 doi:10.20944/preprints202005.0052.v1

Social-Group-Optimization Assisted Kapur's Entropy and Morphological Segmentation for Automated Detection of COVID-19 Infection from Computed Tomography Images

Nilanjan Dey[1,*], V. Rajinikanth[2], Simon James Fong [3] M. Shamim Kaiser [4] Mufti Mahmud [5,*]

[1,*] Department of Information Technology, Techno India College of Technology, Kolkata-700156, West Bengal, India
[2] Department of Electronics and Instrumentation Engineering, St. Joseph's College of Engineering, Chennai 600119, India
[3] Department of Computer and Information Science, University of Macau, Taipa, China; and DACC Laboratory, Zhuhai Institutes of Advanced Technology of the Chinese Academy of Sciences, China
[4] IIT, Jahangirnagar University, Savar, 1342 - Dhaka, Bangladesh
[5,*] Dept. of Computing & Technology, School of Science & Technology, Nottingham Trent University, Nottingham, NG11 8NS, UK

[*] Co-corresponding author. Emails: neelanjan.dey@gmail.com (N. Dey); mufti.mahmud@ntu.ac.uk, muftimahmud@gmail.com (M. Mahmud)

Abstract

The Coronavirus disease (COVID-19) caused by a novel coronavirus, SARS-CoV-2, has been declared as a global pandemic. Due to its infection rate and severity, it has emerged as one of the major global threats of the current generation. To support the current combat against the disease, this research aims to propose a Machine Learning based pipeline to detect the COVID-19 infection using the lung Computed Tomography scan images (CTI). This implemented pipeline consists of a number of sub-procedures ranging from segmenting the COVID-19 infection to classifying the segmented regions. The initial part of the pipeline implements the segmentation of the COVID-19 affected CTI using Social-Group-Optimization and Kapur's Entropy thresholding, followed by k-means clustering and morphology-based segmentation. The next part of the pipeline implements feature extraction, selection and fusion to classify the infection. PCA based serial fusion technique is used in fusing the features and the fused feature vector is then employed to train, test and validate four different classifiers namely Random Forest, k-Nearest Neighbors (KNN), Support Vector Machine with Radial Basis Function, and Decision Tree. Experimental results using benchmark datasets show a high accuracy ($>$ 91%) for the morphology-based segmentation task and for the classification task the KNN offers the highest accuracy among the compared classifiers ($>$ 87%). However, this should be noted that this method still awaits clinical validation, and therefore should not be used to clinically diagnose the ongoing COVID-19 infection.

 doi:10.20944/preprints202005.0052.v1

Introduction

Lung infection caused by COVID-19 has emerged as one of the major diseases and has affected over 2.5 million population globally[1], irrespective of their race, gender and age. The infection and the morbidity rate caused by this novel corona virus is increasing rapidly [1,2]. Due to its severity and progression rate, the recent report of the World Health Organization (WHO) declared it as Pandemic [3]. Even though an extensive number of precautionary schemes have been implemented; the occurrence rate of COVID-19 infection is rising rapidly due to various circumstances.

The origin of COVID-19 is due to a virus called Severe Acute Respiratory Syndrome-Corona Virus-2 (SARS-CoV-2) and this syndrome initially started in Wuhan, China, in December 2019 [4]. The outbreak of COVID-19 has appeared as a worldwide problem and a considerable amount of research works are already in progress to determine solutions to manage the disease infection rate and spread. Further, the recently proposed research works on (i) COVID-19 infection detection [5–8], (ii) handling of the infection [9,10] and (iii) COVID-19 progression and prediction [11–13] has helped to get more information regarding the disease.

The former research and the medical findings discovered that COVID-19 initiates disease in the human respiratory tract and builds severe acute pneumonia. The existing research also confirmed that the premature indications of COVID-19 are subclinical and it necessitates a committed medical practice to notice and authenticate the illness. The frequent medical grade analysis engages in a collection of samples from infected persons and sample supported examination and confirmation of COVID-19 using Reverse Transcription-Polymerase Chain Reaction (RT-PCR) test and image-guided assessment employing lung Computed Tomography scan images (CTI), and the Chest X-ray [14–17]. When the patient is admitted with a COVID-19 infection, the doctor will initiate the treatment process to cure the patient using the prearranged treatment practice which will decrease the impact of pneumonia.

Usually experts recommend a chain of investigative tests to identify the cause, position, and harshness of pneumonia. The preliminary examinations, such as blood tests and pleural-fluid assessment are performed clinically to detect the severity of the infection [18–20]. The image assisted methods are also frequently implemented to sketch the disease in the lung, which can be additionally examined by an expert physician or a computerized arrangement to recognize the severity of the pneumonia. Compared to chest X-ray, CTI is frequently considered due to its advantage and the 3-D view. The research work published on COVID-19 also confirmed the benefit of the CT in detecting the disease in the respiratory tract and pneumonia [21,22].

Recently, more COVID-19 detection methods have been proposed for the progression stage identification of COVID-19 using the RT-PCR and imaging methods. Most of these existing works combined RT-PCR with the imaging procedure to confirm and treat the disease. The recent work of Rajinikanth et al. [8] developed a computer-supported method to assess the COVID-19 lesion using the lung CTI. This work implemented few operators assisted steps to achieve superior outcomes during the COVID-19 evaluation.

Machine Learning (ML) approaches are well-known for their capabilities in recognizing patterns in data. In recent years ML has been applied to a variety of tasks including biological data mining [23,24], medical image analysis [25], financial forecasting [26], trust management [27], anomaly detection [28,29], disease detection [30,31], natural language processing [32] and strategic game playing [33]. The proposed research aims to develop a ML-driven pipeline to extract and detect the COVID-19 infection from lung CTI with an improved accuracy. This work initially implements a series of procedures for an automated extraction of the COVID-19

[1] www.worldometers.info/coronavirus/, as of 23/04/2020

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

infection from the benchmark lung CTI [34]. This work executed a sequence of techniques, such as tri-level thresholding based on Social Group Optimization based Kapur's Entropy (SGO-KE), k-means clustering based separation, morphology-based segmentation to extract COVID-19 infection. Later, the segmentation accuracy of the proposed method is confirmed by executing a comparative study among the extracted COVID-19 infection with the Ground Truth (GT) images. During this work, 78 numbers of images from the benchmark dataset are considered and the proposed procedure is implemented using grayscale images of dimension $256 \times 256 \times 1$ pixels and the mean segmentation accuracy achieved in this work is $> 91\%$. Finally, the achieved classification accuracy is $> 87\%$.

Motivation

The proposed research work is motivated by the former image examination works existing in the literature [35–38]. During the mass disease screening operation, the existing medical data amount will gradually increase and to reduce the data burden, it is essential to employ an image segregation system to categorize the existing medical data into two or multi-class, to assign the priority during the treatment implementation. The recent works in the literature confirm that the feature-fusion based methods will improve the classification accuracy without employing the complex methodologies [39–41]. Classification task implemented using the features of the original image and the Region-Of-Interest (ROI) offered superior result on some image classification problems and this procedure is recommended when the similarity between the normal and the disease class images are more [23, 25, 30, 42, 43]. Hence, for the identical images, it is necessary to employ a segmentation technique to extract the ROI from the disease class image with better accuracy [44–46]. Finally, the fused features of the actual image and the ROI are fused to attain enhanced classification accuracy.

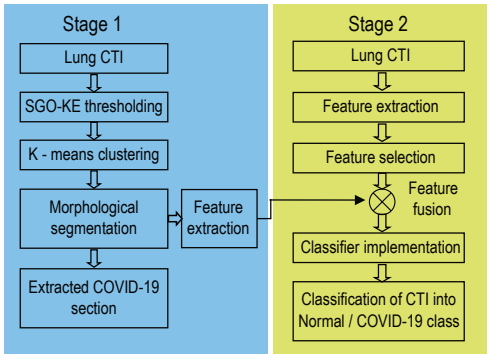

Figure 1. The number of image processing stages implemented in the proposed work.

Methodology

This section of the work presents the methodological details of the proposed scheme. Like the former approaches, this work also implemented two different phases to improve the detection accuracy.

Proposed pipeline

This work consists of the following two stages as depicted in Figure 1. These are–

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

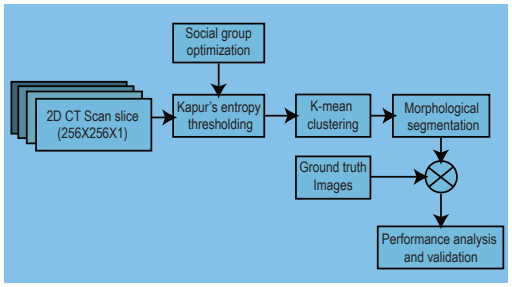

Figure 2. Image segmentation framework to extract COVID-19 infection from 2D lung CT scan image.

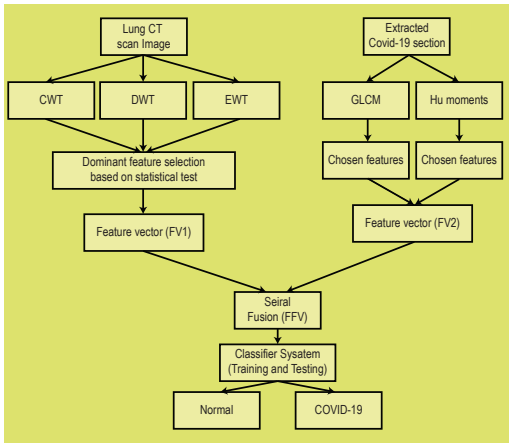

Figure 3. Proposed ML scheme to detect COVID-19 infection.

- Implementation of an image segmentation method to extract the COVID-19 infection,
- Execution of a ML scheme to classify the considered lung CTI database into normal/COVID-19 class.

The detail of these two stages are given below–

Stage 1: Figure 2 depicts the image processing system proposed to extract the pneumonia infection in the lung due to COVID-19. Initially, the required 2D slices of the lung CTI are collected from an open source database [47]. All the collected images are resized into $256 \times 256 \times 1$ pixels and the normalized images are then considered for evaluation. In this work, Social-Group-Optimization (SGO) and Kapur's Entropy (SGO-KE) based tri-level threshold is initially applied to enhance the lung section. Then, K-means clustering is employed to segregate the thresholded image into background, artifact and the lung segment. The unwanted lung sections are then removed using a morphological segmentation procedure and the extracted binary image of the lung is then compared with its related GT provided in the database. Finally, the essential performance measures are computed and based on which the performance of the proposed COVID-19 system is validated.

Stage 2: Figure 3 presents the proposed ML scheme to separate the considered lung CTI into normal/COVID-19 class. This system is constructed using two different images, such as (i) the original test image (normal / COVID-19 class) and (ii) The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

binary form of the COVID-19 section. The various procedures existing in the proposed ML scheme is depicted in Figure 3.

Segmentation of COVID-19 infection

This procedure is implemented only for the CTI associated with the COVID-19 pneumonia infection. The complete details on various stages involved in this process are depicted in Figure 1. The series of procedures implemented in this Figure is used to extract the COVID-19 infection from the chosen test image with better accuracy. The pseudo-code of the implemented procedure is depicted in Algorithm 1.

Algorithm 1: Pseudo-code of the implemented procedure.

1. Initialization: number of agents ($N = 30$), search dimension ($D = 3$), objective func. ($F_{KE}(Th)$), total iteration ($Iter_{max} = 2500$) and stopping criteria;
2. Randomly initialize the agents in the D-dimensional search space and compute $F_{KE}(Th)$ for each agent;
3. Find the best $F_{KE}(Th)$ for an agent and make $F_{KE}(Th) = g_{best}$. Improve the knowledge level of other agents for attained g_{best};

for $I = 1 : N$ **do**
  for $J = 1 : D$ **do**
    $x_{new_{I,J}} = x_{old_{I,J}} * \zeta + \mathcal{R} * (g_{best,J} - x_{old_{I,J}})$;
  end
end

where, $\zeta = 0.2$, $\mathcal{R}$ = random value [0,1] and $x_{new_{I,J}}$ = updated knowledge of agents;

5. Initiate the Acquiring phase to update all the agents based on the attained value of g_{best};
6. Check condition:

if *Is* $Iter_{max}$ *reached (or) all the agents attained* $F_{KE}(Th)$ **then**
  Stop and declare the thresholded image;
else
  Repeat steps 2 to 4, till is attained;
end

7. Execute the KMC on the thresholded image and separate the image into three groups, such as background, normal section and abnormal section.
8. Implement the morphology assisted segmentation of lung-infection based on a chosen marker value.
9. Compare extracted section and GT and compute the performance values.

Image thresholding Initially, the enhancement of the infected pneumonia section is achieved by implementing a tri-level threshold based on SGO and the KE. In this operation, the role of the SGO is to randomly adjust the threshold value of the chosen image till KE is maximized. The threshold which offered the maximized KE is considered as the finest threshold. The related information on the SGO-KE implemented in this work can be found in [48]. The SGO parameters discussed in Dey et al. [49] is considered in the proposed work to threshold the considered CTI.

The proposed SGO-KE threshold is defined below:

Entropy in an image is the measure of its irregularity and for a considered image, the Kapur's thresholding can be used to identify the optimal threshold by maximizing its entropy value. Let, $Th = \{t_1, t_2, ..., t_{n-1}\}$ denotes the threshold vector of chosen

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

image. The Kapur's entropy (KE) to be maximized is given by equation 1

$$KE_{max} = F_{KE}(Th) = \sum_{j=1}^{n} G_j^C. \tag{1}$$

For a tri-level thresholding problem, the expression will be given by equation 2

$$G_i^C = \sum_{j=1}^{t_i} \frac{P_j^C}{w_{i-1}^C} \ln\left(\frac{P_j^C}{w_{i-1}^C}\right), \tag{2}$$

where, $i = 1, 2, 3$, $j = 1, t_1, t_2$ G_i^C is entropy, P_j^C is the probability distribution for intensity, C is the image class ($C = 1$ for the gray scale image) and w_{i-1}^C is the probability occurrence.

During the tri-level thresholding, a chosen approach is employed to find the $F_{KE}(Th)$, by randomly varying the thresholds ($Th = \{t_1, t_2, t_3\}$). In this research, the SGO is employed to adjust the thresholds to find the $F_{KE}(Th)$.

Segmentation based on K-means clustering and morphological process The COVID-19 infection from the enhanced CTI is then separated using the K-means clustering (KMC) technique and this approach helps to segregate the image into various regions [50, 51]. In this work, the enhanced image is separated into three sections, such as the background, normal image section and the COVID-infection. The essential information on KMC and the morphology-based segmentation can be found in [52]. The extracted COVID-19 is associated with the artifacts and hence, a morphological enhancement and segmentation discussed in [53, 54] is implemented to extract the pneumonia infection, with better accuracy.

KMC helps to split the u-observations of image into K-groups. Let, given set of observations, of aspect 'd'. Then, KMC try to split given u-inspection in to K-groups; $Q(Q_1, Q_2, ..., Q_K)$ for $(K \leq u)$; to shrink the within-cluster sum of squares depicted by equation 3:

$$\arg\min_{Q} \sum_{i=1}^{K} ||O_i - \mu_i||^2 = \arg\min_{Q} \sum_{i=1}^{K} |Q_i| Var(Q_i) \tag{3}$$

where O is the number of observations, Q is the number of splits and μ_j is the mean of points in Q_i .

Performance computation The outcome of the morphological segmentation is in the form of binary and this binary image is then compared against the binary form of the GT and then the essential performance measures, such as accuracy, precision, sensitivity, specificity, and F1-Score are computed. A similar procedure is implemented on all the 78 images existing in the benchmark COVID-19 database and the mean value of these measures are then considered to confirm the segmentation accuracy of the proposed technique. The essential information on these measures is clearly presented in [55, 56].

Implementation of Machine Learning Scheme

The ML procedure implemented in this research is briefed in this section. This scheme implements a series of procedures on the original CTI (normal/COVID-19 class) and the segmented binary form of the COVID-19 infection as depicted in Figure 2. The main objective of this ML scheme is to segregate the considered CTI database into normal/COVID-19 class images. The process is shown in algorithm 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

Algorithm 2: Pseudo-code of the ML Scheme

1. Extract the features from the original image (CWT, DWT, EWT) as well as the binary segment of the COVID-19 infection (Haralick, Hu moments);
2. Implement a suitable statistical procedure (student's t-test) for feature vectors and sort features based on its p-value and the t-value;
3. Select all the features having the p-value < 0.05 as the dominant features;
4. Sort the values of the 1D FV by implementing the PCA analysis;
5. With the help of the serial-fusion technique, fuse the sorted features in order to get a 1D fused-feature-vector;
6. Train and test the considered classifier system using the chosen feature vector;
7. Validate the classifier system using the considered image database and compute the performance measured attained for the proposed system;
8. Analyse the classification accuracy and identify the best classifier for the proposed system;

Initial processing This initial processing of the considered image dataset is individually executed for the test image and the segmented COVID-19 infection. The initial processing involves extracted the image features using a chosen methodology and formation of a one-dimensional feature vector using the chosen dominant features.

Grayscale image feature-vector The accuracy of disease detection using the ML technique depends mainly on the considered image information. In the literature, a number of image feature extraction procedures are discussed to examine a class of medical images [35–37,39–42]. The well-known image feature extraction methods, such as Complex-Wavelet-Transform (CWT), Discrete-Wavelet-Transform (DWT) as well as Empirical-Wavelet-Transform (EWT) are considered in 2-D domain to extract the features of the normal/COVID-19 class gray scale images in this work. The information on the CWT, DWT and EWT are clearly discussed in the earlier works [56]. After extracting the essential features using these methods, a statistical evaluation and student's t-test based validation is implemented to select the dominant features to create the essential feature vectors, such as FV_{CWT} (34 features), FV_{DWT} (32 features) and FV_{EWT} (3 features) are considered to get the principle feature-vector set ($FV1$=69 features)by sorting arranging these features based on its p-value and t-value. The implementation of the feature selection process and $FV1$ creation is implemented as discussed in [56].

- CWT: This function was derived from the Fourier transform and is represented using complex-valued scaling function and complex-valued wavelet as defined below;

$$\psi_C(t) = \psi_R(t) + \psi_I(t) \tag{4}$$

  where $\psi_C(t)$,$\psi_R(t)$ and $\psi_I(t)$ represent the complex, real and image parts respectively.

- DWT: This approach evaluates the non-stationary information. When a wavelet has the function $\psi(t) \in W^2(r)$, then its DWT (denoted by $DWT(a,b)$) can be written as

$$DWT(a,b) = \frac{1}{\sqrt{2^a}} \int_{-\alpha}^{\alpha} x(t)\psi^* \left(\frac{t - b2^a}{2^a} \right) dt \tag{5}$$

  where $\psi(t)$ is the principle wavelet, the symbol $*$ denote the complex conjugate, a and b $(a, b \in R)$ are scaling parameters of dilation and transition respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

- EWT: The Fourier spectrum of EWT of range 0 to π is segmented into M regions. Each limit is denoted as ω_m (where $m = 1, 2, ..., M$) in which the starting limit $\omega_0 = 0$ and final limit is $\omega_M = \pi$. The translation phase T_m is centered around ω_m has width of $2\Phi_m$ where $\Phi_m = \lambda\omega_m$ for $0 < \lambda < 1$. Other information on EWT can be found in [57].

FV2 The essential information from the binary form of COVID-19 infection image is extracted using the feature extraction procedure discussed in Bhandary et al. [35] and this work helped to get the essential binary features using the Haralick and Hu technique. This method helps to get 27 numbers of features ($F_{Haralick} = 18$ features and $F_{Hu} = 9$ features) and the combination of these features helped to get the 1D feature-vector ($FV2$=27 features).

- *Haralick features*: Haralick features are computed using a Gray Level Co-occurrence Matrix (GLCM). GLCM is a matrix, in which the total rows and columns depends on the gray-levels (G) of the image. In this, the matrix component $P(i, j|\Delta x, \Delta y)$ is the virtual frequency alienated by a pixel space $(\Delta x, \Delta y)$. If μ_x and μ_y represents the mean and σ_x and σ_y represents the standard deviation of P_x and P_y, then–

$$\begin{aligned} \mu_x &= \sum_{i=0}^{G-1} iP_x(i), \\ \mu_y &= \sum_{j=0}^{G-1} jP_y(j), \\ \sigma_x &= \sum_{i=0}^{G-1} (P_x(i) - \mu_x(i)) \\ \sigma_y &= \sum_{j=0}^{G-1} (P_y(j) - \mu_y(j)). \end{aligned} \tag{6}$$

  where $P_x(i)$ and $P_y(j)$ matrix components during the i-th and j-th entry, respectively.

  These parameters can be used to extract the essential texture and shape features from the considered grayscale image.

- *Hu moments*: Let, for a two-dimensional (2D) image, the 2D $(i + j)$-th order moments can be defined as;

$$M_{ij} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^i y^j f(x, y) dx dy \tag{7}$$

  for $i, j = 0, 1, 2, ...$

  If, the image function $f(x, y)$ is a piecewise continuous value, then the moments of all order exists and the moment sequence M_{ij} is uniquely determined. Other information on Hu moments can be found in [35].

FFV In this work, the original test image helped to get the $FV1$ and the binary form of the COVID-19 helps to get the $FV2$. To implement a classifier, it is essential to have a single feature vector with a pre-defined dimension. In this work, a SFF based on the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

PCA is implemented to attain a 1D FFV (69+27=96 features) by combining the FV1 and FV2 and this feature set is then considered to train, test and validate the classifier system implemented in this study. The complete information on the feature fusion based on the serial fusion can be found in [35,58].

Classification In classification is one of the essential parts in a verity of ML and Deep Learning (DL) techniques implemented to examine a class of medical datasets. The role of the classifier is to segregate the considered medical database into two-class and multi-class information using the chosen classifier system. In the proposed work, the classifiers, such as Random-Forest (RF), k-Nearest Neighbors (KNN), SVM-RBF, and Decision Tree (DT) are considered. The essential information on the implemented classifier units can be found in [35,36,48,56]. A five-fold cross validation is implemented and the best result among the trial is chosen as the final classification result.

Validation From the literature, it can be noted that the performance of the ML and DL based data analysis is normally confirmed by computing the essential performance measures [35,36]. In this work, the common performance measures, such as accuracy (equation 8), precision (equation 9), sensitivity (equation 10), specificity (equation 11), F1-Score (equation 13) and Negative Predictive Value (NPV) (equation 12) computed.

The mathematical expression for these values is as follows:

$$\text{Accuracy} = \frac{(T_P + T_N)}{(T_P + T_N + F_P + F_N)} \tag{8}$$

$$\text{Precision} = \frac{T_P}{(T_P + F_P)} \tag{9}$$

$$\text{Sensitivity} = \frac{T_P}{(T_P + F_N)} \tag{10}$$

$$\text{Specificity} = \frac{T_N}{(T_N + F_P)} \tag{11}$$

$$\text{F1-Score} = \frac{2T_P}{(2T_P + F_N + F_P)} \tag{12}$$

$$\text{NPV} = \frac{T_N}{(T_N + F_N)} \tag{13}$$

where T_P= true positive, T_N= true negative, F_P= false positive and F_N=false negative.

COVID-19 dataset

The clinical level diagnosis of the COVID-19 pneumonia infection is normally assessed using imaging procedure. In this research, the lung CTI are considered for the examination and these images are resized into $256 \times 256 \times 1$ pixels to reduce the computation complexity. This work considered 400 numbers of grayscale lung CTI (200 normal and 200 COVID-19 class images) for the assessment. This research initially considered the benchmark COVID-19 database of [47] for the assessment. This dataset consists 100 numbers of 2D lung CTI along with its GT and in this research; only 78 images are considered for the assessment and the remaining 22 images are discarded due to its poor resolution and the associated artifacts. The remaining COVID-19 CTI (122 images) are collected from the Radiopaedia database [59] from cases 3 [60], 8 [61], 23 [62], 10 [63], 27 [64] 52 [65], 55 [66] and 56 [67].

The normal class images of the 2D lung CTI have been collected from LIDC-IDRI [68–70] and the RIDER-TCIA [70,71] database and the sample images of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

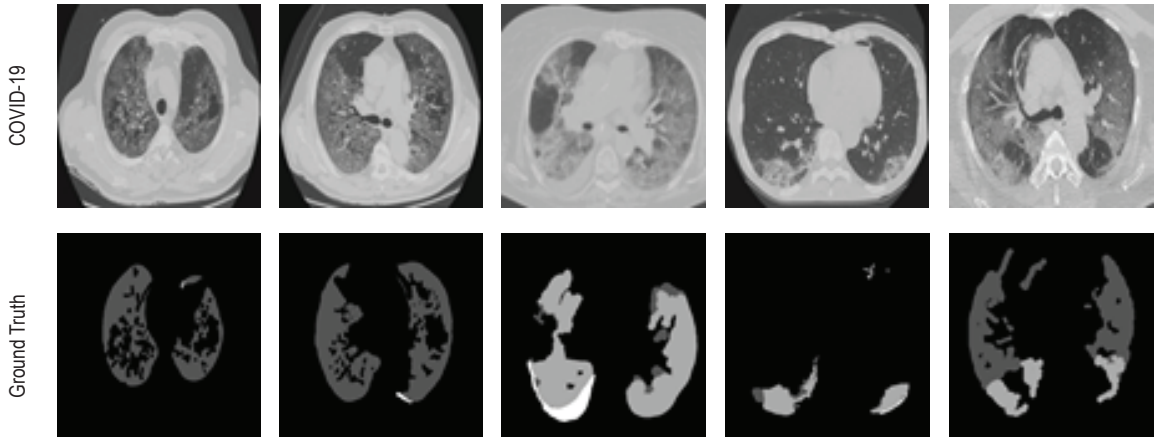

Figure 4. Sample test images of COVID-19 and the GT collected from [24].

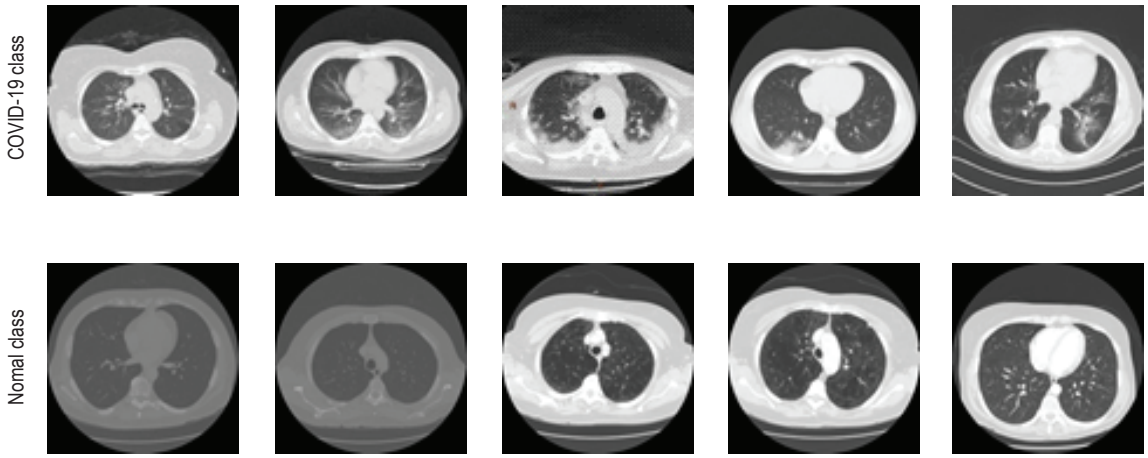

Figure 5. Sample test images of COVID-19 and normal group.

the collected dataset is depicted in Figures 4 and 5. Figure 4 presents the test image and the related GT of the benchmark CTI. Figure 5 depicts the images of the COVID-19 [59] and normal lung [68, 71] CTI considered for the assessment.

Results and Discussion

The experimental results obtained in the proposed work are presented and discussed in this section. This developed system is executed using a workstation with configuration-Intel i5 2.GHz processor with 8GB RAM and 2GB VRAM equipped with the MATLAB (www.mathworks.com). Experimental results of this study confirm that this scheme requires a mean time of 173 ± 11sec to process the considered CTI dataset and the processing time can be improved by using a workstation with higher computational capability. The advantage of this scheme is, it is a fully automated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

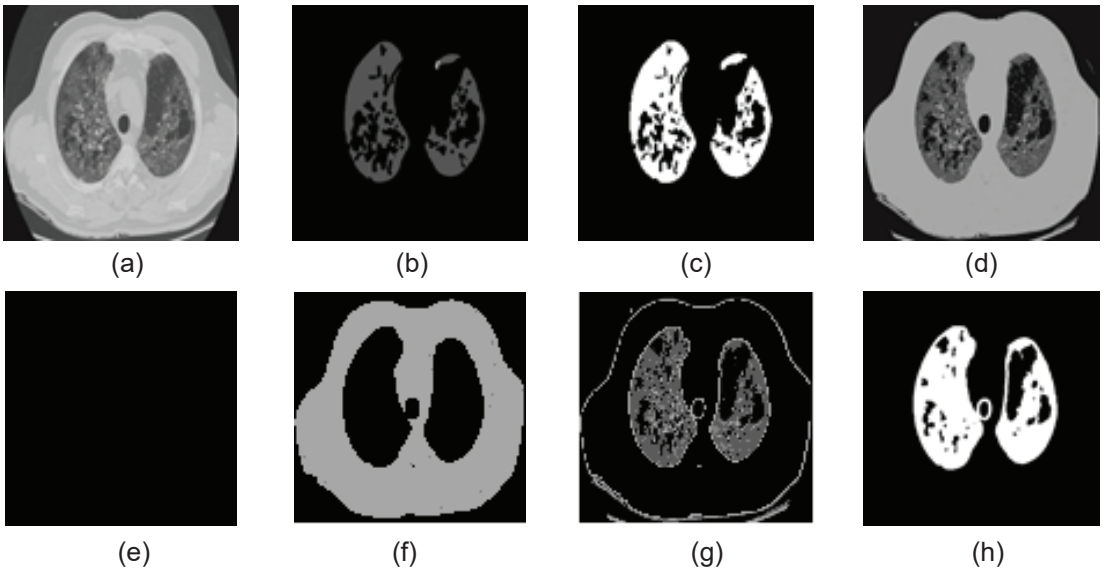

Figure 6. Results attained with the benchmark COVID-19 database. (a) Sample test image, (b) FT image, (c) Binary GT, (d) SGO-KE thresholded image, (e) Background, (f) Artifact, (g) Lung section, (h) Segmented COVID-19 infection.

practice and will not require the operator assistance during the execution. The proposed research initially executes the COVID-19 infection segmentation task using the benchmark dataset of [47]. The results attained using a chosen trial image is depicted in Figure 6. Figure 6(a) depicts the sample image of dimension $256 \times 256 \times 1$ and Figures 6(b) and 6(c) depicts the actual and the binary form of the GT image. The result attains with the SGO-KE-based tri-level threshold is depicted in Figure 6(d). Later, the k-means clustering is employed to segregate Figure 6(d) into three different section*s and the separated images are shown in Figure 6(e)-(g). Finally, a morphological segmentation technique is implemented to segment the COVID-19 infection from Figure 6(g) and the attained result is presented in Figure 6(h). After extracting the COVID-19 infection from the test image, the performance of the proposed segmentation method is confirmed by implementing a comparative examination between the binary GT existing in Figure 6(c) with Figure 6(h) and the essential performance values are then computed based on the pixel information of the background (0) and the COVID-19 section (1). For this image, the values attained are as follows; $T_P = 5865$ pixels, $F_P = 306$, $T_N = 52572$, and $F_N = 1949$ and these values offered; accuracy=96.28%, precision=95.04%, sensitivity= 75.06%, specificity=99.42%, F1-Score=83.88% and NPV=96.43%.

A similar procedure is implemented for other images of this dataset and means performance measure attained for the whole benchmark database (78 images) is depicted in the Figure 7. From this figure, it is evident that the segmentation accuracy attained for this dataset is higher than 91% and in the future, the performance of the proposed segmentation method can be validated against other thresholding and segmentation procedures existing in the medical imaging literature.

The methodology depicted in Figure 3 is then implemented by considering the entire database of the CTI prepared in this research work. This dataset consists of 400 grayscale images with dimension $256 \times 256 \times 1$ pixels and the normal/COVID-19 class images have a similar dimension to confirm the performance of the proposed technique. Initially, the proposed ML scheme is implemented by considering only the grayscale

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

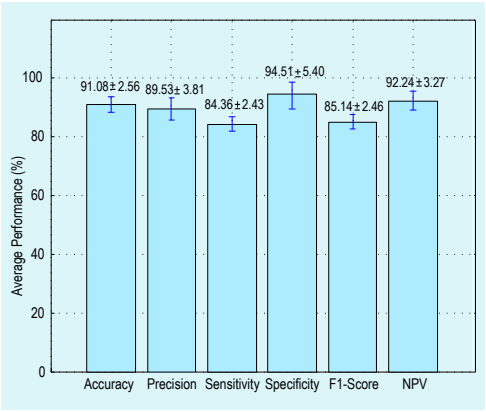

Figure 7. Mean performance measure attained with the proposed COVID-19 segmentation procedure.

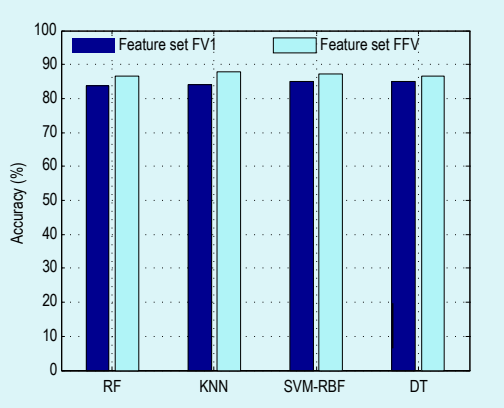

Figure 8. Detection accuracy attained in the proposed system with various classifiers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2020 doi:10.20944/preprints202005.0052.v1

Table 1. Disease detection performance attained with the proposed ML scheme

Features	Classifier	TP	FN	TN	FP	Acc.	Prec.	Sens.	Spec.	F1-Sc.	NPV
FV1 (1×69)	RF	163	37	172	28	83.75	85.34	***81.50***	86.00	83.37	***82.30***
	KNN	159	41	177	23	84.00	87.36	79.50	88.50	83.24	81.19
	SVM-RBF	161	39	179	21	***85.00***	***88.46***	80.50	***89.50***	***84.29***	82.11
	DT	160	40	168	32	82.00	83.33	80.00	84.00	81.63	80.77
FFV (1×96)	RF	169	31	178	22	86.75	***88.48***	84.50	***89.00***	86.45	85.17
	KNN	178	22	173	27	***87.75***	86.83	***89.00***	86.50	***87.90***	***88.72***
	SVM-RBF	172	28	177	23	87.25	88.20	86.00	88.50	87.09	86.34
	DT	174	26	172	28	86.50	86.14	87.00	86.00	86.57	86.89

Legend: TP: True Positive; FN: False Negative; TN: True Negative; FP: False Positive; Acc.: Accuracy; Prec.: Precision; Sens.: Sensitivity; Spec.: Specificity; F1-Sc.: F1-Score; NPV: Negative Predictive Value.

image features (FV1) with a dimension 1×69 and the performance of the considered classifier units, such as RF, KNN, SVM-RBF, and DT are computed. During this procedure, 70% of the database (140+140=280 images) is considered for training and 30% (60+60=120 images) are considered for testing. After checking its function, each classifier is separately validated by using the entire database and the attained results are recorded. Here, a five-fold cross validation is implemented for each classifier and the best result attained is considered as the final result. The obtained results are depicted in Table 1 (first three rows). The results reveal that the classification accuracy attained with SVM-RBF is superior (85%) compared to the RF, KNN and DT. Also, the RF technique helped to get the better values of the sensitivity and NPV compared to other classifiers.

To improve the detection accuracy, the feature vector size is increased by considering the FFV (1×96 features) and a similar procedure is repeated. The obtained results (as in Table 1, bottom three rows) with the FFV confirm that the increment of features improves the detection accuracy considerably and the KNN classifier offers an improved accuracy (higher than 87%) compared to the RF, SVM-RBF, and DT. The precision and the F1-Score offered by the RF is superior compared to the alternatives. The experimental results attained with the proposed ML scheme revealed that this methodology helps to achieve better classification accuracy on the considered lung CTI dataset. The accuracy attained with the chosen classifiers for FV1 and FFV is depicted in Figure 8. The future scope of the proposed method includes - (i) Implementing the proposed ML scheme to test the clinically obtained CTI of COVID-19 patients, (ii) Enhancing the performance of implemented ML technique by considering the other feature extraction and classification procedures existing in the literature, and (iii) Implementing and validating the performance of the proposed ML with other ML techniques existing in the literature, and (iv) Implementing an appropriate DL architecture to attain better detection accuracy on the benchmark as well as the clinical grade COVID-19 infected lung CTI.

Conclusion

The aim of this work has been to develop an automated detection pipeline to recognize the COVID-19 infection from lung CTI. This work proposes an ML-based system to achieve this task. The proposed system executed a sequence of procedures ranging from image pre-processing to the classification to develop a better COVID-19 detection tool. The initial part of the work implements an image segmentation procedure with; SGO-KE thresholding, k-means clustering based separation, morphology based COVID-19 infection extraction and a relative study between the extracted COVID19

sections with the GT. Th segmentation assisted to achieve an overall accuracy higher than 91% on a benchmark CTI dataset. Later, an ML scheme with essential procedures such as feature extraction, feature selection, feature fusion, classification is implemented on the considered data and the proposed scheme with the KNN classifier achieved an accuracy higher than 87%.

Acknowledgements

The authors of this paper would like to thank Medicalsegmentation.com and Radiopaedia.org for sharing the clinical grade COVID-19 images.

Compliance with Ethical Standards

Funding: This research received no external funding.
Conflicts of Interest: All authors declare that they have no conflict of interest.
Ethical Approval: All procedures reported in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.
Informed Consent: This study used secondary data, therefore, the informed consent does not apply.
Authors and Contributors: This work was carried out in close collaboration between all co-authors. ND, VR, MSK, and MM first defined the research theme and contributed an early design of the system. ND and VR further implemented and refined the system development. ND, VR, SJF, MSK, and MM wrote the paper. All authors have contributed to, seen and approved the final manuscript.

References

1. WHO. WHO, editor. Coronavirus. WHO; 2020. Last accessed: 10th April 2020. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

2. (ALA) ALA. ALA, editor. Coronavirus Update- Worldometer. ALA; 2020. Available from: https://www.worldometers.info/coronavirus/.

3. WHO. WHO, editor. WHO/Europe | Coronavirus disease (COVID-19) outbreak - WHO announces COVID-19 outbreak a pandemic. WHO; 1948. Last access date: 22-04-2020. Available from: https://bit.ly/3bvuX8S.

4. Li Q, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New England Journal of Medicine. 2020 Mar;382(13):1199–1207.

5. Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, et al. Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT. Radiology;0(0):200823.

6. Chua F, Armstrong-James D, Desai SR, Barnett J, Kouranos V, Kon OM, et al. The role of CT in case ascertainment and management of COVID-19 pneumonia in the UK: insights from high-incidence regions. The Lancet Respiratory Medicine. 2020 Mar;0(0). EPub ahead of print. doi: 10.1016/S2213-2600(20)30132-6.

7. Santosh KC. AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data. Journal of Medical Systems. 2020;44:93.

8. Rajinikanth V, et al. Harmony-Search and Otsu based System for Coronavirus Disease (COVID-19) Detection using Lung CT Scan Images. CoRR. 2020;abs/2004.03431. Available from: https://arxiv.org/abs/2004.03431.

9. Liu KC, Xu P, Lv WF, Qiu XH, Yao JL, Gu JF, et al. CT manifestations of coronavirus disease-2019: A retrospective analysis of 73 cases by disease severity. European Journal of Radiology. 2020;126:108941.

10. Yang R, Li X, Liu H, Zhen Y, Zhang X, Xiong Q, et al. Chest CT Severity Score: An Imaging Tool for Assessing Severe COVID-19. Radiology: Cardiothoracic Imaging. 2020;2(2):e200047.

11. Fong SJ. Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak. International Journal of Interactive Multimedia and Artificial Intelligence. 2020;6(1):132–140.

12. Fong SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E. Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Applied Soft Computing. 2020;p. 106282.

13. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases. 2020 Mar;0(0). Publisher: Elsevier.

14. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. Radiology;0(0):200432.

15. Zhou Z, Guo D, Li C, Fang Z, Chen L, Yang R, et al. Coronavirus disease 2019: initial chest CT findings. European Radiology. 2020 Mar;EPub ahead of print. doi: 10.1007/s00330-020-06816-7.

16. Yoon SH, Lee KH, Kim JY, Lee YK, Ko H, Kim KH, et al. Chest Radiographic and CT Findings of the 2019 Novel Coronavirus Disease (COVID-19): Analysis of Nine Patients Treated in Korea. Korean Journal of Radiology. 2020;21(4):494–500.

17. Li K, Fang Y, Li W, Pan C, Qin P, Zhong Y, et al. CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). European Radiology. 2020 Mar;EPub ahead of print. doi: 10.1007/s00330-020-06817-6.

18. Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, et al. CT Imaging Features of 2019 Novel Coronavirus (2019-nCoV). Radiology. 2020 Feb;295(1):202–207.

19. CT outperforms lab diagnosis for coronavirus infection;. Last accessed date : 22-04-2020. Available from: https://bit.ly/3aoTQBD.

20. Borges do Nascimento IJ, Cacic N, Abdulazeem HM, von Groote TC, Jayarajah U, Weerasekara I, et al. Novel Coronavirus Infection (COVID-19) in Humans: A Scoping Review and Meta-Analysis. Journal of Clinical Medicine. 2020 Apr;9(4):941. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

21. Bernheim A, Mei X, Huang M, Yang Y, Fayad ZA, Zhang N, et al. Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection. Radiology. 2020 Feb;p. 200463.

22. Wang Y, Dong C, Hu Y, Li C, Ren Q, Zhang X, et al. Temporal Changes of CT Findings in 90 Patients with COVID-19 Pneumonia: A Longitudinal Study. Radiology. 2020 Mar;p. 200843.

23. Mahmud M, Kaiser MS, Hussain A, Vassanelli S. Applications of Deep Learning and Reinforcement Learning to Biological Data. IEEE Transactions on Neural Networks and Learning Systems. 2018 Jun;29(6):2063–2079.

24. Mahmud M, Kaiser MS, Hussain A. Deep Learning in Mining Biological Data. arXiv:200300108 [cs, q-bio, stat]. 2020 Feb;p. 1–36. ArXiv: 2003.00108. Available from: http://arxiv.org/abs/2003.00108.

25. Ali HM, Kaiser MS, Mahmud M. Application of Convolutional Neural Network in Segmenting Brain Regions from MRI Data. In: Liang P, Goel V, Shan C, editors. Brain Informatics. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2019. p. 136–146.

26. Orojo O, Tepper J, McGinnity TM, Mahmud M. A Multi-recurrent Network for Crude Oil Price Prediction. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI); 2019. p. 2940–2945.

27. Mahmud M, Kaiser MS, Rahman MM, Rahman MA, Shabut A, Al-Mamun S, et al. A Brain-Inspired Trust Management Model to Assure Security in a Cloud Based IoT Framework for Neuroscience Applications. Cognitive Computation. 2018 Oct;10(5):864–873.

28. Yahaya SW, Lotfi A, Mahmud M. A consensus novelty detection ensemble approach for anomaly detection in activities of daily living. Appl Soft Comput. 2019;83:105613.

29. Yahaya SW, Lotfi A, Mahmud M, Machado P, Kubota N. Gesture Recognition Intermediary Robot for Abnormality Detection in Human Activities. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI); 2019. p. 1415–1421.

30. Noor MBT, Zenia NZ, Kaiser MS, Mahmud M, Al Mamun S. Detecting Neurodegenerative Disease from MRI: A Brief Review on a Deep Learning Perspective. In: Liang P, Goel V, Shan C, editors. Brain Informatics. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2019. p. 115–125.

31. Miah Y, Prima CNE, Seema SJ, Mahmud M, Kaiser MS. Performance Comparison of Machine Learning Techniques in Identifying Dementia from Open Access Clinical Datasets. In: Proc. ICACIN 2020. Springer; 2020. p. 69–78.

32. Rabby G, Azad S, Mahmud M, Zamli KZ, Rahman MM. TeKET: a Tree-Based Unsupervised Keyphrase Extraction Technique. Cogn Comput. 2020;EPub ahead of print. doi: 10.1007/s12559-019-09706-3.

33. Silver D, et al. Mastering the game of Go with deep neural networks and tree search. nature. 2016;529(7587):484.

34. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, et al. Lung Infection Quantification of COVID-19 in CT Images with Deep Learning. CVPR. 2020;.

35. Bhandary A, Prabhu GA, Rajinikanth V, Thanaraj KP, Satapathy SC, Robbins DE, et al. Deep-learning framework to detect lung abnormality – A study with chest X-Ray and lung CT scan images. Pattern Recognition Letters. 2020;129:271–278.

36. Pugalenthi R, Rajakumar MP, Ramya J, Rajinikanth V. Evaluation and Classification of the Brain Tumor MRI using Machine Learning Technique. Journal of Control Engineering and Applied Informatics. 2019 Dec;21(4):12–21. Number: 4.

37. Celik Y, Talo M, Yildirim O, Karabatak M, Acharya UR. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. Pattern Recognition Letters. 2020;133:232 – 239.

38. Sharif M, Amin J, Raza M, Anjum MA, Afzal H, Shad SA. Brain tumor detection based on extreme learning. Neural Computing and Applications. 2020 Jan;EPub ahead of print. doi: 10.1007/s00521-019-04679-8.

39. Amin J, Sharif M, Raza M, Mussarat Y. Detection of Brain Tumor based on Features Fusion and Machine Learning. Journal of Ambient Intelligence and Humanized Computing. 2018;.

40. Amin J, Sharif M, Gul N, Yasmin M, Shad SA. Brain tumor classification based on DWT fusion of MRI sequences using convolutional neural network. Pattern Recognition Letters. 2020;129:115 – 122.

41. Sharif M, Amin J, Nisar MW, Anjum MA, Nazeer M, Shad SA. A unified patch based method for brain tumor detection using features fusion. Cognitive Systems Research. 2020;59:273–286.

42. Das A, Acharya RU, Panda SS, Sabut SK. Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques. Cognitive Systems Research. 2019 May;54:165–175.

43. Wu YH, Gao SH, Mei J, Xu J, Fan DP, Zhao CW, et al. JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. arXiv:200407054 [cs, eess]. 2020 Apr;p. 1–11. ArXiv: 2004.07054. Available from: http://arxiv.org/abs/2004.07054.

44. Hou B, Kang G, Zhang N, Liu K. Multi-target Interactive Neural Network for Automated Segmentation of the Hippocampus in Magnetic Resonance Imaging. Cognitive Computation. 2019 Oct;11(5):630–643.

45. Zhan J, Zhao H, Zheng P, Wu H, Wang L. Salient Superpixel Visual Tracking with Graph Model and Iterative Segmentation. Cognitive Computation. 2019 Jun;.

46. Xie J, Yu L, Zhu L, Chen X. Semantic Image Segmentation Method with Multiple Adjacency Trees and Multiscale Features. Cognitive Computation. 2017 Apr;9(2):168–179.

47. Artificial Intelligence AS. CT Dataset for COVID-19; 2020. Last access date 22-04-2020. Available from: http://medicalsegmentation.com/covid19/.

48. Dey N, Rajinikanth V, Shi F, Tavares JMRS, Moraru L, Karthik KA, et al. Social-Group-Optimization based tumor evaluation tool for clinical brain MRI of Flair/diffusion-weighted modality. Biocybernetics and Biomedical Engineering. 2019;39(3):843 – 856.

 doi:10.20944/preprints202005.0052.v1

49. Dey N, Rajinikanth V, Ashour AS, Tavares JMRS. Social Group Optimization Supported Segmentation and Evaluation of Skin Melanoma Images. Symmetry. 2018 Feb;10(2):51.

50. Kowsalya N, Kalyani A, Chalcedony CJ, Sivakumar R, Janani M, Rajinikanth V. An Approach to Extract Optic-Disc from Retinal Image Using K-Means Clustering. In: 2018 Fourth International Conference on Biosignals, Images and Instrumentation (ICBSII); 2018. p. 206–212.

51. Bose S, Mukherjee A, Madhulika, Chakraborty S, Samanta S, Dey N. Parallel image segmentation using multi-threading and k-means algorithm. In: 2013 IEEE International Conference on Computational Intelligence and Computing Research; 2013. p. 1–5.

52. Dey N, Ashour A. Classification and Clustering in Biomedical Signal Processing. IGI Global; 2016.

53. Tian Z, Dey N, Ashour AS, McCauley P, Shi F. Morphological segmenting and neighborhood pixel-based locality preserving projection on brain fMRI dataset for semantic feature extraction: an affective computing study. Neural Computing and Applications. 2018 Dec;30(12):3733–3748.

54. Wang Y, Shi F, Cao L, Dey N, Wu Q, Ashour AS, et al. Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. Current Bioinformatics. 2019;14(4):282–294.

55. Chaki J, Dey N. Texture Feature Extraction Techniques for Image Recognition, Voice In Settings. SpringerBriefs in Computational Intelligence. Springer Singapore; 2020.

56. Acharya UR, Fernandes SL, WeiKoh JE, Ciaccio EJ, Fabell MKM, Tanik UJ, et al. Automated Detection of Alzheimer's Disease Using Brain MRI Images– A Study with Various Feature Extraction Techniques. Journal of Medical Systems. 2019 Aug;43(9):302.

57. Maheshwari S, Pachori RB, Acharya UR. Automated Diagnosis of Glaucoma Using Empirical Wavelet Transform and Correntropy Features Extracted From Fundus Images. IEEE Journal of Biomedical and Health Informatics. 2017;21(3):803–813.

58. Kala S, Ezhilarasi M. Comparative Analysis of Serial and Parallel Fusion on Texture Features for Improved Breast Cancer Diagnosis. Current Medical Imaging Reviews. 2018;14(6):957–968.

59. Moore CM, Bell DJ, et al. COVID-19 | Radiology Reference Article | Radiopaedia.org; 2020. Last access date: 22-04-2020. Available from: https://radiopaedia.org/articles/covid-19-3.

60. Bahman R. Radiopaedia, editor. Cases by R. Bahman: Radiopaedia.org rID: 74560. Radiopaedia; 2020. Last access date: 22-04-2020. Available from: https://radiopaedia.org/cases/covid-19-pneumonia-3.

61. Hosseinabadi F. Radiopaedia, editor. Case courtesy of Dr Fateme Hosseinabadi: Radiopaedia.org rID: 74868. Radiopaedia; 2020. Last access date: 22-04-2020. Available from: https://radiopaedia.org/cases/covid-19-pneumonia-8.

 doi:10.20944/preprints202005.0052.v1

62. Smith D. Radiopaedia, editor. Case courtesy of Dr Derek Smith: Radiopaedia.org rID: 75249. Radiopaedia; 2020. Last access date: 22-04-2020. Available from: https://radiopaedia.org/cases/covid-19-pneumonia-23.

63. Bahman R. Radiopaedia, editor. Cases by R. Bahman: Radiopaedia.org rID: 74879. Radiopaedia; 2020. Last access date: 22-04-2020. Available from: https://radiopaedia.org/cases/covid-19-pneumonia-10.

64. Cetinoglu K. Radiopaedia, editor. Case courtesy of Dr Kenan Cetinoglu: Radiopaedia.org rID: 75281. Radiopaedia; 2020. Last access date: 22-04-2020. Available from: https://radiopaedia.org/cases/covid-19-pneumonia-27.

65. Feger J. Radiopaedia, editor. Case courtesy of Dr Joachim Feger: Radiopaedia.org rID: 75541. Radiopaedia; 2020. Last access date: 22-04-2020. Available from: https://radiopaedia.org/cases/covid-19-pneumonia-52.

66. TaghiNiknejad M. Radiopaedia, editor. Case 55, courtesy of Dr Mohammad TaghiNiknejad: Radiopaedia.org rID: 75606. Radiopaedia; 2018. Last access date: 22-04-2020. Available from: https://radiopaedia.org/cases/covid-19-pneumonia-55.

67. TaghiNiknejad M. Radiopaedia, editor. Case courtesy of Dr Mohammad TaghiNiknejad: Radiopaedia.org rID: 75607. Radiopaedia; 2020. Last access date: 22-04-2020. Available from: https://radiopaedia.org/cases/covid-19-pneumonia-56.

68. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of Digital Imaging. 2013 Dec;26(6):1045–1057.

69. Armato SG, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans: The LIDC/IDRI thoracic CT database of lung nodules. Medical Physics. 2011 Jan;38(2):915–931.

70. Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. Radiology. 2009 Jul;252(1):263–272.

71. Zhao B, Schwartz LH, Kris MG. Data From RIDER_Lung CT. The Cancer Imaging Archive; 2015. Available from: https://wiki.cancerimagingarchive.net/x/XIRXAQ.