



Computer-aided diagnosis of COVID-19 from chest X-ray images using histogram-oriented gradient features and Random Forest classifier

Malathy Jawahar¹ · J. Prassanna² · Vinayakumar Ravi³ · L. Jani Anbarasi² · S. Graceline Jasmine² · R. Manikandan⁴ · Ramesh Sekaran⁵ · Suthendran Kannan⁶

Received: 5 September 2021 / Revised: 30 January 2022 / Accepted: 28 April 2022 /
Published online: 10 May 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

The decision-making process is very crucial in healthcare, which includes quick diagnostic methods to monitor and prevent the COVID-19 pandemic disease from spreading. Computed tomography (CT) is a diagnostic tool used by radiologists to treat COVID patients. COVID x-ray images have inherent texture variations and similarity to other diseases like pneumonia. Manually diagnosing COVID X-ray images is a tedious and challenging process. Extracting the discriminant features and fine-tuning the classifiers using low-resolution images with a limited COVID x-ray dataset is a major challenge in computer aided diagnosis. The present work addresses this issue by proposing and implementing Histogram Oriented Gradient (HOG) features trained with an optimized Random Forest (RF) classifier. The proposed HOG feature extraction method is evaluated with Gray-Level Co-Occurrence Matrix (GLCM) and Hu moments. Results confirm that HOG is found to reflect the local description of edges effectively and provide excellent structural features to discriminate COVID and non-COVID when compared to the other feature extraction techniques. The performance of the RF is compared with other classifiers such as Linear Regression (LR), Linear Discriminant Analysis (LDA), K-nearest neighbor (kNN), Classification and Regression Trees (CART), Random Forest (RF), Support Vector Machine (SVM), and Multi-layer perceptron neural network (MLP). Experimental results show that the highest classification accuracy (99.73%) is achieved using HOG trained by using the Random Forest (RF) classifier. The proposed work has provided promising results to assist radiologists/physicians in automatic COVID diagnosis using X-ray images.

Keywords COVID-19 · Classification · Random Forest · HOG · Features extraction

✉ Vinayakumar Ravi
vravi@pmu.edu.sa

1 Introduction

Coronavirus disease 2019 (COVID-19) is a deadly infectious disease that was reported first in Wuhan, China, in December 2019 [24, 28, 47]. This COVID-19 virus has created a significant outbreak because no cure was discovered during the early days. The biological structure has a stable, one-stranded RNA, and it is challenging to treat this disease due to its mutating features. COVID-19 is currently the source of death of thousands worldwide and has primarily affected various countries such as the United States, Spain, India, Italy, China, the United Kingdom, Iran, etc. Several varieties of COVID-19 are found in humans, cats, dogs, pigs, poultry and rodents. COVID-19 symptoms include sore throat, fever, headache, headache, nose flushing, and cough. The virus can infect people with weak immune systems and cause them to die. This COVID-19 infectious disease spreads from one human to another and spreads rapidly across the globe. This can mainly spread through physical touch, breath contact, hand contact, or mucus contact from one person to another. This virus is part of a family that causes acute breathing symptoms. The structure of the virus includes spikes on the crown on the outer surface. Severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) also fall into the same category [36]. Such conditions also cause acute lung injury and acute breathing distress [10]. Twelve million nine hundred sixty-four thousand eight hundred nine people were infected worldwide, resulting in 570,288 deaths as of July 15 2020. The current scenario reveals that COVID-19's mortality rate seems to be a maximum threat for people with chronic health problems and for the elderly. The virus is spread by coughing, sneezing, and respiratory droplets [44] from person to person. This virus's typical signs include fever, inflammation, respiratory abnormality, conditions that can cause catastrophic consequences, including pneumonia, multiple organ failure, and death [36, 41]. Laboratory tests are a time-consuming testing tool that has a high cost and requires a well-developed research laboratory. Computer tomography (CT) helps provide treatment quickly. The main challenges include:

- Analyzing major discriminant features than the current research works to attain high accuracy.
- Lesser image data with varying image *sizes pose* challenges due to the low-resolution images.
- Computation time should be minimal.

COVID-19's superior outcomes on chest CTs are multifocal, patchy, consolidated, ground-glass opacity, and a peripheral distribution of a 'crazy-paved' pattern. Ground-glass opacities (GGO) are common among many diseases, such as measles, severe SARS, and MERS [8, 48]. Further investigation is required to improve the radiologist's performance. This study's primary motivation is to detect COVID-19 using handcrafted features and to perform an extensive experiment to suggest the best quality training with the best classifier.

There are main contributions of the proposed work computer-aided detection of COVID-19 using CT lung x-ray images:

- A robust, handcrafted feature extraction technique that is presented for the classification of COVID-19 detection using the Lung X-Ray dataset.
- Three different feature extraction techniques are explored, namely Haralick features using GLCM matrix, HOG features and Invariant Hu Moments.

- This model generated *a* handcrafted discriminative feature set compared to *another* study.
- Seven different classifiers are employed to classify the input image into COVID and Non-COVID.
- Comparative analysis of the multifeature and multiclassifier models is performed in terms of accuracy and receiver operating characteristic (ROC) analysis.
- The proposed work attained 99.7% accuracy in diagnosing the X-ray COVID-19 images.

The experimental basis of this research work is organized in the following sequence: Section 2 highlights the related work in this area of research. Section 3 elucidates the methodology of the proposed work. Section 4 discusses the performance *comparison* of the *multi feature* and multiclassifier models for classifying the COVID-19 Lung X-ray images with the *current results*. Conclusion and future works *summary* is given in Section 5.

2 Related work

Many researchers have conducted investigations to classify COVID-19 disease in different ways. Few have attempted to recognize the COVID-19 virus [39], the effects of the disease in humans [25], its treatment [47], possible conditions during treatment and the discharge [11], the alternative mesenchymal stem cell therapy treatment [7], a kit to protect the doctors against the COVID-19 virus [15], and possible effects on children [46]. Studies presume that most of the symptoms and infections occur in adult patients [23].

Togacar et al. [26] propose a classification for pneumonia disease using chest X-ray images. Deep Learning techniques like AlexNet, VGG-16, and VGG-19 neural network models were used that achieved an accuracy of 99.41%. Sousa et al. [42] categorized the pneumonia diagnosis in infants by CAD from radiographic images using SVM, KNN, and Naïve Bayes, where SVM outperformed the other. The dataset compiled by Kermanyet et al. [22] was analyzed to test the efficiency of child pneumonia classification using the Neural Convolution Network method and achieved 90.5% accuracy.

Turker et al. [44] proposed the use of Decision tree, Linear Discriminant (LD), SVM, subspace discriminant (SD) system to detect COVID-19, which would involve preprocessing and feature extraction, selection and classification, and holdout validation. The dataset used in this paper included healthy and COVID-19 X-ray images that are a public dataset collected from the Github (<https://github.com/UCSD-AI4H/COVID-CT>) website. The dataset included 87 X-ray images with COVID-19 disease, including 26 females, 41 males, and 20 undetermined. Similarly took Healthy data from the Kaggle (<https://www.kaggle.com/luisblanche/COVIDCT>) site.

COVID-19 and other typical and viral pneumonia were analyzed using 1020 CT slices from 108 laboratory-proven patients who were suggested by Ali Abbasian Ardakani et al. [2] based on a deep learning process. AlexNet, VGG-16, VGG-19, Squeeze Net, GoogleNet, MobileNet-V2, ResNet-18, ResNet-50, ResNet-101, and Xception were used to analyze the COVID-19 infection. ResNet-101 and Xception achieved the highest accuracy of all networks. ResNet-101 is used as the high sensitivity model for characterizing and diagnosing COVID-19 infections. Figure 1 shows sample COVID-19 X-ray data. The training accuracy attained by AlexNet, VGG-16, VGG-19, GoogleNet were 82%, 83.70%, 87.13%, 84.80, respectively.

Flu-like symptoms are the initial symptoms of the COVID-19 coronavirus; the dataset included data after confirming the diagnosis through swab samples. Patients with chronic lung

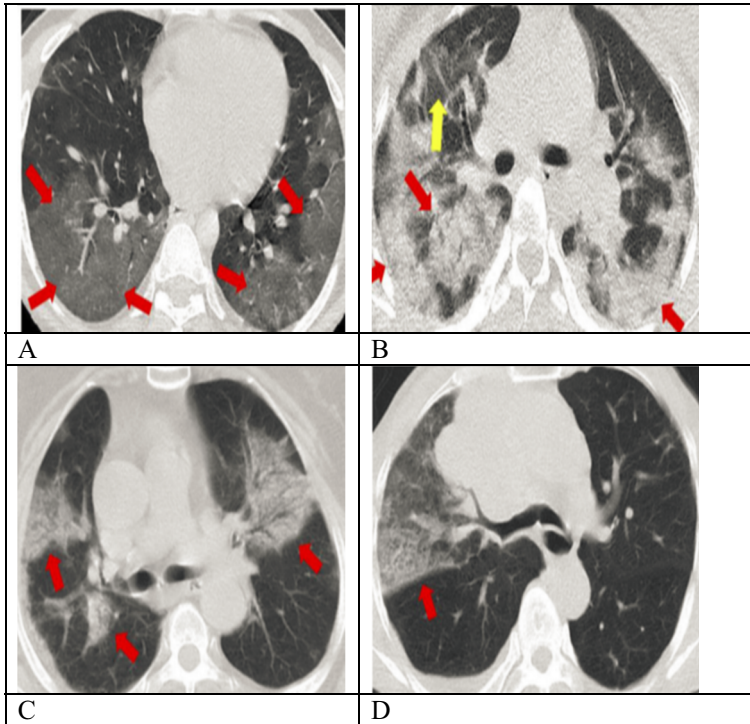


Fig. 1 CT sample images of patients with pneumonia [2]

diseases were excluded with negative real-time polymerase chain reaction (RT-PCR). The High-resolution computed tomography (**HRCT**) images taken from September 2019 to December 2019 in the university hospitals analyzed images that measured 512×512 pixels. A high spatial algorithm is used to reconstruct images in the transverse plane. The images would then be transformed into a grey-scale image and analyzed by the radiologist. The regions are then captured and resized to 60×60 pixels for further study.

Using deep learning techniques, this paper [30] classified the COVID-19 chest image, pneumonia, and normal chest datasets. Each print is preprocessed before training using deep learning models. The dataset is reconstructed with the Fuzzy technique and the Stacking technique during preprocessing. These preprocessed images are trained using SqueezeNet, MobileNetV2, and deep learning models and are later classified using an SVM classifier.

These authors used three-class X-ray images namely regular, pneumonia, and COVID-19 from the publicly accessible datasets. A COVID-19 dataset, shared with 76 [9] pictures labeled with COVID-19, was posted on the GitHub website by a researcher called Joseph Paul Cohen of the University of Montreal. The second dataset included 219 X-ray images available on the Kaggle website [34]. These authors combined these two datasets to create a new dataset that contains 219 X-Ray images. Pneumonia datasets are also taken from the public group [16], including 53 patients. This study included 295 images in the COVID-19 class, 65 in the Normal category, and 98 in pneumonia [1] class for a total of 458 chest X-ray images. 70% of the datasets are used as for training, and 30% are used for testing.

Tanvir et al. [29] proposed the CovXNets deep learning-aided scheme to detect the COVID-19 and other pneumonia using chest X-ray images. Trained different variations of

CovXNets with various X-ray images. Implemented a discriminative, gradient-based localization was later to classify the abnormal regions of the radiation images. Detailed research achieved COVID / Normal 97.4%, COVID / Viral Pneumonia 96.9%, COVID / Bacterial Pneumonia 94.7% and COVID / Viral / Bacterial Pneumonia 90.2%. Minimal preprocessing min-max normalization and resizing to uniform shapes are involved in the testing. The datasets used in this study included 5856 images, collected 2780 bacterial pneumonia, 1583 normal X-rays, 1493 non-COVID viral pneumonia X-rays from Guangzhou Medical Center, China [22]. The other dataset included 305 X-rays of COVID-19 collected from Sylhet Medical College, Bangladesh. The final dataset was created by combining all the COVID-19 X-rays. Five-Fold Cross-validation schemes are used to evaluate the proposed process in training and testing phases.

This scheme used the open-source dataset compiled by Cohen et al. [9], which included 192 X-ray images of COVID-19 patients; a total of 337 images are in the dataset. The images are obtained using websites like Radiopedia.org and Figure1.com. This set included the Posterior Anterior (PA) and Anterior-Posterior Supine (AP Supine) views of the pulmonary images [12]. This study analyzed only the PA views for the training and testing process. The PA view has 147 X-ray images of different ages. Five thousand eight hundred sixty-three healthy chest X-ray images taken from the Kaggle website. This study analyzed four classes: COVID-19, Normal, Bacterial Pneumonia, and Viral Pneumonia, including One hundred forty-seven images of normal X-rays from the dataset. The data is split in the ratio of 70% data for training and 30% for testing purposes. The dataset was resized to 224×224 pixels; RGB re-ordering is applied, resulting in $224 \times 224 \times 3$ image format. Image augmentation was performed with a rotation range of 20, horizontal and vertical flip rotation range for the small dataset. Data leakage is a crucial when the dataset includes the images of the same patients taken on different days. Splitting is performed manually to avoid this at the individual patient level at 70% and 30% for the training and testing processes. The proposed model involved deep learning based on Convolution Neural Network, known as nCOVnet, which includes 24 layers where the first layer is the input layer, and the other 18 layers are the combination of Convolution + ReLU and Max Pooling layers. The VGG16 [40] model was used in this scheme, which was trained in ImageNet data collection. A threshold drop-out of 0.5 is used with a fully dense layer with an activation feature and a size of 64 units. The training accuracy is up to 97%, and the COVID-19 patient are correctly classified with 97% accuracy during testing. Various works have been performed in recent years to classify COVID-19 using Lung X-ray images. It is a challenging task due to its inherent texture variations and similarity to other diseases like pneumonia. Several studies have developed classifications for COVID-19 based on computer vision algorithms. Ullah et al. [45] provided an overview of the scalable telemedicine services commonly used for COVID-19 and other patients. Mobile networks, social networks, and software-based services are the three types of telehealth systems now available. The *designs* are detailed in terms of their operating principle, technology, utilized equipment, and the types and methodologies of service given throughout the paper. Hou et al. [17] introduced a new diagnosis platform based on a deep convolutional neural network (DCNN) that can assist radiologists in *diagnosing* COVID-19 pneumonia on chest X-ray. The DCNN additionally uses the explainable approach to choose instances to explain the behavior of training-learning models to improve prediction accuracy. Md. Milon Islam et al. [19] discussed the various existing portable monitoring devices and respiratory safety systems frequently used to help coronavirus-infected people. To identify the appropriate technology for

COVID-19 infected individuals, a comparison discussion with possible future trends is also conducted.

3 Proposed work

3.1 Overall structure

Figure 2 shows the overall flow structure of the proposed COVID-19 classification system. A machine learning hand-crafted feature extraction technique using Histogram Oriented Gradient with Random Forest is proposed for the classification of X-ray image into COVID and Non-COVID.

The proposed model is developed using the system configuration as given in Table 1. Python 3.6 is used to structure the original data set using the feature extraction and classification methods. Jupyter Notebook is the interfaces program used in compiling Python.

The X-ray image is converted into grayscale. Then, the image is preprocessed using the Wiener filter. Lim and Oppenheim defined the wiener filter [3] based on the reduction of the mean square error between the estimated signal $W'(\omega)$ and the original signal $W(\omega)$. as

$$S(\omega) = \frac{W_x(\omega)}{W_x(\omega) + W_y(\omega)} \quad (1)$$

Where $W_x(\omega)$ and $W_y(\omega)$ represents the noise-free and the noisy-background signal that is stationary and uncorrelated. Compute the preprocessed COVID-19 image after assigning the transfer function as

$$W'(\omega) = X(\omega)S(\omega) \quad (2)$$

Two feature extraction methods (GLCM, HOG) are deployed and fused with Hu moments.

3.2 Grey-level co-occurrence matrix

A second-order statistical texture function is derived from the COVID-19 images based on neighbor pixels' displacement d and pixel orientation θ . The likelihood of a pair of grey levels occurring at the distance d with direction in the image is described as function $f(i,j|d, \theta)$. Using the GLCM matrix, Haralick et al. [13] analyzed 14 parameters namely Contrast, Autocorrelation, Cluster prominence, Correlation, Dissimilarity, Energy (uniformity), Cluster shade,

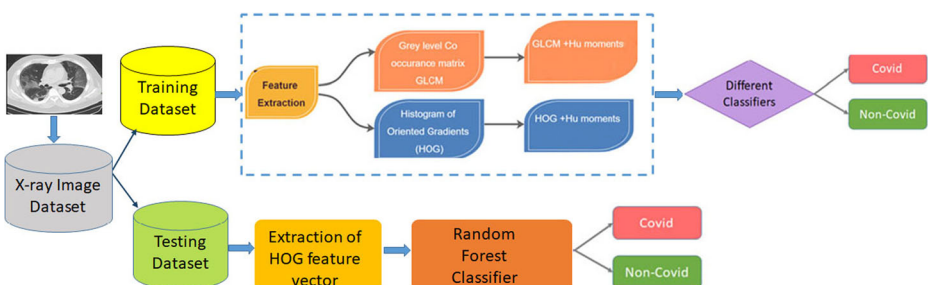


Fig. 2 Proposed COVID-19 Classification model using Chest X-ray images

Table 1 System configuration of the proposed work

Item	Configuration
Processor	Intel(R) Xeon(R) Gold 6230 CPU @ 2.10 GHz
Graphics Card	NVIDIA QUADRO RTX 5000 16 GB
RAM size	64.0 GB
Hard disk size	2 TB

Entropy, Maximum probability, Homogeneity, Information measure of correlation Sum of square (variance), Difference variance, Inverse difference normalized (INN) and Inverse difference moment normalized [31, 35, 43].

3.3 Hu moments

Hu [18] introduces the moment invariants in which six absolute orthogonal invariants were derived from COVID-19 images irrespective of position, size, and orientation and irrespective of parallel projections. The moment invariants provide adequate measures to track COVID-19 image patterns for image position, scaling and image rotation, given that they are noise free and continuous. Hu moment invariants are extensively used in COVID-19 image registration, reconstruction and pattern recognition. The two dimensional $(p + q)$ th order moment is given as follows in the Eq. (3)

$$Hm_{x,y} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P^x q^y f(p,q) dpdq \text{ where } x,y = 0, 1, 2 \dots \quad (3)$$

The COVID-19 image function is defined as a blockwise continuous bounded function $f(p,q)$ and the Hu moment $Hm_{x,y}$ is uniquely identified by the function $f(p,q)$; the $f(p,q)$ can also be uniquely identified by the moment sequence $Hm_{x,y}$. The moment features will not be when they are translated, rotated, or scaled. Central moments must be computed to obtain the invariant features and are defined as follows in Eq. 4

$$C\mu_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (p - p')^x (q - q')^y f(p,q) dpdq \text{ where } x,y = 0, 1, 2 \dots \quad (4)$$

The pixel points (p', q') are the centroid of the image $f(p,q)$. The normalized central moment's Eq. (5)

$$n_{x,y} = \frac{C\mu_{xy}}{C\mu_{00}^{\frac{x+y}{2}}} \quad (5)$$

3.4 Histogram of oriented gradients (HOG) features

HOG is a feature descriptor mostly used to extract features from image data. HOG deals with the structure of an object's shape. HOG identifies the edge direction by extracting the removing and the gradient and breaking the image into smaller regions. The HOG will eventually generate a histogram separately with pixel values for each of these regions.

The total gradients magnitude is computed using the Eq. (6) and the orientation is computed as given in Eqs. (7, 8)

$$M = \sqrt{(Q_x)^2 + (Q_y)^2} \quad (6)$$

$$\tan(\phi) = \frac{Q_x}{Q_y} \quad (7)$$

$$\phi = \text{atan}\left(\frac{Q_x}{Q_y}\right) \quad (8)$$

Accommodating changes in contrast and illumination, which basically require grouping cells in larger, spatially connected blocks, must be normalized locally. The HOG descriptor computes the vector of the normalized cell histograms components from each block. Each cell's final descriptor is assisted more than once through the block overlap. Figure 3 shows the input and gradient images of the COVID-19 X-Ray obtained using HOG. Figure 4 shows the Block Normalization Scheme of HOG Algorithm.

3.5 Machine learning classifiers

All four different feature vectors were given as input to the classifiers, namely Linear regression (LR) [37], Linear Discriminant Analysis (LDA) [4, 27], K-Nearest Neighbour kNN [2, 23, 26, 42, 48]; (<https://www.kaggle.com/luisblanche/COVIDCT>) Classification and Regression Trees (CART) [20, 33], Support Vector Machine (SVM) [5, 38], Multi-layer perceptron Neural Network (NN) [14, 21] and Random Forest (RF).

3.5.1 Random Forest

Breiman proposed the RF classifier based on multiple decision trees. RF classifier is robust to noise and outliers. RF gives various useful estimates such as variable importance, Out-of-Bag

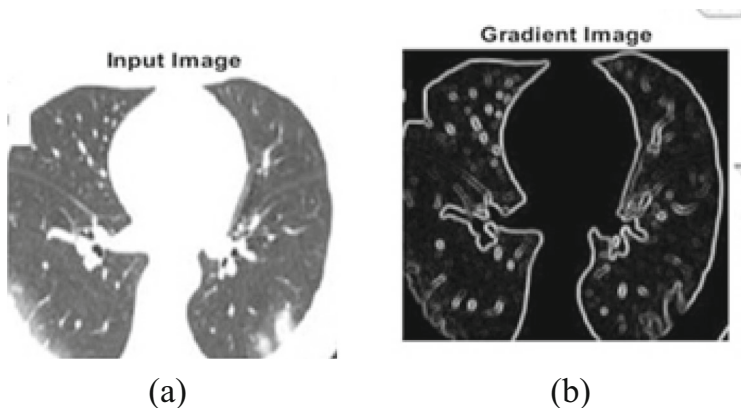


Fig. 3 The input image and the gradient image of the COVID-19 X-Ray

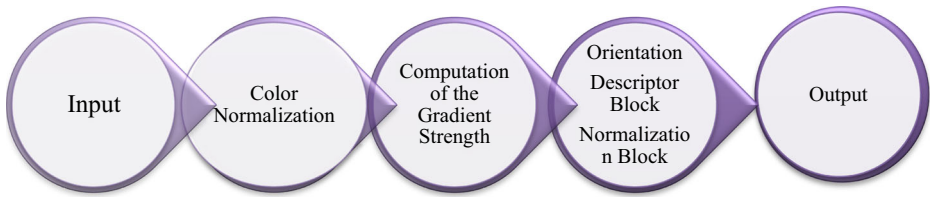


Fig. 4 Block normalization scheme of HOG algorithm

error estimate (unbiased estimate of the classification error), strength, correlation etc., and is faster than other bagging and boosting algorithms. Every tree can be considered a single classifier and computed as a unit to identify the input function for final classification. RF changes the way each sub-tree learns by randomly creating a different subsets of features such that the sub-trees have less correlation. The number of features that can be searched at each split and the number of the trees to be constructed are given as parameters to the algorithm. For example, if the number of images (L) in the dataset (T) with the ‘ F ’ number of features, the size of features selected at each split is typically equal to \sqrt{F} . RF divides each node randomly employing selected features. An Error has been predicted for every tree’s out-of-bag portion; each variable’s feature variable permutation was also computed. The same computation was carried out after permuting each feature variable. Splitting stops when the feature variable standard deviation difference equals 0 [6, 37]. The node impurity measure is computed using the Gini index in RF. Gini(T) is defined as Eq. 9:

$$Gini(T) = 1 - \sum_{j=1}^n P_j^2 \quad (9)$$

Where P_j represents dataset T relative frequency with n classes. Random forest hyperparameters include the number of decision trees in the forest ($n_estimators = [150, 250, 350, 500]$), the number of features considered by each tree when splitting a node ($max_features = ['auto', 'sqrt']$), max number of levels in each decision tree ($max_depth = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$), min number of data points placed in a node before the node is split ($min_samples_split = [2, 4, 8]$), min number of data points allowed in a leaf node ($min_samples_leaf = [2, 4, 6]$), method for sampling data points ($bootstrap = [true (with) or false (without replacement)]$) are tuned using GridSearchCV from Sklearn python package. In our study there are $(4 * 2 * 10 * 3 * 3 * 2) * cv = 1440$ training.

The working of the Random Forest algorithm is explained in the flow diagram given in Fig. 5.

4 Experimental results and discussion

Kaggle included COVID-19 CT Lung X-ray images where 349 CT images with 153, 491, and 1853 are the minimum, average and maximum width. These images belong to 216 patients; This includes 397 images from 55 people, whereas LUNA includes 36 images, MedPix has 195 images, 202 are from PMC, and Radiopaedia has 30 of COVID-19 negative images. The width of 124, 383 and 1485 is the minimum, average and maximum. A 10-fold cross-validation technique is executed to achieve consistent evaluation for classification accuracy on each classifier. The classifier’s performance is assessed based on the average classification accuracy of the 10 folds.

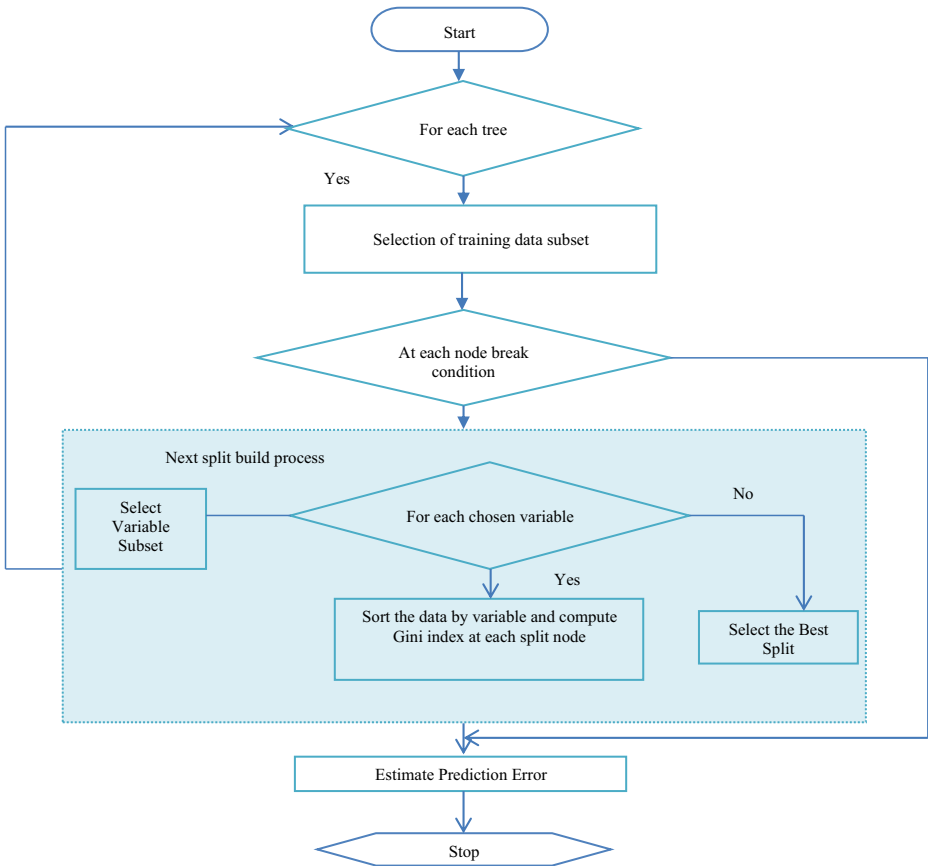


Fig. 5 Process flow of Random Forest algorithm

This research work uses three feature extraction methods namely HOG, GLCM and Hu moments to classify COVID-19 X-ray images. The dataset was partitioned into 80% training and 20% testing subsets. Images were initially preprocessed using Wiener the filter and from the preprocessed data HOG, GLCM and Hu Moments features are extracted. Extracted HOG, GLCM and Hu Moments features are given as input to different classifiers, namely LR, LDA, KNN, CART, RF, SVM and MLP. Hyperparameters are significant because they govern a machine learning model's overall behaviour. The final goal is to discover the best hyperparameter combination that minimises a preset loss function and produces better results. A hyperparameter is a value for a parameter that influence the learning process. Other factors, such as node weights, Regularization constant, kernel type, and constants, are fine tuned in SVMs and k in K-NN. In a neural network, the number of layers, units per layer, and regularisation are all computed to reduce the errors in training; the Bias and Variance are also tuned for better accuracy. Many research studies showed that a larger decision tree forest has high prediction accuracy. Since determining the right hyper-parameter can be time-consuming, in this study, authors used GridSearchCV from *the* sklearn package to find the best model parameters. The GridSearchCV returned best RF model parameters as bootstrap method for sampling data points with `max_feature` as 'auto', `max_depth` as '70', `min_samples_leaf` as '4',

Table 2 The Accuracy obtained for Haralick features (GLCM), GLCM and Hu features, and GLCM and HOG features

Haralick (GLCM) (13 features) – Exp. 1	GLCM+Hu (20 features) - Exp. 2	GLCM+HOG – Exp. 3
LR: 0. 676 (0. 055)	LR: 0. 949 (0. 054)	LR: 0. 995 (0. 006)
LDA: 0. 659 (0. 066)	LDA: 0. 959 (0. 029)	LDA: 0. 995 (0. 006)
KNN: 0. 603 (0. 061)	KNN: 0. 991 (0. 013)	KNN: 0. 591 (0. 003)
CART: 0. 719 (0. 066)	CART: 0. 991 (0. 013)	CART: 0. 991 (0. 008)
RF: 0. 768 (0. 035)	RF: 0. 997 (0. 008)	RF: 0. 997 (0. 005)
SVM: 0. 572 (0. 060)	SVM: 0. 822 (0. 182)	SVM: 0. 567 (0. 052)
MLP: 0. 519 (0. 049)	MLP: 0. 569 (0. 428)	MLP: 0. 618 (0. 153)

min_samples_split as ‘4’ and n_estimators as ‘250’. The bootstrapping method *assists* RF to overcome model overfitting issues.

The RF classifier trained with the GLCM and the HOG feature set efficiently discriminates between the COVID and NON-COVID X-ray images. The RF classifier’s performance evaluated using GLCM, HOG, and fusing Hu moments showed higher classification accuracy than the other classifiers. The RF classifier achieved better results by employing a majority voting strategy, bagging on data samples, and randomly creating different subsets of features. Tables 2 and 3 show the mean and standard deviation of the classification accuracy obtained for seven classifiers trained with Haralick and HOG features using the 10-fold cross-validation method. From Table 2, it can be observed that the first experiment (Exp. 1) with Haralick features did not show any progressive improvement for all the seven classifiers; the best accuracy (76.8%) is obtained using RF. However, Haralick features combined with Hu moments (Exp. 2) give a classification accuracy above 90% for LR, LDA, KNN, RF, whereas SVM attained 82.2%; a low accuracy of 56% is noted for MLP. The HOG features fused with Haralick in the third experiment (Exp. 3) gave the best accuracy of above 99% for LR, LDA, CART, RF. In comparison, MLP and SVM illustrated with lower accuracy of 61.8% and 56.7%, respectively. Figure 6 shows the classification accuracy using (a) Multi-Class Classifier for GLCM + Hu, (b) Multi-Class Classifier for GLCM, (c) Multi-Class Classifier for GLCM + HOG, (d) Multi-Class Classifier for GLCM + HOG + HU, (e) Multi-Class Classifier for Hu Moments, (f) Multi-Class Classifier for HOG, and (g) Multi-Class Classifier for HOG + Hu.

The results obtained with various combinations of HOG features, illustrated in Table 3, confirm the HOG features’ excellent performance with all the seven classifiers.

Figure 7a illustrates the distribution of the classification accuracy data for the seven classifiers trained using the HOG feature extraction methods. Figure 7b shows the classification accuracy obtained for the Random Forest classifier trained with all the feature extraction

Table 3 The accuracy obtained for HOG features, HOG + Hu features and Haralick + HOG + Hu features

HOG (3872 features) – Exp. 4	HOG+Hu (3879 features) - Exp. 5	GLCM+HOG+Hu features - Exp. 6
LR: 0. 997 (0. 005)	LR: 0. 994 (0. 006)	LR: 0. 995 (0. 006)
LDA: 0. 997 (0. 005)	LDA: 0. 995(0. 006)	LDA: 0. 995 (0. 006)
KNN: 0. 997 (0. 005)	KNN: 0. 993 (0. 006)	KNN: 0. 591 (0. 003)
CART: 0. 993 (0. 006)	CART: 0. 995 (0. 006)	CART: 0. 9973 (0. 0053)
RF: 0. 997 (0. 005)	RF: 0. 997 (0. 005)	RF: 0. 9973 (0. 0053)
SVM: 0. 997 (0. 005)	SVM: 0. 995 (0. 006)	SVM: 0. 567 (0. 052)
MLP: 0. 997(0. 005)	MLP: 0. 994 (0. 006)	MLP: 0. 686 (0. 224)

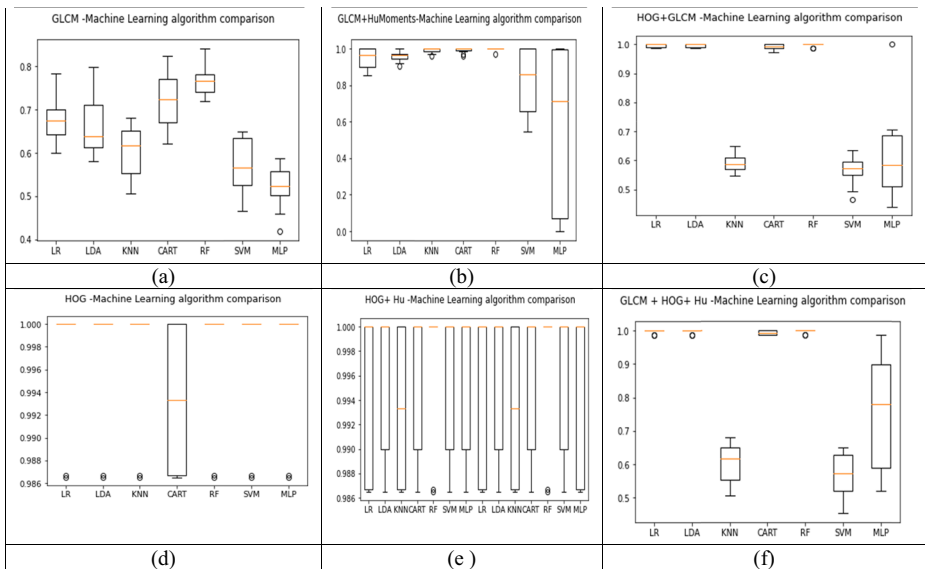
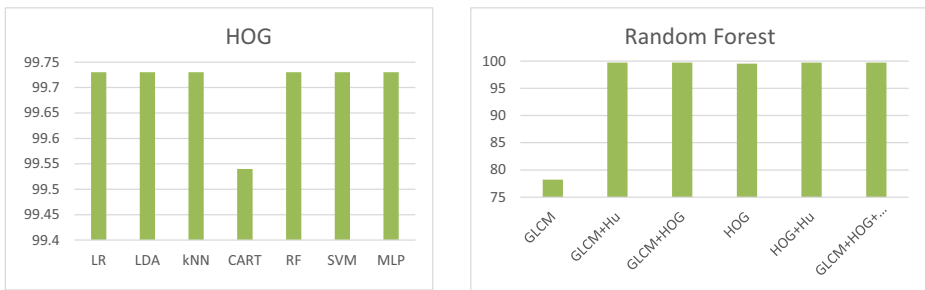


Fig. 6 Classification Accuracy using (a) Multi-Class Classifier for GLCM (b) Multi-Class Classifier for GLCM + Hu Moments (c) Multi-Class Classifier for HOG + GLCM (d) Multi-Class Classifier for HOG (e) Multi-Class Classifier for HOG + Hu Moments (f) Multi-Class Classifier for GLCM + HOG + Hu

techniques. We can observe that Random Forest achieved a superior classification accuracy (99.73%) with the HOG feature vector. Experimental results show that we can successfully use the multiple features set of GLCM with Hu texture features trained using Random Forest classifier for COVID-19 classification.

The ROC plot (Fig. 8) also confirms the potential of using the HOG feature technique coupled with Random Forest to classify COVID-19 X-ray images. Ali Abbasian Ardakani et al. [2] reported a similar result of 99% accuracy using ResNet101. Deep learning (DL) models are less interpretable and massive huge data and high computational resources. DL models may result in an overfit when trained with a smaller dataset. Most medical domain problems have a smaller and imbalanced training dataset. The efficiency of the proposed work is evaluated by comparing it with various existing research studies reported in current literature and is given in Table 4. It can be seen from the above results (Tables 3 and 4) that the proposed



a. The Accuracy achieved by different classifiers trained using HOG features

b. The Accuracy achieved by different feature extraction methods using Random Forest classifier

Fig. 7 a The Accuracy achieved by different classifiers trained using HOG features. b The Accuracy achieved by different feature extraction methods using Random Forest classifier

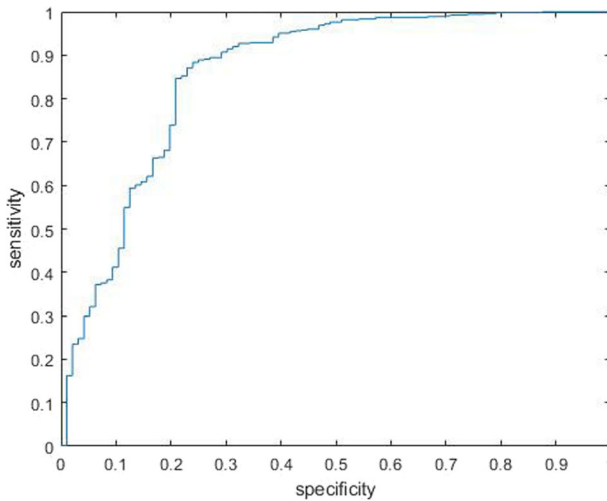


Fig. 8 ROC plot for RF trained with HOG features

machine learning model based on Random Forest trained using HOG features achieved promising results on the classification of the COVID-19 dataset. HOG is found to reflect the local description of edges effectively and, hence, provide an excellent localization property when compared to the other feature extraction techniques. An experimental analysis of the HOG method proved to provide a better generalization ability in discriminating between the COVID and NON-COVID images and yielded a higher classification accuracy when trained with all the classifiers.

The proposed optimized Random Forests trained with HOG feature system can be very helpful to the radiologist and physicians for treating the COVID patients.

5 Conclusions

Classifying the COVID dataset is an essential issue in the current scenario. One of the vital aspects is identifying the significant features that discriminate between the COVID and non-COVID images. A machine learning approach based on CT images with promising potential is presented to distinguish COVID-19 infections from normal chest lung x-ray images. This study used a dataset comprising 349 COVID-19 and 397 non-COVID-19 images. Both an individual and a hybrid method that combines multi-feature vector HOG(3872 features), GLCM (13 features), and Hu moments (7 features) are investigated. Experimental results show that the classification results also increase as the dimension of the features increases. HOG achieved the highest classification accuracy of 99.7% when trained with different classifiers, namely LR, LDA, KNN, CART, RF, SVM, and MLP. The Random Forest classifier, an ensemble classifier that combines multiple decision trees as the base classifier, is similarly found to provide more accurate predictions than other different classifiers. The proposed study demonstrated an efficient feature extraction technique and intelligent classifier that automatically provide faster learning to classify chest X-ray images. The study showed HOG to extract unique features and RF to provide better generalization performance for classifying the COVID-19 lung X-ray images. The performance of the proposed system is

Table 4 Comparison of the proposed accuracy with the existing algorithm

Ref No	COVID-19 Images	Type of Images	Class	Link	Accuracy	Algorithm
[44]	87	Chest X-ray im-ages	COVID-19 and Healthy X-ray images	Github and Kaggle	100	Residual Exemplar Local Binary Pattern
[2]	1020	Chest X-ray im-ages	COVID-19 and viral pneumonia diseases	Alexion, Toshiba Medical System, Japan	82.0	AlexNet
[2]	1020	Chest X-ray im-ages	COVID-19 and viral pneumonia diseases	Alexion, Toshiba Medical System, Japan	83.7	VGG-16
[2]	1020	Chest X-ray im-ages	COVID-19 and viral pneumonia diseases	Alexion, Toshiba Medical System, Japan	87.13	VGG-19
[2]	1020	Chest X-ray im-ages	COVID-19 and viral pneumonia diseases	Alexion, Toshiba Medical System, Japan	84.80	GoogleNet
[2]	1020	Chest X-ray im-ages	COVID-19 and viral pneumonia diseases	Alexion, Toshiba Medical System, Japan	99.4	ResNet-101
[29]	2780	Chest X-ray im-ages	coronavirus, pneumonia, and normal X-ray imagery.	Guangzhou Medical Center, China	97.4	CovXNets
[32]	337	Chest X-ray im-ages	COVID-19 and Healthy X-ray images	Cohen et al. [9]	97.97	nCOVnet
Proposed	746	Chest X-ray im-ages	COVID-19 and Healthy X-ray images		99.73	Handcrafted features HOG trained with Random Forest

Bold indicates the proposed method

compared with other popular deep learning models such as by AlexNet, VGG-16, VGG-19, GoogleNet. The proposed method based on a Random Forest classifier trained using HOG feature vector achieves an accuracy of 99.73%, a significant improvement compared to existing classification schemes. This research demonstrated that radiologists and physicians could use the HOG trained with the Random Forest classifier as an automatic diagnostic system to classify the COVID X-ray dataset. Future investigation of our study may include the application of the present framework to the medical diagnosis of other diseases using different input image modalities. In addition, the framework can be further customized and applied to other domains where the presence of misinformation is of high concern, such as pharmaceutical drug information, pandemic management, financial advisories, online healthcare services, and cyber frauds.

Acknowledgements Malathy Jawahar acknowledges CSIR-CLRI for conducting this research work (A/2022/LPT/MLP/1667).

Data availability Data will be shared upon reasonable request.

Code availability Programming codes available upon reasonable request.

Declarations

Conflict of interest The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Ahmed A (2019) Pneumonia Sample X-Rays, GitHub. <https://www.kaggle.com/ahmedali2019/pneumonia-sample-xrays>. Accessed 10 Mar 2020
2. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A (2020) Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. *Comput Biol Med* 121:103795
3. Bahoura M, Rouat J (2001) Wavelet speech enhancement based on the teager energy operator. *IEEE Signal Proc Lett* 8(1):104–12. <https://doi.org/10.1109/97.889636>. Accessed Jan 2020
4. Barman U, Ridip Dev Choudhury (2020) Smartphone image based digital chlorophyll meter to estimate the value of citrus leaves chlorophyll using Linear Regression, LMBP-ANN and SCGBP-ANN. *J King Saud Univ-Comput Inf Sci*. <https://doi.org/10.1016/j.jksuci.2020.01.005>
5. Batool FE, Attique M, Sharif M, Javed K, Nazir M, Abbasi AA, Iqbal Z, Riaz N (2020) Offline signature verification system: a novel technique of fusion of GLCM and geometric features using SVM. *Multimed Tools Appl* 4:1–20
6. Breiman L (2001) Random Forests. *Mach Learn* 45(1):5–32
7. Chen J, Hu C, Chen L, Tang L, Zhu Y, Xu X, Chen L, Gao H, Lu X, Yu L (2020) Clinical study of mesenchymal stem cell treating acute respiratory distress syndrome induced by epidemic Influenza A (H7N9) infection, a hint for COVID-19 treatment. *Engineering* 6(10):1153–1161
8. Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, Cui J, Xu W, Yang Y, Fayad ZA, Jacobi A, Li K, Li S, Shan H (2020) CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology* 295:202–207
9. Cohen JP (2020) COVID-19 Chest X-Ray dataset or CT dataset, GitHub. <https://github.com/iecc8023/COVID-chestxray-dataset>. Accessed 10 Mar 2020
10. Cortegiani A, Ingoglia G, Ippolito M, Giarratano A, Einav S (2020) A systematic review on the efficacy and safety of chloroquine for the treatment of COVID-19. *J Crit Care* 57:279–283

11. Darlenski R, Tsankov N (2020) COVID-19 pandemic and the skin - what should dermatologists know? *Clin Dermatol* 38(6):785–787
12. Hall LO, Paul R, Goldgof DB, Goldgof GM. Finding COVID-19 from chest x-rays using deep learning on a small dataset. arXiv preprint arXiv: 2004020602020
13. Harlick RM, Shanmugam K, Dinstein IH (1973) Texture features for image classification [J]. *IEEE Trans Syst Man Cybernetics* 3(6):610–621
14. Haykin S (1994) *Neural networks: a comprehensive foundation*. Prentice Hall PTR
15. Holland M, Zaloga DJ, Friderici CS (2020) COVID-19 personal protective equipment (PPE) for the emergency physician. *Visual J Emerg Med* 19:100740
16. Hosseney M, Kooraki S, Gholamrezaezhad A, Reddy S, Myers L (2020) Radiology perspective of coronavirus disease 2019 (COVID-19): lessons from severe acute respiratory syndrome and Middle East respiratory syndrome. *Am J Roentgenol* 1–5. <https://doi.org/10.2214/AJR.20.22969>
17. Hou J, Gao T (2021) Explainable DCNN based chest X-ray image analysis and classification for COVID-19 pneumonia detection. *Sci Rep* 11(1):1–15
18. Huang Z, Leng J (2010) Analysis of Hu's moment invariants on image scaling and rotation. 2010 2nd International Conference on Computer Engineering and Technology, vol 7. IEEE
19. Islam Md, Milon et al (2020) Wearable technology to assist the patients infected with novel coronavirus (COVID-19). *SN Comput Sci* 1(6):1–9
20. Janeliukstis R, Rucevskis S, Chate A (2020) Condition monitoring with defect localisation in a two-dimensional structure based on linear discriminant and nearest neighbour classification of strain features. *Nondestructive Testing and Evaluation* 35(1):48–72
21. Jawahar M, Jani Anbarasi L, Graceline Jasmine S, Narendra M, Venba R, Karthik V (2021) A machine learning-based multi-feature extraction method for leather defect classification. In: Smys S, Balas VE, Kamel KA, Lafata P (eds) *Inventive Computation and Information Technologies*, vol 173. Springer, Singapore
22. Kermany D, Zhang K, Goldbaum M (2018) Labeled optical coherence tomography (OCT) and chest X-ray images for classification. *Mendeley Data* 2. <https://doi.org/10.17632/rscbjbr9sj.2>
23. Lai C-C, Liu YH, Wang C-Y, Wang Y-H, Hsueh S-C, Yen M-Y, Ko W-C, Hsueh P-R (2020) Asymptomatic carrier state, acute respiratory disease, and pneumonia due to severe acute respiratory syndrome coronavirus 2 (SARSCoV-2): facts and myths. *J Microbiol Immunol Infect* 53(3):404–412
24. Lau H, Khosrawipour V, Kocbach P, Mikolajczyk A, Ichii H, Schubert J, Bania J, Khosrawipour T (2020) Internationally lost COVID-19 cases. *J Microbiol Immunol Infect* 53(3):454–458
25. Li Z, Ge J, Yang M, Feng J, Qiao M, Jiang R, Bi J, Zhan G, Xu X, Wang L (2020) Vicarious traumatization in the general public, members, and non-members of medical teams aiding in COVID-19 control. *Brain Behav Immun* 88:916–919
26. Li C, Yang Y, Ren L (2020) Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. *Infect Genet Evol* 82:104285
27. *Linear Discriminant Analysis - A Brief Tutorial*
28. Lippi G, Plebani M, Henry BM (2020) Thrombocytopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: a meta-analysis. *Clin Chim Acta* 506:145–148
29. Mahmud T, Rahman MdA, Fattah SA (2020) CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. *Comput Biol Med* 122:103869
30. Togaçar M, Ergen B, Comert Z (2020) COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. *Comput Biol Med* 121:10380
31. Özcan C, Ersoy KO, Oğul İÜ (2020) Fast texture classification of denoised SAR image patches using GLCM on Spark. *Turk J Electr Eng Comput Sci* 28(1):182–95
32. Panwar H et al (2020) Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. *Chaos Solitons Fractals* 138:109944
33. Rahbari D, Nickray M (2020) Task offloading in mobile fog computing by classification and regression tree. *Peer-to-Peer Networking and Applications* 13(1):104–22
34. Rahman T, Chowdhury M, Khandakar A (2020) COVID-19 Radiography Database, Kaggle. <https://www.kaggle.com/tawsifurrahman/COVID19-radiography-database/data#>. Accessed 20 Apr 2020
35. Rizal RA, Susanto M, Chandra A (2020) Classification of borax content in tomato sauce through images using GLCM. *Sinkron* 15:6–92
36. Rodriguez-Morales AJ, Cardona-Ospina JA, Gutiérrez-Ocampo E, Villamizar-Peña R, Holguin-Rivera Y, Escalera-Antezana JP, Alvarado-Arnez LE, Bonilla-Aldana DK, Franco-Paredes C, Henao-Martínez AF (2020) Clinical, laboratory and imaging features of COVID-19: a systematic review and meta-analysis. *Travel Med Infect Dis* 34:101623

37. Sahin EK, Colkesen I, Kavzoglu T (2020) A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping. *Geocarto Int* 11(4):341–363
38. Sharon JJ, Anbarasi LJ, Raj BE (2018) DPSO-FCM based segmentation and Classification of DCM and HCM Heart Diseases. 2018 Fifth HCT Information Technology Trends (ITT). IEEE
39. Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R (2020) COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. *J Adv Res* 24:91–98
40. Simonyan K, Zisserman. A very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 140915562014
41. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R (2020) World Health Organization declares global emergency, A review of the 2019 novel coronavirus (COVID-19). *Int J Surg* 76:71–76
42. Sousa RT, Marques O, Soares FAAMN, Sene IIG, De Oliveira LLG, Spoto ES (2013) Comparative performance analysis of machine learning classifiers in detection of childhood pneumonia using chest radiographs. *Procedia Comput Sci* 18:2579–2582. <https://doi.org/10.1016/j.procs.2013.05.444>
43. Tan J, Zhang S, Cao W, Gao Y, Li LC, Huo Y, Liang Z (2020) A multi-stage fusion strategy for multi-scale GLCM-CNN model in differentiating malignant from benign polyps. In: *Medical Imaging 2020: Computer-Aided Diagnosis*, vol 11314. International Society for Optics and Photonics, p 113141S
44. Tuncer T, Dogan S, Ozyurt F (2020) An automated Residual Exemplar Local Binary Pattern and iterative Relief based COVID-19 detection method using chest X-ray image. *Chemometr Intell Lab Syst* 203:104054
45. Ullah SM, Azmat et al (2021) Scalable telehealth services to combat Novel Coronavirus (COVID-19) pandemic. *SN Comput Sci* 2(1):1–8
46. Yang P, Liu P, Li D, Zhao D (2020) Corona Virus Disease 2019, a growing threat to children? *J Infect* 80(6):671–693
47. Zhang J-f, Yan K, Ye H-h, Lin J, Zheng J-j, Cai T (2020) SARS-CoV-2 turned positive in a discharged patient with COVID-19 arouses concern regarding the present standard for discharge. *Int J Infect Dis* 97: 212–214
48. Zu ZY, Jiang MD, Xu PP, Chen W, Ni QQ, Lu GM, Zhang LJ (2020) Coronavirus disease 2019 (COVID-19): a perspective from China. *Radiology* 296(2):E15–E25

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Malathy Jawahar¹ · J. Prassanna² · Vinayakumar Ravi³ · L. Jani Anbarasi² · S. Graceline Jasmine² · R. Manikandan⁴ · Ramesh Sekaran⁵ · Suthendran Kannan⁶

Malathy Jawahar
malathyj@clires.in

J. Prassanna
prassanna.j@vit.ac.in

L. Jani Anbarasi
janiyanbarasi.l@vit.ac.in

S. Graceline Jasmine
graceline.jasmine@vit.ac.in

R. Manikandan
srmanim75@gmail.com

Ramesh Sekaran
sramsaran1989@gmail.com

Suthendran Kannan
k.suthendran@klu.ac.in

¹ Leather Process Technology Division, CSIR-Central Leather Research Institute, Adyar, Chennai 600020, India

² School of Computer Science and Engineering, Vellore Institute of Technology, 600 127 Chennai, India

³ Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia

⁴ School of Computing, SASTRA Deemed University, Thanjavur, India

⁵ Department of Information Technology, Velgapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India

⁶ Department of Information Technology, Kalasalingam Academy of Research and Education, Srivilliputhur, India