Electronic Letters on Computer Vision and Image Analysis 19(3):85-97;2020

Covid19 Identification from Chest X-ray Images using Machine Learning Classifiers with GLCM Features

Sudeep D. Thepade, Shalakha V. Bang, Piyush R. Chaudhari, Mayuresh R. Dindorkar

Computer Engineering Department, Pimpri Chinchwad College of Engineering, SPPU, Pune, India

Received 2 August 2020; Revised 3 September 2020; Accepted 19 October 2020

Abstract

From staying quarantined at home, practicing work from home to moving outside wearing masks and carrying sanitizers, every individual has now become so adaptive to so called 'New Normal' post series of lockdowns across the countries. The situation triggered by novel Coronavirus has changed the behaviour of every individual towards every other living as well as non-living entity. In the Wuhan city of China, multiple cases were reported of pneumonia caused due to unknown reasons. The concerned medical authorities confirmed the cause to be Coronavirus. The symptoms seen in these cases were not much different than those seen in case of pneumonia. Earlier the research has been carried out in the field of pneumonia identification and classification through X-ray images of chest. The difficulty in identifying Covid19 infection at initial stage is due to high resemblance of its symptoms with the infection caused due to pneumonia. Hence it is trivial to well distinguish cases of coronavirus from pneumonia that may help in saving life of patients. The paper uses chest X-ray images to identify Covid19 infection in lungs using machine learning classifiers and ensembles with Gray-Level Cooccurrence Matrix (GLCM) features. The advocated methodology extracts statistical texture features from X-ray images by computing a GLCM for each image. The matrix is computed by considering various stride combinations. These GLCM features are used to train the machine learning classifiers and ensembles. The paper explores both the multiclass classification (X-ray images are classified into one of the three classes namely Covid19 affected, Pneumonia affected and normal lungs) and binary classification (Covid19 affected and other). The dataset used for evaluating performance of the method is open sourced and can be accessed easily. Proposed method being simple and computationally effective achieves noteworthy performance in terms of Accuracy, F-Measure, MCC, PPV and Sensitivity. In sum, the best stride combination of GLCM and ensemble of machine learning classifiers is suggested as vital outcome of the proposed method for effective Covid19 identification from chest X-ray images.

Key Words: Coronavirus, Covid19, Chest X-ray, Texture, Feature Extraction, Gray-Level Cooccurrence Matrix, Haralick Features, Machine Learning, Random Forest, Logistic, Multiple Layer Perceptron, Ensemble

1 Introduction

The novel Coronavirus 2019 shortly called Covid19 has been affirmed as pandemic by World Health Organization (WHO) in March 2020. It is caused due to Severe Acute Respiratory Syndrome and hence called SARS-CoV2. The outbreak of this virus in China, led to its spread across multiple countries and soon it became a Pandemic. Till date it has caused large number of deaths in human. The virus is still undefeated by medical professionals owing to its new emergence, late detection, fewer testing knowledge, lack of

Correspondence to: sudeepthepade@gmail.com

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.1277
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Spain

experience in medical professionals and most important its high resemblance with already known Pneumonia disease. The Pneumonia disease caused by bacteria, fungi, and viruses spreads infection in one or both lungs - Alveoli get filled with air or fluid pus which makes it difficult for person to breathe. The symptoms seen in this case include cough, shortness of breath, fatigue, fever, sweating etc. Very similar to these are the symptoms seen in persons affected by Covid19.

The chest radiographs i.e. the chest X-ray images of patient infected with the Covid19 exhibits characteristic patterns that are found in X-ray images of pneumonia infected patient. Imaging departments have also confirmed that radiological findings of Covid19 on chest X-ray images are those very much similar to the pneumonia. Copious amount of research has taken place in identification and classification of pneumonia using chest X-ray images. Although Coronavirus samples are getting tested widely using Reverse Transcription Polymerase Chain Reaction (RT-PCR), it alone might not offer brake on the global spread of Covid19. Thus, chest X-ray images can prove to be a considerate tool for triage of Covid19 infected patients.

The ultra-rapid transmission of this virus is creating immense pressure on the global medical and health authorities. In such gigantic provocative situation, there is an urgent need to devise a method that can well distinguish Covid19 suspected patients from Normal or Pneumonia infected patients by making use of chest X-ray images of concerned individuals. Even after diagnosing chest X-ray images one finds it difficult to distinguish between pneumonia infected and suspected Covid19 patients. Researchers around the world are attempting to devise some texture based cognitive methods for identification of Covid19. Thus, current research work focuses on extracting texture features from chest X-ray images using Second-order statistical moments-based techniques. The Haralick Features (GLCM) are extricated from each chest X-ray image to gain some distinguishable insights related to texture of chest X-ray image. The advocated technique aims at early detection of suspected Covid19 cases by categorizing X-ray images into one of the three mentioned classes Covid19, Normal and Pneumonia infected. Moreover, binary classification categorizes images into Covid19 infected or Non-Covid19 infected.

The main contributions of the paper and proposed method are

- Ability to identify Covid19 infection from chest X-ray images considering high similarity with pneumonia.
- Decision about the GLCM stride combinations giving more efficient Covid19 identification model.
- Performance assessment of Machine Learning classifiers for better Covid19 identification.
- Proposing the best possible ensemble of machine learning classifiers for more efficient Covid19 identification.

More details regarding literature survey in Pneumonia and Covid identification are elaborated in section 2. Section 3 throws light on entire flow of schemed approach and section 4 presents essential data on experimental environment and performance metrics used. Results achieved by proposed method are discussed in section 5 while conclusion is outlined in section 6.

2 Literature Survey

The Novel Coronavirus Pandemic resulted in immense loss of health, wealth, and economy. The medical practitioners or authorities are facing lot of challenges in detection and treatment of patients affected with Covid19. The radiology experts are in progressive search of finding effective and early identification methods for Covid19. The research carried out till date in identification of Covid19 and Pneumonia uses various texture features based cognitive methods. Few of such methods are summarized below.

Nanditha Krishna et al. advocates extraction of all 14 Haralick texture features from chest X-ray images in [1]. Authors have found that 3 texture features namely variance, sum average and sum variance provide discriminative features to classify X-ray images into two classes - normal lungs or lungs affected with pneumonia. The results are validated on a dataset obtained from Bangalore based medical college and hospital. The dataset consists of total of 22 images (11 normal lungs and 11 pneumonia affected lungs).

Nitin Singh et al. have proposed a state-of-art technique in [2] that uses dataset consisting of chest X-ray images of normal people and those infected with pneumonia. Authors have used the combination of wavelet transform method and Gray-Level Cooccurrence Matrix method to extract six time and frequency domain

features for detecting Pneumonia. The feature matrix created by fetching statistical GLCM texture features from an image, is used to train algorithms like K-nearest neighbours (KNN) & Support Vector Machine (SVM), achieving better accuracy of 92.6% with weighted KNN model.

Tulin Ozturk et al. have developed a model in [3] that performs two class classification (covid19/No-findings) as well as three class classification (covid19/No findings/Pneumonia). Authors have used deep learning model, darknet-19 - a classifier model. The inputs are standardized using Batch Normalization method. The performance of this fully automated method is assessed by various parameters of Accuracy, Sensitivity, F-Measure and Precision. Results obtained by using proposed model can be improvised by training the model on larger dataset containing a greater number of covid19 xray images.

In [4] Prabira Kumar Sethy et al. makes use of transfer learning approach for extracting deep features using 13 different pretrained CNN models. The ResNet50 model outperforms other 12 CNN models achieving highest accuracy of 95.33%. The method also compares this approach with other traditional methods of LBP+SVM, HOG+SVM and GLCM+SVM by evaluating performance on these combinations. The author finds out that best classification accuracy is obtained by LBP+SVM followed by GLCM+SVM. The chest X-ray images are collected from GitHub that consists of in all 381 images.

Abhishek Sharma et al. have attempted to automate the diagnosis process of pneumonia infection by considering histogram calculation and OTSU thresholding [5]. The image dataset of Japan Society of Radiological Technology (JSRT) is used for analysis of proposed method, consisting of 40 chest X-ray images. Here the resized input chest X-ray image is histogram equalized and then the abdomen area (region of interest) is cropped. From this region of interest, pneumonia clouds are detected by using OTSU thresholding followed by computing ratio of healthy region to entire lung region.

Abolfazl Zargari Khuzani et al. have proposed a method in [6] which distinguishes a Covid19 patient from pneumonia patient using machine learning classifier. The method extracts global image features from entire X-ray image without lesion segmentation, which reduces the need of huge training data. The dataset is categorized as images belonging to covid19 class, Normal class, and Pneumonia class. A dimensionality reduction method, Principal Component Analysis (PCA) is explored for synthesizing a set of optimal features. Texture features like Wavelet, GLCM and GLDM are explored to construct a feature array from each X-ray image of chest in both spatial and frequency domains.

3 Proposed Covid19 Identification model from Chest X-ray Images using Machine Learning Classifiers with GLCM Features

The present work comes up with identification of Covid19, putting forward a method that marks the extraction of texture features from jpeg 8-bit input X-ray image [9] using Gray Level Cooccurrence Matrix (GLCM). The Cooccurrence matrix is created for an entire single input chest X-ray image. Five different mix of strides i.e. (1), (1, 2), (1, 2, 4, 8), (1, 2, 4, 8, 16, 32) and (1, 2, 4, 8, 16, 32, 64, 128) are considered for computing GLCM. Six statistical features extracted from GLCM are contrast, homogeneity, dissimilarity, correlation, angular second moment and energy. These features are united to constitute a solitary feature vector. Finally, the multiclass classification and two class classification performed using Random Forest classifier.

Initially X-ray images of varied sizes are pre-processed. Size of the Co-occurrence matrix depends upon the maximum number of unique or distinct intensity values of pixels in an image. Greater the number of distinct intensity values, highly accurate will be the extracted textural information. But more number of distinct gray levels increases the size of co-occurrence matrix resulting in increase of computational cost and time. Hence, to reduce the size of GLCM matrix, all the intensity values of image are quantized to 16 levels and these images are further considered for construction of GLCM matrix.

The GLCM is a second order statistical method that considers relationship between two pixels. The GLCM of an image is computed for an offset given in equation 1.

$$\text{Offset} = (\text{Stride}, \text{Angle}) = (\delta, \theta) \tag{1}$$

Where, δ = [1], [1, 2], [1, 2, 4, 8], [1, 2, 4, 8, 16, 32], [1, 2, 4, 8, 16, 32, 64, 128]

$$\theta = [0, \frac{\Pi}{4}, \frac{\Pi}{2}, \frac{3\Pi}{4}]$$

The statistical Haralick features are extracted from the cooccurrence matrix created from an image. The properties are explained below along with the mathematical equations used. Let 'N' be the number of distinct pixel values or intensity values in an image, 'i' be the row index, 'j' be the column index and M_{ij} be the $(i, j)^{th}$ entry in cooccurrence matrix.

The energy is the measure of orderliness, which is said to detect disorders in the image. It is the addition of all squared elements in the GLCM as given in equation 2. The minimum value of energy is 0 while the maximum value is 2. Greater the uniform or smooth the image is, greater will be the value of energy.

$$\text{energy} = \sqrt{\sum_{i,j=0}^{N-1} M_{ij}^{2}} \tag{2}$$

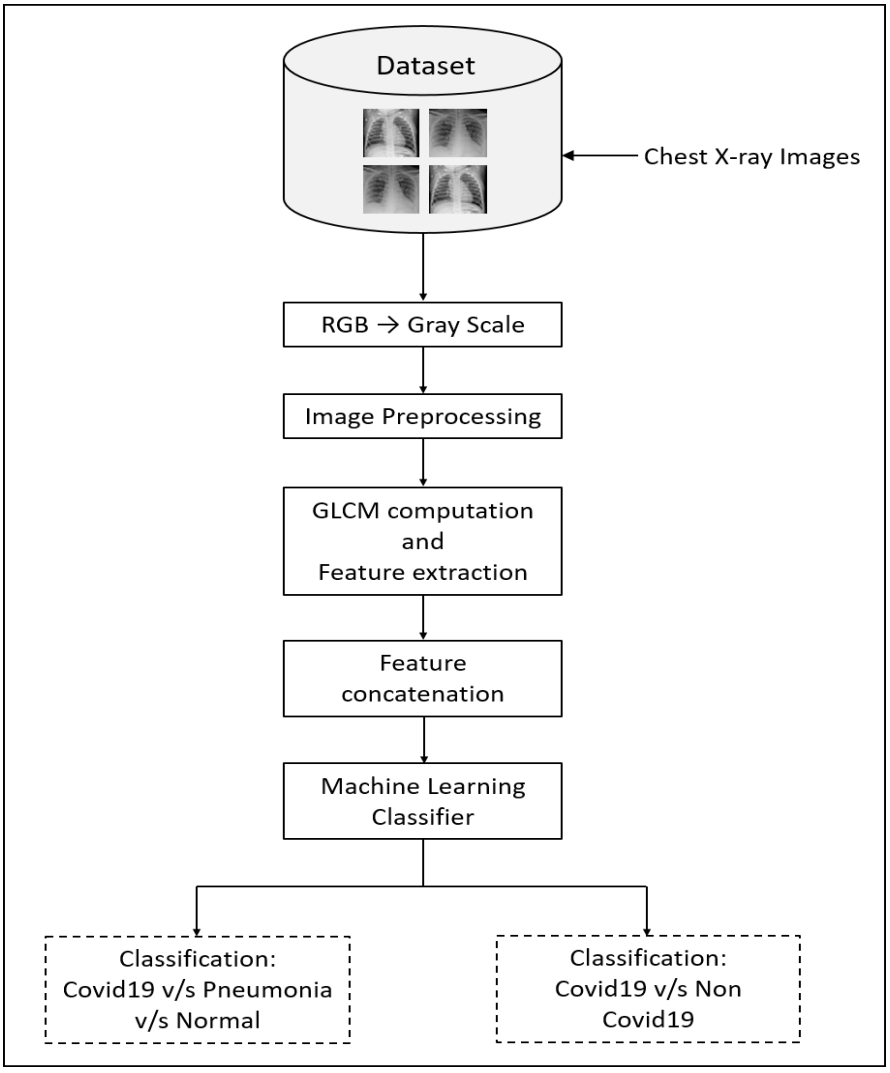

Figure 1: Block Diagram Describing Procedure of Identification of Covid19 using GLCM Feature Extraction Method

Contrast measures the intensity difference between the neighbouring pixels in cooccurrence matrix. It returns local level variations in an image. It is also called as 'Difference moment' and is represented as equation 3. It is the variation between largest and smallest value in a continuous group of pixels. The constant image will have contrast equal to 0.

$$\text{contrast} = \sqrt{\sum_{i,j=0}^{N-1} M_{ij}(i-j)^2} \tag{3}$$

Homogeneity generally referred as 'Inverse Difference Moment' measures similarity of pixel values. It has the highest value when all the pixel values in an image are alike. The range of homogeneity varies along [0, 1]. Its value is 1 for diagonal GLCM. It is strongly but inversely proportional to the contrast measure. The weights decrease exponentially from the diagonal. The equation 4 depicts homogeneity of an image.

$$\text{homogeneity} = \sum_{i,j=0}^{N-1} \frac{1}{1+(i-j)^2} M_{ij} \tag{4}$$

Dissimilarity is also called as Difference Average. It calculates the mean of difference between the gray level distributions of an image, which is exhibited as equation 5.

$$\text{dissimilarity} = \sum_{i,j=0}^{N-1} M_{ij} |i-j| \tag{5}$$

Angular Second Moment is the square of energy. It can be calculated using formula given in equation 6.

$$\text{ASM} = \sum_{i,j=0}^{N-1} M_{ij}^2 \tag{6}$$

The Correlation feature describes amount of linear interdependence between the neighbouring pixels. It measures how closely the neighbouring pixels are connected. The range of Correlation property lies between [-1, 1]. A perfectly positive correlated image has a value of 1 while a value of -1 specifies perfectly negative correlation. For a constant image its value is not defined i.e. NaN. Let μ and σ are mean and standard deviation respectively, and then correlation is calculated as equation 7.

$$\text{correlation} = \sum_{i,j=0}^{N-1} \frac{(i-\mu i)(j-\mu j)}{\sqrt{\sigma i^2 \sigma j^2}} \tag{7}$$

4 Experimentation Environment

The programming language used for extraction of features from input chest X-ray images is Python. The Scikit image library of python has useful functions to extract various features from GLCM of an image. The extracted GLCM features with various stride combinations are used to train considered machine learning classifiers using Waikato Environment for Knowledge Analysis (WEKA) [7]. Eight machine learning classifiers (Random Forest, Random Tree, REP Tree, Logistic, Simple Logistic, Multilayer Perceptron, Bayes Net and Naive Bayes) along with two ensembles 'Random Forest + Logistic + Simple Logistic' and 'Logistic + Simple Logistic + Multilayer Perceptron' are explored for Covid19 identification from chest X-ray images. The results are assessed on 10-fold cross validation.

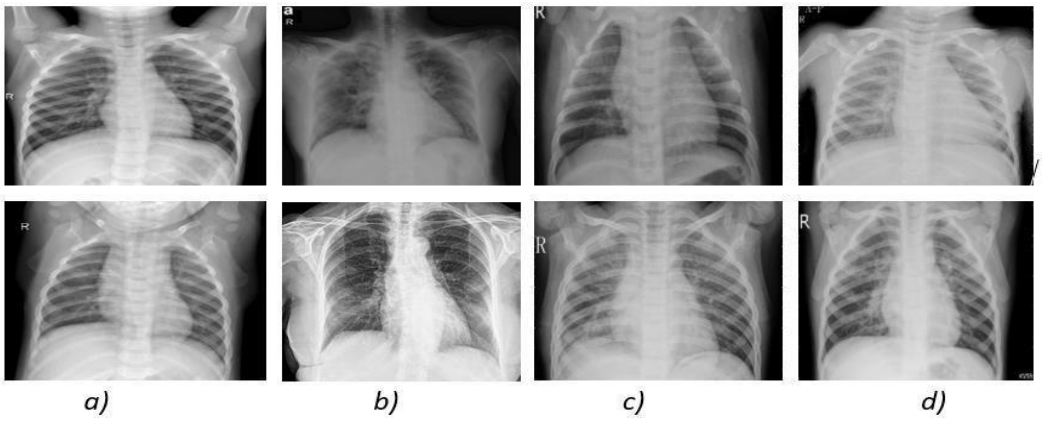

Figure 2: Specimen X- Ray images from dataset [8,9] a) X-ray images of normal lungs, b) X-ray images of Covid19 affected lungs, c) X-ray images of bacterial Pneumonia affected lungs, d) X-ray images of viral Pneumonia affected lungs

Covid19 dataset [8, 9] used for experimentation consists of labelled chest X-ray images that are bifurcated into 4 classes - class Covid19, class Pneumonia Bacterial, class Pneumonia Viral and class Normal. Total 240 chest X-ray images belonging to 50 patients are present in this dataset. Composition of the dataset is as follows: Covid19 class (60 images), Normal class (60 images), Pneumonia viral class (60 images) and Pneumonia bacterial class (60 images). All images are not of same size. The size ranges from 912 * 456 pixels to 2721 * 2438 pixels. Specimens of all categories are shown in figure 2. For present work, chest X-ray images belonging to Pneumonia viral and Pneumonia bacterial classes are merged, and new class is created named as Pneumonia for performing three class classification. Further, dataset from these images is curated for two class classification where images belong to either class Covid19 or class Non-Covid19. In this classification, Non-Covid class consists of 180 images (60 Normal, 60 Pneumonia viral and 60 Pneumonia bacterial). Covid class originally consisted of only 60 images. For balancing purpose, the images in Covid class are rotated by 90° and 180° to make total as 180.

The implementation results of propounded methodology are conveyed through classification Accuracy, Positive Predictive Value (PPV), F-measure, Sensitivity and Matthews Correlation Coefficient (MCC); respectively calculated as equations (8), (9), (10), (11) and (12) based on the confusion matrix obtained for each classifier.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$\text{Positive Predictive Value (PPV)} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F Measure} = \frac{2 * PPV * \text{Sensitivity}}{PPV + \text{Sensitivity}} \tag{11}$$

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{TP * TN - FP * FN}{\sqrt{(FP + TP)(FN + TP)(FP + TN)(FN + TN)}} \tag{12}$$

Where TP denotes true positive count , TN denotes true negative count , FP and FN denotes the count of false positive and false negative respectively. The Positive Predictive Value (PPV), F-Measure, Sensitivity & Matthews Correlation Coefficient (MCC) were considered for the Covid19 class as well as for the weighted average of all four classes.

5 Results and Discussion

The empirical observations of proposed Covid19 identification method experimented as three class classification (Covid19 v/s Pneumonia v/s Normal) and two class classification (Covid19 v/s Non-Covid19) approaches are given in subsections 5.1 and 5.2 respectively. These subsections communicate that the GLCM stride mix - (1, 2, 4, 8, 16, 32, 64, 128) shows better ability of Covid19 identification as compared to other four stride combinations attempted for GLCM feature extraction. Further for the GLCM stride combination: (1, 2, 4, 8, 16, 32, 64, 128), the performance appraises of eight machine learning classifiers and two majority voting-based ensembles is shown in subsection 5.3; indicating that the Logistic classifiers prove to be better choice in Covid19 identification. The ensemble combination: 'Random Forest + Logistic + Simple Logistic' is chosen based on the better performing individual machine learning classifiers i.e. Logistic family and have shown better ability of Covid19 identification from chest X-ray images.

5.1 Performance of GLCM Stride Combinations in Proposed Covid19 Identification Method using Random Forest with Three Class Classification Approach (Covid19 v/s Normal v/s Pneumonia):

Bar graph in figure 3 shows the percentage value of performance metrics achieved for all considered GLCM stride combinations of 1, (1, 2), (1, 2, 4, 8), (1, 2, 4, 8, 16, 32) and (1, 2, 4, 8, 16, 32, 64, 128) used for X-ray image feature extraction with Random Forest in proposed Covid19 identification method with three class classification approach. The accuracy increases gradually as observation moves from stride combination 1 to (1, 2, 4, 8, 16, 32, 64, 128). The maximum value of percentage accuracy is reached when all the distances starting from 1 to 128 are considered. The maximal percentage accuracy observed in case of three class classifications (Covid19 v/s Pneumonia v/s Normal) is 85%.

Stride	Accuracy	PPV	Sensitivity	F-measure	MCC
1	76.667	76.200	76.700	76.200	61.500
(1, 2)	77.917	77.400	77.900	77.400	64.000
(1, 2, 4, 8)	82.500	82.500	82.500	82.300	71.800
(1, 2, 4, 8, 16, 32)	83.333	83.500	83.300	83.100	73.100
(1, 2, 4, 8, 16, 32, 64, 128)	85.000	85.000	85.000	84.900	75.300

Table 1: Values (%) of performance metrics obtained for five considered GLCM stride combinations with Random Forest classifier in three class classification approach

Similar observation is noted for other performance metrics like PPV, F-Measure, Sensitivity and MCC as evident from figure 3. The exact values (%) obtained for each measure are tabulated in table 1. This indicates as observed from all performance metrics as Accuracy, PPV, F - measure, sensitivity and MCC; the GLCM stride (1, 2, 4, 8, 16, 32, 64, 128) gives better Covid19 identification ability in proposed method over other stride combinations for three class classification approach.

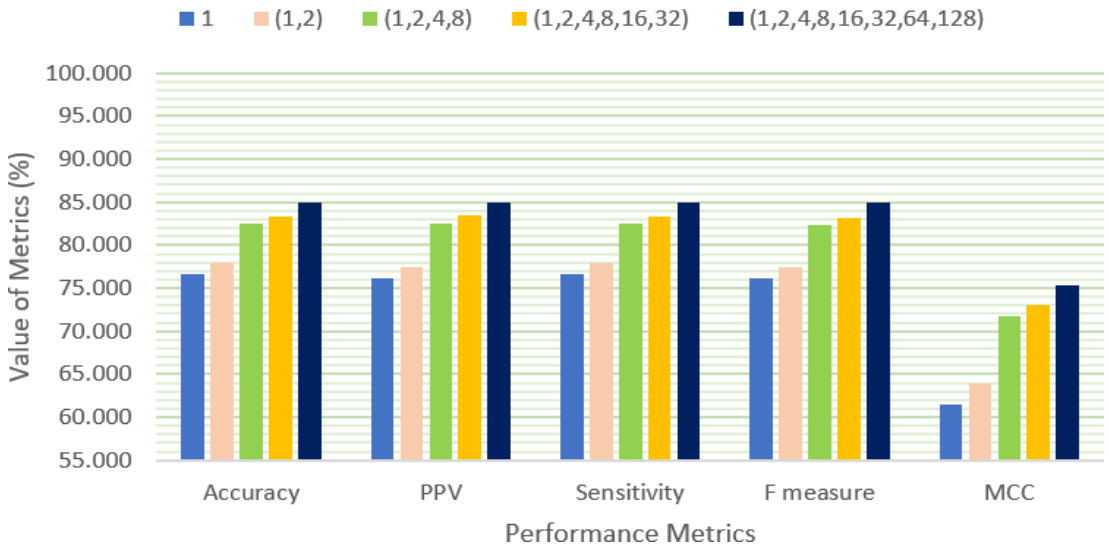

Figure 3: Performance of considered GLCM stride combinations with Random Forest classifier in proposed three class classification approach with reference to Accuracy, PPV, F-Measure, Sensitivity and MCC

5.2 Performance of GLCM Stride Combinations in Proposed Covid 19 Identification Method using Random Forest with Two Class Classification Approach (Covid19 v/s Non-Covid19)

The graph shown in figure 4 gives the percentage values of performance measures obtained for each GLCM stride mix in binary classification scenario with proposed Covid19 identification method. The greatest value of accuracy observed here is 94.444% with the GLCM stride combination: (1, 2, 4, 8, 16, 32, 64, 128); commenting its better Covid19 identification capability over other 4 considered combinations of GLCM strides.

Stride	Accuracy	PPV	Sensitivity	F measure	MCC
1	83.056	83.100	83.100	83.100	66.100
(1,2)	88.333	88.300	88.300	88.300	76.700
(1,2,4,8)	90.833	90.900	90.800	90.800	81.700
(1,2,4,8,16,32)	93.056	93.100	93.100	93.100	86.100
(1,2,4,8,16,32,64,128)	94.444	94.500	94.400	94.400	88.900

Table 2: Values (%) of performance metrics obtained for five considered GLCM stride combinations with Random Forest classifier in two class classification approach

Similar observations are acknowledged with reference to superior ability of GLCM stride: (1, 2, 4, 8, 16, 32, 64, 128) for present Covid19 identification model. Values of other considered performance metrics are also visualized in figure 4. Table 2 depicts the definite values obtained for all considered performance metrics measured in percentage.

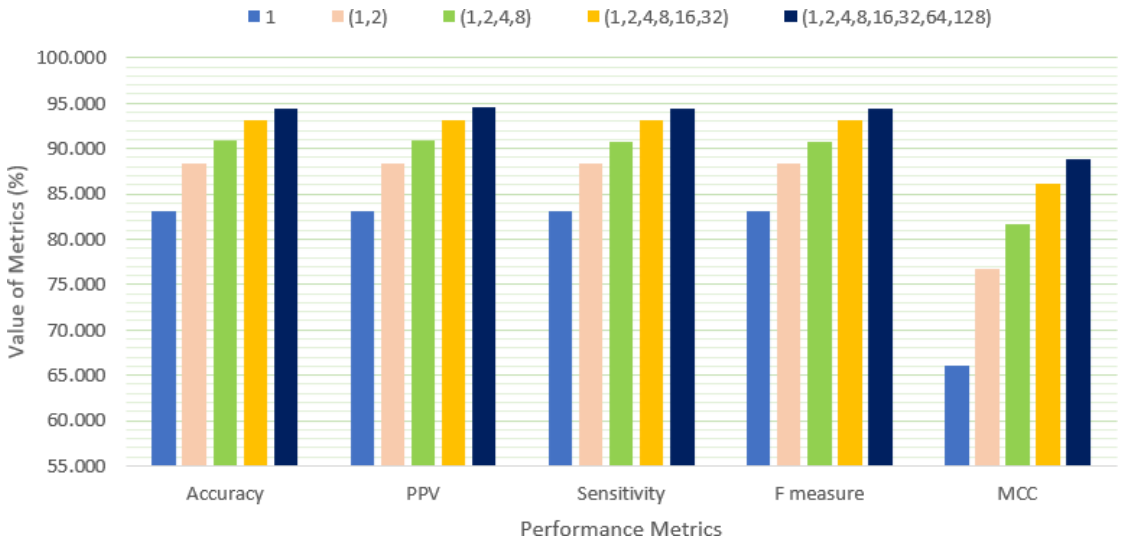

Figure 4: Performance of considered GLCM stride combinations with Random Forest classifier in proposed two class classification approach with reference to Accuracy, PPV, F-Measure, Sensitivity and MCC

From subsection 5.1 and 5.2 it can be presumed that the GLCM stride combination: (1, 2, 4, 8, 16, 32, 64, 128) gives overall best performance for all considered measures in both multiclass and binary classification context with binary classification performing superior to three class classification. Henceforth binary classification model (Covid19 v/s Non-Covid19) is evaluated on various machine learning classifiers and their ensembles. These results are discussed in subsection 5.3.

5.3 Performance Appraise of Machine Learning Classifiers and Ensembles in Proposed Covid19 Identification Method with Two class Classification Approach (Covid19 v/s Non-Covid19)

The performance of proposed Covid19 identification model is evaluated further using other machine learning classifiers and ensembles of classifiers. The model is evaluated on eight assorted machine learning classifiers alias Random Tree, Random Forest, REP Tree, Logistic, Simple Logistic, Bayes Net, Multilayer Perceptron and Naive Bayes classifiers & two majority voting-based ensemble combinations alias 'Random Forest + Logistic + Simple Logistic' and 'Logistic + Simple Logistic + Multilayer Perceptron. Here the earlier observed superior GLCM stride combination (1, 2, 4, 8, 16, 32, 64, 128) is used for feature extraction of chest X-ray images.

Machine Learning Classifiers	Accuracy	PPV	F Measure	Sensitivity	MCC	Average of All Performance Metrics
Random Forest	94.44	94.50	94.40	94.40	88.90	93.33
Random Tree	89.72	89.90	89.70	89.70	79.70	87.74
REP Tree	86.11	86.20	86.10	86.10	72.30	83.36
Logistic	98.61	98.60	98.60	98.60	97.20	98.32
Simple Logistic	95.28	95.30	95.30	95.30	90.60	94.36
Multilayer Perceptron	95.83	95.90	95.80	95.80	91.70	95.01
Bayes Net	82.78	82.80	82.80	82.80	65.60	79.36
Naive Bayes	78.61	79.70	78.40	78.60	58.33	74.73
Ensemble: 'Random Forest+Logistic +Simple Logistic'	99.17	99.20	99.20	99.20	98.30	99.01
Ensemble: 'Logistic+ Simple Logistic+ Multilayer Perceptron'	98.33	98.30	98.30	98.30	96.70	97.99

Table 3: Values (%) of performance metrics obtained for eight considered machine learning classifiers and two ensembles

Table 3 shows the values obtained for performance metrics across considered classifiers and ensembles. All the performance metrics like Accuracy, PPV, F-Measure, Sensitivity and MCC show that the better Covid19 identification is observed when Logistic classifier is used followed with Multilayer Perceptron, Simple Logistic and Random Forest classifiers. The assessment of machine learning classifiers experimented in proposed Covid19 identification method with various performance metrics is graphically shown in figure 5.

Further to boost the performance of proposed Covid19 identification model the majority voting based ensembles of better performing machine learning classifiers is done as 'Random Forest + Logistic + Simple Logistic' and 'Logistic + Simple Logistic + Multilayer Perceptron' as shown in table 3 and figure 6. Here it is observed that the best Covid19 identification capability in proposed method is demonstrated by ensemble - 'Random Forest + Logistic + Simple Logistic' as indicated by all performance metrics used - Accuracy, PPV, F Measure, Sensitivity and MCC.

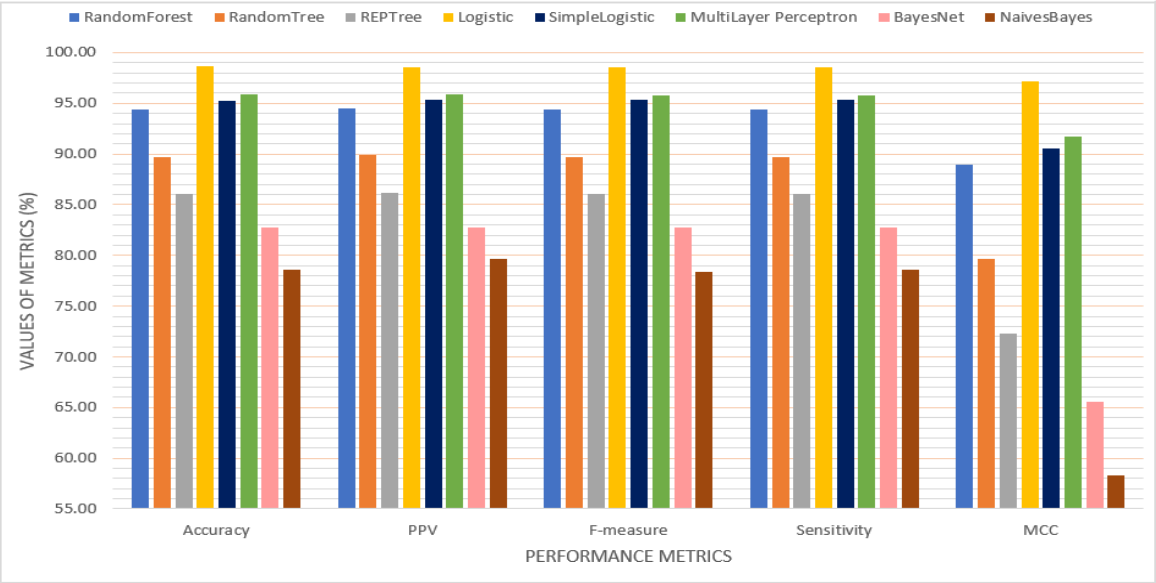

Figure 5: Performance Appraise of considered machine Learning Classifiers in proposed covid19 identification Method with reference to Accuracy, PPV, F-Measure, sensitivity and MCC

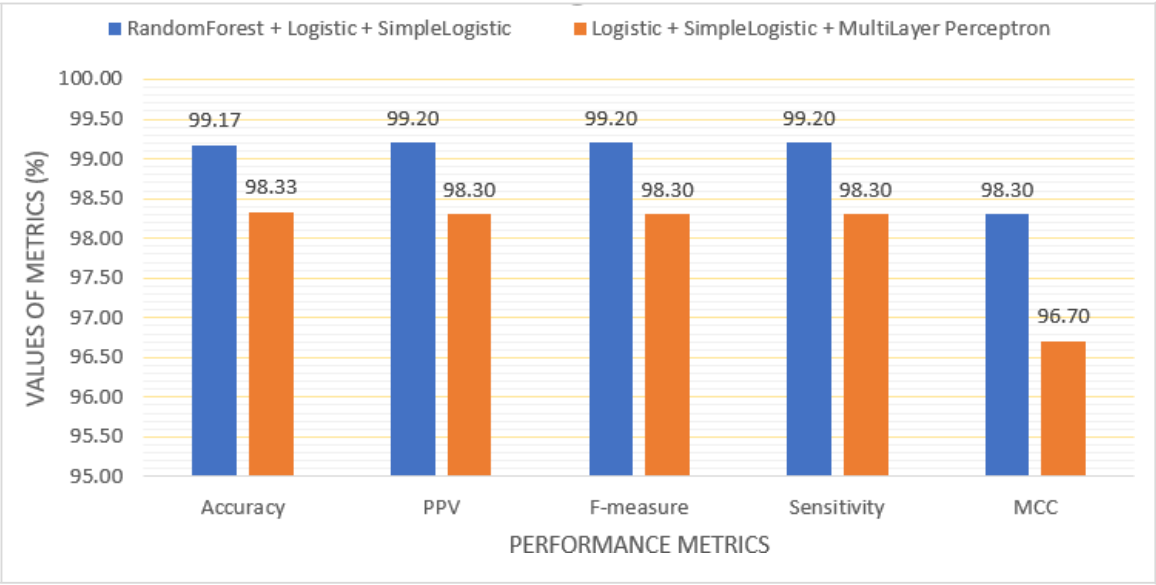

Figure 6: Comparison of performance of considered machine learning classifier ensembles in terms of performance metrics used for evaluation in proposed Covid19 identification method

Thus, the proposed method of Covid19 identification from chest X-ray images using machine learning classifiers with GLCM Features has given impressive performance on a test X-ray image dataset for GLCM stride combination (1, 2, 4, 8, 16, 32, 64, 128) and ensemble 'Random Forest + Logistic + Simple Logistic' as indicated by the performance metrics as Accuracy (99.17%), PPV (99.20%), F-Measure (99.205%), Sensitivity (9.205%) and MCC (98.30 %).

Further the comparison of proposed model with other existing Covid19 diagnostic models on grounds of methodology used, number of X-ray images used in experimentation, performance metrics used for evaluation and percentage accuracy achieved. It is summarized in table 4.

Methodology	Number of X-ray images	Performance Metrics	Classifiers	Accuracy
ResNet50 [4]	Covid (127) Normal (127) Pneumonia (127)	Accuracy, Sensitivity, FPR, F1 score	SVM	95.33%
Deep Learning model DarkCovidNet [3]	Covid (127) Normal (500) Pneumonia (500)	Accuracy, PPV, F-Measure, Sensitivity	DarkCovidNet-19	98.08% (Binary Classification) 87.02% (Multiclass Classification)
LBP [4]	Covid (127) Normal (127) Pneumonia (127)	Accuracy, Sensitivity, FPR, F1 score	SVM	93.4%
Neural Network + PCA [6]	Normal (140) Pneumonia (140) COVID-19 (140)	Accuracy, Precision, Sensitivity, F-Score	Neural Network based	94%
GLCM [4]	Covid (127) Normal (127) Pneumonia (127)	Accuracy, Sensitivity, FPR, F1 score	SVM	93.2%
Proposed method GLCM	Covid19 (180) Non-Covid19 (180)	Accuracy, PPV, F-Measure, Sensitivity, MCC	**Logistic**	**98.61%**
Proposed method GLCM	Covid19 (180) Non-Covid19 (180)	Accuracy, PPV, F-Measure, Sensitivity, MCC	**Ensemble of Logistic, Simple Logistic and RandomForest**	**99.17%**

Table 4: Comparison of proposed Covid19 identification model with existing Covid19 identification and classification methodologies

6 Conclusion

The Covid19 pandemic has threatened the entire world with exponentially increasing infected cases. The higher number of deaths in Covid19 due to lack of ability to detect the infection in initial stages is causing concerns in world. The chest X-ray images of Covid19 affected individuals are showing high resemblance with chest X-ray images of infections caused due to pneumonia, making it difficult to correctly diagnose the Covid19 infection in early stages using chest X-rays.

This paper presents effective method for identification of Covid19 by extracting GLCM texture features from chest X-ray images and classifying them using various machine learning classifiers and their ensembles. The method proposed in paper provides a means to classify highly similar X-ray images of pneumonia and Covid19 patients. The proposed method based on GLCM feature extraction gives impressive accuracy and thus may help medical practitioners in this global emergency.

The performance increases drastically with consideration of additional number of strides initially and then this improvement becomes marginal. Among all the GLCM stride combinations used for feature extraction, the stride of (1, 2, 4, 8, 16, 32, 64, 128) is observed to be better suited for proposed Covid19 identification method as validated using performance metrics - Accuracy, Positive Predictive Value (PPV), Sensitivity, F-Measure and Matthews Correlation Coefficient (MCC). The cases where more computational power is available and more time permitted, bigger stride sizes may be considered.

The paper has also explored performance appraise of assorted machine learning classifiers on the best observed GLCM stride combination for Covid19 identification. Among the experimented eight machine learning classifiers (Random Tree, Random Forest, REP Tree, Logistic, Simple Logistic, Multilayer Perceptron, Naive Bayes and Bayes Net); the Logistic classifier has shown better ability of Covid19 identification as indicated by all performance metrics.

Based on best performing individual machine learning classifiers, the proposed method is further experimented with two majority voting-based ensemble combinations of machine learning classifiers as 'Random Forest + Logistic + Simple Logistic' and 'Logistic + Simple Logistic + Multilayer Perceptron'. It is observed that the Covid19 identification from chest X-ray images is efficiently achieved by the ensemble 'Random Forest + Logistic + Simple Logistic'.

Thus the work explored in paper results into an efficient method having ability to identify Covid19 infection from chest X-ray (considering the high similarity with pneumonia) using GLCM features having stride of (1, 2, 4, 8, 16, 32, 64, 128) used with ensemble of 'Random Forest + Logistic + Simple Logistic'.

References

[1]. N. Deepa, Nanditha Krishna, G. Kumar, "Feature Extraction and Classification of X-Ray Lung Images Using Haralick Texture Features", *Smart and Innovative Trends in Next Generation Computing Technologies*, 899-907, 2018.

[2]. Nitin Singh, Rajneesh Sharma, Amit Kukker, "Wavelet Transform Based Pneumonia Classification of Chest X- Ray Images", *2019 International Conference on Computing, Power and Communication Technologies (GUCON)*, 2019.

[3]. Tulin Ozturk, et al., "Automated detection of COVID-19 cases using deep neural networks with X-ray images", *Computer in Biology and Medicine*, Vol. 121, 103792, 2020.

[4]. Prabira Kumar Sethy, et al., "Detection of Coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine", International *Journal of Mathematical, Engineering and Management Sciences*, 5(4):643-651, 2020.

[5]. Abhishek Sharma, Daniel Raju, Sutapa Ranjan, "Detection of pneumonia clouds in chest X-ray using image processing approach", *2017 Nirma University International Conference on Engineering (NUiCONE)*, 1-4, 2017.

[6]. Abolfazl Zargari Khuzani, Morteza Heidari, Ali Shariati, "COVID-Classifier: An efficient machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images", *medRxiv 2020.05.09.20096560*, 2020.

[7]. Mark Hall, et al., "The WEKA data mining software: An update", *ACM SIGKDD Explorations Newsletter*, 11(1):10-18, 2009

[8]. Joseph Paul Cohen, Paul Morrison, Lan Dao, "COVID-19 Image Data Collection", *arXiv:2003.11597*, 2020.

[9]. COVID-19 dataset available at website https://github.com/ieee8023/covid-chestxray-dataset (last referred on 10 May 2020).