Turkish Journal of Computer and Mathematics Education Vol.12 No.9 (2021),1384-1392

Research Article

SVM-kNN- IPSO ensemble method for Diagnosis of Novel Coronavirus (COVID-19) with CT images

Wial Abbas Hanon[1], Tahseen A. Wotifi[2] , Mohammed G. Al-Hamiri [3]

[1]Computer Center- University of Babylon-Iraq
[2]Computer Center- University of Babylon-Iraq
[3]Computer Center- University of Babylon-Iraq
[1]wailh@uobabylon.edu.iq, [2]tahseen.ali@uobabylon.edu.iq,[3] m.ghadhban@uobabylon.edu.iq

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: New coronavirus epidemic- COVID- 19 is still growing. This epidemic disease not only includes high mortality due to viral infection but also caused the psychological disaster in all parts of the world. The paper provides the early Coronavirus stage detection COVID-19, with the methods of machine learning. Support vector machine (SVM) is a two-class classifier which in the recent years attracted a significant attention. The performance of this classifier depends on the amount of its parameters such as C (Penalty Factor) and the existing parameter in kernel. Also the selection of a suitable kernel function has a significant affect in its performance improvement. Besides the mentioned cases, performing the feature selection process not only causes to improve the mentioned performance improvement but also causes to reduce the computation complexity and training time. In this paper, we used the improved partial swarm optimization algorithm (IPSO) to optimize the SVM. Findings illustrated that proposed method could be utilized for diagnosing disease of COVID-19 as the assistant system. Promisingly, the proposed method can be regarded as a useful clinical decision tool for the physicians.

Keywords: Support vector machine (SVM), Partial Swarm Optimization algorithm (PSO), multi-objective optimization, COVID-19.

1. Introduction

New coronavirus-COVID-19 which was discovered in China in year 2019 for the first time and became a world epidemic from that time, now became the most difficult human tests in the modern history of world. As the confirmed cases of new coronavirus epidemic- COVID- 19 is growing, this virus kills more victims by overshadowing the health systems, shaken the foundations of world economy and led to sustained geopolitical developments [1-2]. Coronavirus disease has reached a state of pandemic. While this pandemic is rapidly growing in all over the world, it has caused fear and anxiety in public particularly among the specific groups such as older people, patient caregivers, healthcare providers and the people with underlying disease conditions. Therefore, more interventions are essential particularly in specific groups that they are at high risk for acute and persistent emotional distress [3].

Todays, the methods of machine learning have been utilized largely in domain of healthcare and in order to have more faster and effective prediction of COVID-19 infected person. Since the basic paradigm of machine learning and SVM are rooted in theory of Vapnik–Chervonenkis also minimization principle of structural risk. SVM tries for assigning tradeoff among reducing error of training set as well as increasing margin for obtaining the best ability of generalization and keep resistant for over fitting. In addition, the main SVM benefit is convex quadratic programing usage that presents just the world minima; therefore, this prevents to be trapped in local minima. Because of the advantageous aspect of it, SVM has been employed for the classification tasks wide range [4].

In paper [5], have improved lung CT diagnosis system based on deep learning for detecting patients with COVID-19 that is able to can extract new pneumonia radiographic features automatically particularly ground-glass opacity (GGO) from radiographs.

In paper [6], have presented the new method of Support Vector Regression for analyzing 5 various tasks correspondent to the new coronavirus. In the paper, instead of the easy line of regression they utilize supported vectors for getting the better accuracy of classification.

In paper [7], have improved the program of AI with analyzing representative images of CT by utilizing the method of deep learning. The study is diagnostic, multicolor and retrospective. They built the neuro network of Inception migration which obtained the accuracy of 82.9 percent.

In paper [8], have presented deep learning based on feature extractor as well as the classifier approach of machine learning for COVID-19 pneumonia computer-aided diagnosis. Some algorithms of ML were trained on features which are extracted by CNNs architectures that are well-established for finding the best learners' and features combination. Taking high image data visual complexity into account, accurate deep extraction of feature is taken as the critical stage in improving deep models of CNN.

The paper utilized 4000 images of CT [9] for classification of COVID-19. Datasets samples were labeled as negative or positive. Methods of selection and extraction of Feature and SVM are utilized during coronavirus images classification.

This paper remaining is ordered as below. Proposed method detailed implementation will be described in section 2. Section 3 defines proposed approach discussions also experimental results are provided Section 4. At last, conclusions are briefed in Section 5.

2. Proposed method

Basic goal is concentrating on proposed technique for grouping COVID-19 in either negative/positive. Proposed method basically includes five steps:

Acquisition of an Image, Pre-processing of an image, extraction of feature, selection of feature and the step of classification.

2-1- Acquisition of an Image

Generally, acquisition of an Image is obtaining the image process from various modalities of imaging like PET scan, CT scan, MRI scan image.

Input image of COVID-19 is taken from CT slices with the suspected pneumonia signs of COVID-19 which are marked by radiologists for every patient (negative, positive) in this study for ensuring that the whole images of patients are able to be uploaded in the capacity. Between them, CT images (training) from a hospital which is able to be shared in validation, training and datasets of test are saved in file of "Data_For_Training_Validation_Test.hdf5". Every part of CT images shape is (256, 256), down-sampling from (512, 512) for saving the space.

In [1] performed the retrospective research and enrolled 201 patients from 2 hospitals in China, From January 18 to February 23, 2020, who underwent the tests of chest CT and RT-PCR that 98 patients the tested positive for COVID-19 (83 females, 118 males with the 42 years middle age). CT images of Patient from hospital were shared between validation, training, datasets of test with ratio of 80 percent: 10 percent :10 percent.

2-2- Pre-processing of an image

Basic goal of stage of enhancement and pre-processing of image is not easily unwanted noise and background information removal but also image pixel amounts alteration. Equalization techniques of histogram are utilized to increase images through field of frequency as well as the field concept of spatial. Significantly, this develops image boundaries interpretability and perception for viewers of human. The Most popular filed techniques of frequency are equalization of histogram and this is utilized for developing images contrast this is appropriate for the all types of images.

2-3- Extraction of feature

Take for granted image is two various space variables function of y and x. It is expressed as f (x, y),

Where

$$x = 0.1. \ldots . N - 1 \tag{1}$$

$$y = 0.1. \ldots . M - 1 \tag{2}$$

G expresses whole intensity levels number in an image. Here, discrete amounts $i = 0.1. \ldots . G - 1$ are able to be taken by the function f (x,y). Pixels in whole image that have intensity of

$$h(i) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \delta(f(x.y).i) \tag{3}$$

$$\delta(j.i) = \begin{cases} 1. & j = i. \\ 0. & j \neq i. \end{cases}$$

(4)

That δ(j, i) expresses delta function of Kronecker.
Central moments are obtained from this for characterizing texture as it is described by equations below:
i) Mean:

$$\mu = \sum_{i=0}^{G-1} ip(i)$$

(5)

(ii) Variance:

$$\sigma^2 = \sum_{i=0}^{G-1} (i-\mu)^{-2} p(i)$$

(6)

(iii) Skewness:

$$\mu_4 = \sum_{i=0}^{G-1} (i-\mu)^{-3} p(i)$$

(7)

(iv) Kurtosis:

$$\mu_4 = \sigma^{-4} \sum_{i=0}^{G-1} (i-\mu)^{-4} p(i) - 3$$

(8)

2-3-1- Texture Features

Feature of Texture is achieved by Gabor wavelets help. Ma and Manjunath suggested the quality features of Gabor for retrieval due to the better performance of it in simultaneous autoregressive model that is multi-resolution, pyramid-structured and tree-structured wavelet transmit the features.

2-3-2- Grey-Level Co-occurrence Matrix (GLCM)

Grey-Level Co-occurrence Matrix is the statistical tool that is robust to extract second-order information of texture from various images. In addition, Grey-Level Co-occurrence Matrix is called as spatial dependence matrix of grey-level. Most essential stages in the process are (a) generating Co-Occurrence Matrix of grey-Level, (b) allocating Offsets (c) obtaining Statistics from Grey-Level Co-occurrence Matrix. Statistics obtained include:

*(*i*)* Contrast =

$$\sum_{i=0}^{N_{p-1}} \sum_{j=0}^{N_{p-1}} (i-j)^2 p(i.j)$$

(9)

(ii) Correlation =

$$\sum_{i=0}^{N_{p-1}} \sum_{j=0}^{N_{p-1}} \frac{(i-\mu)(j-\mu)p(i.j)}{\sigma_i \sigma_j}$$

(10)

(iii) Energy =

$$\sum_{i=0}^{N_{p-1}} \sum_{j=0}^{N_{p-1}} p^2(i.j) \tag{11}$$

(iv) Homogeneity =

$$\sum_{i=0}^{N_{p-1}} \sum_{j=0}^{N_{p-1}} \frac{p(i.j)}{1+|i-j|} \tag{12}$$

2-4- Classification and selection of feature

Usually, features are chosen with procedures of respective search. Search procedures number has been proposed already. The mostly utilized algorithms of feature selection are selection of sequential forward, selection of sequential backward, bound and branch, PSO, GA, Recursive Feature Elimination based on SVM (SVM-RFE). In contrary, utilizing the more features number causes to more costs of computation. For keeping a tradeoff among computational cost and accuracy, we have described the optimization technique of nature inspired called IPSO and the Whale Optimization algorithm for determining optimum features selection for the suitable corona virus image classification.

2-4-1- Improved SVM-RBF kernel Classifier

The improvements were made in the hyper plane by considering the soft margin C, the gamma (γ) value has been given accordingly and the classification was done such that the positive and negative were classified whereas the neutral can also be found. The general classifier with including weights of the sentences repeated can be given as,

$$f(s) = \sum_{i=1}^{n} \propto_i k(s.s') \tag{13}$$

f (s) is hyper plane. K (s,s') is function of kernel. ∝i expresses sentences weights. Vector of feature based on weights is able to be described as,

$$k(s.s') = \exp\left[-\gamma \|s-s'\|_2^2\right] \tag{14}$$

i.e.

$$k(s.s') = \exp\left[-\frac{\|s-s'\|^2}{2\sigma^2}\right] \tag{15}$$

where $\|s-s'\|^2$ is distance of Euclidean among them. σ is free parameter. From that classifier of RBF is able to be:

$$f(s) = \sum_{i=1}^{n} \propto_i \exp\left[-\frac{\|s-s'\|}{2\sigma^2}\right] \tag{16}$$

where, b is a constant.

2-4-2- Improved PSO and WOA-Based SVM

IPSO-WOA-SVM includes two steps: at first, the approach based on IPSO for SVM parameter optimization is improved for finding better kernel function initial parameters then WOA is employed for continuing training of SVM and find the best SVM parameters. The algorithm of IPSO-WOA-SVM is able to be illustrated in detail as below.

Algorithm 1:

Stage 1 – Initializing whale and particles with the size of population, generations, inertia weight, hyper parameters C and kernel range.

Stage 2 –employing IPSO for finding better kernel and initial parameters C.

Stage 3 – assessing every particle fitness.

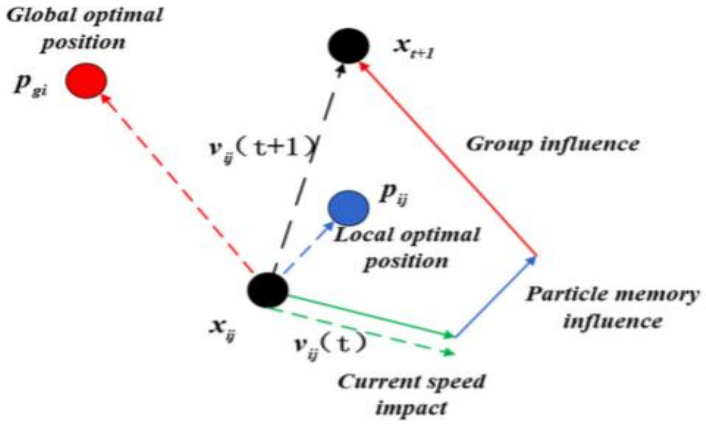
Stage 4 – comparison of values of fitness with detecting global best and local best particle.

Step 5 – Updating every particle position and velocity until fitness function converges value.

Step 6 – After converging, global best particle in swarm is fed to the classifier of SVM to train.
Step 7 – testing and Training the classifier of SVM.

2-4-3- Improved PSO

Therefore, the research utilizes heuristic algorithm of PSO for obtaining parameter optimization of SVM. In algorithm of PSO, one feasible optimization issue solution is known as particle. Position of particle is updated with constantly seeking the own optimal solution of it and existing population optimal solution. Iterative search is performed till global optimal solution is recognized. Every particle that is denoted as x_i [x_i1, x_{i2}, x_{i3},..., x_{in}] expresses the point in n-dimensional space of search. Suppose that pbesti [p_{i1}, p_{i2}, p_{i3},..., p_{in}] presents ith particle optimal solution and gbest [g_1, g_2, g_3,..., g_n] expresses population global optimal solution. ith particle movement velocity after iterations of t is donated as $v_i(t)$ [$v_{i1}(t)$, $v_{i2}(t)$, $v_{i3}(t)$,...,$v_{in}(t)$]. This is able to be updated as Figure 1 illustrates



Traditional algorithm of PSO lacks the ability of convergence in process of calculation. Thus, the factor k of shrinkage is added to an algorithm for suppressing and controlling velocity of particle. Shrinkage factor equation is as below:

$$k = \frac{2}{|2 - c - \sqrt{c^2 - 4c}} c = c_1 + c_2 . c > 4 \quad (17)$$

Algorithm of IPSO is as below:

$$v_{ij}(t+1) = ku_{ij}(t) + c_1 r_{1j}(t) * \left[p_{ij}(t) - x_{ij}(t)\right] + c_2 r_{2j}(t) * \left[p_{gj}(t) - x_{ij}(t)\right] \quad (18)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (19)$$

where t is the current iteration of the algorithm; $x_{ij}(t)$ is the current position of P_{ij} ;v_{ij} (t + 1) is velocity vector that applied to P_{ij} at time t; c_1 and c_2 are random values that represent the exploration and diversity component of the algorithm, c_1 is a cognitive learning factor, c_2 is a social learning factor. They usually follow a uniform distribution within the range [0, 1], $P_{ij}(t)$ is the local best of particle, $P_{gi}(t)$ is the global best of particle.

2-4-4- Hybrid Classifier of (KNN-SVM)

Classifier of SVM is equal to the classifier of KNN that selects (K = 1, i.e., 1NN) one point of representative for the vectors of support in per level. Class step (positive and negative class) an algorithm calculates distance from samples of test to SVM optimal hyper plane in space of feature. For distance of condition is greater than mentioned threshold, sample of test is grouped on SVM; otherwise algorithm of 1NN is utilized. In Hybrid classifier, we train SVM classifier. In the part of testing, we will compute the nearest neighbor (like vector of support) to point of query by utilizing KNN.

3. Evaluation result

In the part, achieved results for proposed method are discussed. First step is acquisition of image now image of input is obtained from the CT slices with suspected COVID-19 pneumonia signs marked by radiologists for each patient. After that, pre-processing of an image and process of enhancement is performed. After taking database, an image is read, after that process of resized is performed (Fig. 2).

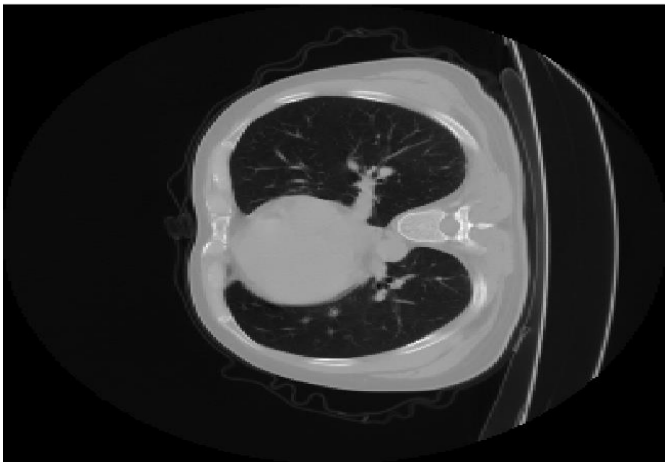


Fig. 2 an image of database

Second part is step of pre-processing. Now, input is given to the technique of pre-processing, the histogram of them is achieved. When utilizing the method, image is improved and increased image quality.

Table 1 illustrates that different kinds of image features of sample 10 image of COVID-19 extracted also the values of them are given. Whole 4000 images are taken as the input to test and train the features of them that are extracted from features extraction techniques of Texture, GLCM. From Table 1, decision is taken whether image is impacted by positive and negative.

Table 1: Sample of 10 lung cancer images extracted features value

Image no.	Contrast	Correlation	Energy	homogeneity	Mean	Standard deviation	Skewness	Kurtosis
1	**0.25919**	**0.97844**	**0.65891**	**0.99383**	**0.82965**	**0.32582**	**-1.3901**	**2.9335**
2	**0.25919**	**0.97844**	**0.65891**	**0.99383**	**0.82965**	**0.32582**	**-1.3901**	**2.9335**
3	**0.25919**	**0.97844**	**0.65891**	**0.99383**	**0.82965**	**0.32582**	**-1.3901**	**2.9335**
4	**0.25919**	**0.97844**	**0.65891**	**0.99383**	**0.82965**	**0.32582**	**-1.3901**	**2.9335**
5	**0.25919**	**0.97844**	**0.65891**	**0.99383**	**0.82965**	**0.32582**	**-1.3901**	**2.9335**
6	**0.26029**	**0.97835**	**0.65886**	**0.9938**	**0.82964**	**0.32588**	**-1.3894**	**2.9312**
7	**0.26029**	**0.97835**	**0.65886**	**0.9938**	**0.82964**	**0.32588**	**-1.3894**	**2.9312**
8	**0.25919**	**0.97844**	**0.65891**	**0.99383**	**0.82965**	**0.32582**	**-1.3901**	**2.9335**
9	**0.25919**	**0.97844**	**0.65891**	**0.99383**	**0.82965**	**0.32582**	**-1.3901**	**2.9335**
10	**0.25919**	**0.97844**	**0.65891**	**0.99383**	**0.82965**	**0.32582**	**-1.3901**	**2.9335**

3-1- Measures of Performance

New proposed method has been performed on MATLAB 2018b. For assessing results, parameter measures of performance are accuracy, specificity, sensitivity, precision, and F1 score criteria.

Selecting a criterion for assessing the method efficiency depends on the issue we are trying to solve. Assume that there are a number of data samples, these data are given to the model one-by-one and a class is taken for each of them as an output. The class predicted by the model and data real class can be shown in a table. This table is called confusion matrix.

Table 2. confusion matrix table

Predicted class label			
Patient	Healthy	Predicted/real	
False positive (FP)	True negative (TN)	Healthy	**Real class label**
True positive (TP)	False negative (FN)	Patient	

True positive: the samples which are accurately diagnosed by the test as patient.
False positive: the samples which are wrongly diagnosed by the test as patient.
True negative: the samples which are accurately diagnosed by the test as healthy.
False negative: the samples which are wrongly diagnosed by the test as healthy.

3-1-1- Accuracy criterion

The ability of a test in the accurate dissociation of healthy and patient cases from the other cases is called accuracy. In order to calculate the accuracy of a test we should derive the total samples of true positive and true negative to the total tested cases. Mathematically, this ratio can be expressed as below (equation 1):

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (20)$$

3-1-2- Specificity criterion

With utilizing the automated method, accurate background covid 19 percentage is able to be detected accurately. Specificity is described as below

$$Specificity = \frac{TN}{TN+FP} \quad (21)$$

3-1-3- sensitivity criterion

Sensitivity criterion is the other kind of criteria to show the rules of efficiency. It calculates the accurate percentage of being patient rules detection. If this amount is higher, the rules have higher detection capability.

$$Sensitivity = \frac{TP}{TP+FN} \quad (22)$$

3-1-4- precision criterion

Precision is another parameter that shows the prediction possibility of being patient and the accuracy of that prediction. If this amount is higher, it shows that it has more rules in diseases' detection.

$$Precision = \frac{TP}{TP+FP} \quad (23)$$

3-1-5- F1 score criterion

F1 score is defined as the harmonic average of sensitivity and precision.

$$F1 = \frac{2\times TP}{2TP+FP+FN} \quad (24)$$

3-2- Parameters' initialization

We set the parameters, in order to evaluate the proposed algorithm performance. Also, we determine the percentage of data classification for training and test equal to 80 and 20. In table 3, the settings related to parameters are shown.

Table 3 initial amounts for parameters

Parameters	Initial values
Train set percentage	80
Test set percentage	20
The number of searching factors in WOA	50
Number of Particle in IPSO	50
upper bound	1
lower bound	0
The maximum repetition IPSO and WOA	1000, 100

6.4. The evaluation of results

After implementing the proposed method in Matlab 2018b environment, we compared the proposed method and the other methods on the dataset. The results of comparing these two methods are mentioned in Table 6.

The swarm size and number of generations play important role in controlling the search ability of IPSO and WOA. Thus, we firstly investigated the impact of the five factors on the performance of IPSO and WOA. Different sizes of swarm from 10 to 50 were evaluated, the detailed results are presented in Table 4. From the table, we can see that the best performance was achieved when the swarm size is 50, where the accuracy, specificity, sensitivity, precision and F1 score are 90.75%, 84.90%, 84.90% , 85.91% and 85.4%, respectively.

Table 4. The detailed results of IPSO-WOA-SVM on the COVID-19 dataset

Sizes of swarm	**IPSO-WOA-SVM**				
	accuracy	specificity	sensitivity	precision	F1 score
20	81.15	73.04	73.04	72.05	72.52
30	92	90.53	90.53	86.64	88.36
40	90.25	79.32	79.32	88	82.66
50	**90.75**	**84.90**	**84.90**	**85.91**	**85.4**
60	70.5	60.51	62.51	59.95	60.51

In Table 6, we compare the performance results of the proposed algorithm for other criteria with other methods on 1500 images the COVID-19 dataset.

Table 6. The comparison results of the proposed algorithm with other methods

Methods	F1 score	precision	sensitivity	specificity	accuracy
[5]	94%	96%	92%	-	94%
[7]	77%	-	81%	84%	82.9%
Proposed method	90.76%	96.24%	87.19%	87.19%	94.26%

4 Conclusion

In the article, lung image of input is taken from the CT slices with suspected COVID-19 pneumonia signs marked by radiologists for each patient (both positive and negative). After that, obtained image is given to the pre-processing in order to the goal of enhancement. Proposed technique of WOA-SVM and IPSO presents the obvious optimized amount for classifier of SVM. Moreover, network was trained successfully for the both two CT scan image classes (positive and negative) with the middle 94.26 percent classification accuracy is achieved by utilizing Whale Optimization Algorithm and IPSO in comparison to state-of-the art in software of MATLAB 2018b. In the future work, we plan to apply the proposed method to other medical diagnosis problems.

References

A. World Economic Forum (2020); Strategic Intelligence: COVID-19; https:// intelligence.weforum.org
B. Li, H., Liu, S. M., Yu, X. H., Tang, S. L., & Tang, C. K. (2020). Coronavirus disease 2019 (COVID-19): current status and future perspective. International journal of antimicrobial agents, 105951.
C. Schoch-Spana, Monica. (2020). COVID-19's Psychosocial Impacts The pandemic is putting enormous stress on all of us but especially on health care workers and other specific groups. Scientific American. March 20. 2020. https:// blogs. sc ientificamerican. Com.
D. Shen, L., Chen, H., Yu, Z., Kang, W., Zhang, B., Li, H., ... & Liu, D. (2016). Evolving support vector machines using fruit fly optimization for medical data classification. Knowledge-Based Systems, 96, 61-75.
E. Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., ... & Chong, Y. (2020). Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. medRxiv.
F. Yadav, M., Perumal, M., & Srinivas, M. (2020). Analysis on novel coronavirus (covid-19) using machine learning methods. Chaos, Solitons & Fractals, 110050.
G. Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., ... & Xu, B. (2020). A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). MedRxiv.
H. Kassani, S. H., Kassasni, P. H., Wesolowski, M. J., Schneider, K. A., & Deters, R. (2020). Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning-Based Approach. arXiv preprint arXiv:2004.10641.

I. Song, J., Wang, H., Liu, Y., Wu, W., Dai, G., Wu, Z., ... & Deng, K. (2020). End-to-end automatic differentiation of the coronavirus disease 2019 (COVID-19) from viral pneumonia based on chest CT. European journal of nuclear medicine and molecular imaging, 1-9.

J. Vijila Rani, K., & Joseph Jawhar, S. (2019). Lung Lesion Classification Scheme Using Optimization Techniques and Hybrid (KNN-SVM) Classifier. IETE Journal of Research, 1-15.

K. Gopi, A. P., Jyothi, R. N. S., Narayana, V. L., & Sandeep, K. S. (2020). Classification of tweets data based on polarity using improved RBF kernel of SVM. International Journal of Information Technology, 1-16.

L. Liang, H., & Zou, J. (2020). Rock image segmentation of improved semi-supervised SVM–FCM algorithm based on chaos. Circuits, Systems, and Signal Processing, 39(2), 571-585.