis Espinoza, (2022): vol. 12,2 309

[19] Gould, Stephen et al.“Region-based segmentation and object detec- tion”Advances in neural information processing systems, (2009):22

[20] Ming-Chuan Chiu, Ho-Yen Tsai et al. “A novel directional object detec- tion method for piled objects using a hybrid region-based convolutional neural network” Advanced Engineering Informatics, (2022): 101448

[21] Suhaila, S., Low, Z.Y et al. “Image Segmentation Module Development for Image Processing Learning Mobile Application”. In Proceedings of the 12th National Technical Seminar on Unmanned System Technology, (2022): pp. 523-537

[22] Jamshidi, Ali, Shahrzad Faghih-Roohi, Siamak Hajizadeh, Alfredo Núñez, Robert Babuska, Rolf Dollevoet, Zili Li, and Bart De Schutter. "A big data analysis approach for rail failure risk assessment." Risk analysis 37, no. 8 (2017): 1495-1507.

[23] Masoudi, Aghdas, Mohammad Davarpanah Jazi, Majid Mohrekesh, and Reza Masoudi Nejad. "An investigation of rail failure due to wear using statistical pattern recognition techniques." Engineering Failure Analysis 134 (2022): 106084

[24] Jamshidi, Ali, Shahrzad Faghih-Roohi, Siamak Hajizadeh, Alfredo Núñez, Robert Babuska, Rolf Dollevoet, Zili Li, and Bart De Schutter. "A big data analysis approach for rail failure risk assessment." Risk analysis 37, no. 8 (2017): 1495-1507.

[25] Rahmani, Amir Masoud, Elham Azhir, Morteza Naserbakht, Mokhtar Mohammadi, Adil Hussein Mohammed Aldalwie, Mohammed Kamal Majeed, Sarkhel H. Taher Karim, and Mehdi Hosseinzadeh. "Automatic COVID-19 detection mechanisms and approaches from medical images: a systematic review." Multimedia Tools and Applications (2022): 1-20.

[26] Kaggle. Luisblanche/covidct [online] Available at: https://www.kaggle.com/luisblanche/covidct

[27] Kaggle.COVID-19 CT scans [online] Available at: https://www.kaggle.com/andrewmvd/covid19-ct-scans

[28] Kaggle. COVID CT [online] Available at https://www.kaggle.com/hgunraj/covidxct?select=2Aimages

[29] Lin, Fen, Yong-Hao Wu, Wen-Jian Deng, Pei-Biao Wu, Jian-Yong Chen, Shao-Huang Tang, Jin-Zhou Wen et al. "Clinical analysis and RNA findings in a family with SARS-CoV-2 infection." World Academy of Sciences Journal 2, no. 5 (2020): 1-1.