Advances and Applications in Mathematical Sciences
Volume 20, Issue 12, October 2021, Pages 3017-3026
© 2021 Mili Publications

CLASSIFICATION OF COVID-19 PATIENT'S CHEST X-RAY IMAGES USING MACHINE LEARNING TECHNIQUES – A COMPARATIVE ANALYSIS

A. A. SHERNAS MOL and M. K SABU

Department of Computer Applications
Cochin University of Science
and Technology, Kalamassery, India
E-mail: shernas2131@gmail.com
sabumk@cusat.ac.in

Abstract

Covid-19 pandemic is a major health thread all over the world. Early detection is the only solution to control the spread of disease. Chest X-rays plays a key role in the diagnosis of Covid-19 since the viral test and the antibody test may take time to get the result. These tests sometimes give the result negative for infected persons. Chest X-rays are also cost effective when compared to other diagnosis tests for Covid-19 patients. Medical image analysis requires more efforts as the data increases rapidly. Due to high risk of work in this area, a Computer aided technique can lead to diagnose Covid-19 accurately than the radiologist. Better solution is to use machine learning techniques for risk assessment and treatment planning. This model can classify Covid-19 patients, Pneumonia patients and healthy patients based on their chest X-rays. Statistical measures are used in machine learning to retrieve the hidden information present in the image that may be used for good decision-making. X-ray images are gray scale images with almost the same textural features. In our model the traditional textural feature Gray Level Co-occurrence matrix (GLCM) is used to extract the information of pixel intensities between neighbouring pixels in a small region in the chest X-ray images of the patients. Then these extracted features of the patients are given to different conventional machine learning techniques like *K*-Nearest neighbor, Naive-Bayes Classifier, Support Vector machine for classification. Comparison of these classifiers are done on the basis of accuracy and found to be less. Then advanced machine learning ensemble methods were tried for classification. The ensemble methods like Random forest and XGBoost are used for classification. The comparative study of the model shows that classifying the X-ray image dataset with the combination of GLCM and ensemble methods gives better result than using GLCM with traditional machine learning methods. Our model has less computation time and it requires less memory (cost effective).

2010 Mathematics Subject Classification: 68U10.

Keywords: ensemble methods, machine learning, SVM, KNN, Naive Bayes, random forest, XGBoost.

Received October 13, 2020; Accepted November 5, 2020

I. Introduction

Covid-19 first appeared in Wuhan, China in December 2019 and become pandemic within two months. The virus causing Covid-19 (novel Coronavirus) is termed as SARS-CoV-2. Some symptoms related to Covid-19 include fever, cough, sore throat, headache, fatigue, muscle pain and shortness of breath [1]. Common test for Covid-19 diagnosis is real time reverse transcription – polymerase chain reaction (RT-PCR). In the initial stage of this pandemic Chinese clinical centres had insufficient test kits with high rate of false negative results and the doctors are forced to make a diagnosis only based on clinical and chest X-ray and CT results [2] [3]. The result of RT-PCR with CT or X-ray images can be used as a screening tool. Timely detection of the disease enables the health workers to give care required and the isolation of the patient to prevent spreading of the disease. However, due to the lack of sufficient chest X-ray images the classification of Covid-19 patients remains the biggest challenge in deep learning. But with these limited number of images we can develop a good machine learning model. Machine learning application increases daily with respect to image-based diagnosis, disease prediction and risk assessment [4]. The proposed model is a fully automatic Covid-19 detection system which can automatically analyse the presence of coronavirus quickly from the focused images with minimal or no human intervention. Comparison of five different classification models is conducted. First challenge for this work was the unavailability of dataset for the Covid-19 patient's chest X-ray. Our dataset consists of 891 images, collected from two different open sources available in the internet. In this study, a machine learning model is proposed for the automatic diagnosis of Covid-19. The proposed model uses texture feature extraction method, GLCM for feature extraction. Then the features are given to machine learning model for classification.

The motivation of this comparative study is to give a review of machine learning algorithms in medical image analysis based on literature review, current work and future directions. Previous related works done in this area are discussed in Section 2. In Section 3 the proposed method is discussed. Section 4 deals with dataset. Experiment and result analysis are discussed in section 5. Finally, conclusions are drawn in Section 6.

II. Related Works

This section offers a brief review of different works that explicate the theoretical aspects of the medical image analysis. Computational models are also discussed. Health Organization (WHO) declared Covid_19 as pandemic. It has the symptoms related to Pneumonia that can be diagnosed by blood test and medical imaging like Chest X-ray and CT. The only way to control the spread of the disease is early diagnosis and to isolate the patients. In [5] Lu and his team worked with GLCM features and Local Binary Pattern (LBP) features on different benchmark datasets and compare their classification performance. GLCM texture feature extraction method is used by Mohanaiah et al. for feature extraction from X-ray images. For histopathological image classification Saban and his team worked with texture features like GLCM, LBP, Local Binary Gray Level Co-occurrence Matrix (LBGLCM), Grey Level Run Length Matrix (GLRLM) and Segmentation based Fractal Texture Analysis (SFTA). GLCM, LBP, Histogram of Oriented Gradients (HOG) and combination of these methods are used to extract features in [8] from *X*-ray images. LBP variants are used for medical image analysis by Nanni and his team. In [10] for the diagnosis of polyp texture feature GLCM is combined with Convolutional Neural Network (CNN) for classification. To improve the classification accuracy of Support Vector Machine (SVM) in multiclass classification problems Saad et al. took SVM kernels like Linear, radial based sigmoid and polynomial to classify chest X-rays based on lung nodules. In [12] SVM is used to identify pneumothorax from Chest X-ray. Then Sobel edge detectors are used to segment the abnormal region. Camlica et al. extract significant features from the lungs using LBP feature and image classification and retrieval was done using SVM. This method delivers comparable classification accuracies but exhibits less computational cost and less memory for storage. Several physical properties of porous media like porosity, average pore size and specific surface area are used to train the CNN. In [15] Corona virus is automatically detected using a novel combination of GLCM, LBP and HOG features from X-ray images belonging to Image Retrieval in Medical applications (IRMA) database. From X-ray images Kumar et al. extract features of Covid-19 and Normal patients and use Random Forest and XGBoost to classify the patients. After dimensionality reduction of features obtained from *K*-means clustering in [17] they classify the images based on SVM and CNN. Singh et al. initially crop and remove the Gaussian noise from the mammogram and give the GLCM features to Ada Boost and Random Forest for classification. In [19] combination of different features like Haralick Statistical, first four moments, Texture and shape are used as features and use Random Forest to classify Osteoarthritis using knee X-ray image. *K*-Nearest Neighbor algorithm is used by Alarabeyyat and his team to detect Breast cancer. In [21] performance analysis of SVM is compared with Naive Bayes to classify Oral X-ray images. In this experiment SVM outperforms Naive Bayes. Abdolshah et al. follows Scale Invariant Feature Transform (SIFT), Feature vectors, Bag of Visual Words (BoVW), and Tree Augmented Naive Bayes (TAN) approach to classify the shipping containers based on X-ray images. In [23] they use different pre-processing methods like circular Fourier filter, multivariate linear regression and independent component analysis before giving the X-ray images to Naive Bayes and Random Forest for classification. Gabor transform and Naive Bayes classifier are used by Ahmad to classify infection and fluid regions with in the lungs. In [25] feature extraction methods like GLCM, Local Directional Pattern (LDP), GLRLM, Grey Level Size Zone Matrix (GLSZM) and Discrete Wavelet Transform (DWT) with SVM is used to classify Covid-19 patients CT images. Yan et al. use XGBoost to identify the ill cases from the clinical data.

III. Proposed Method

The workflow of the proposed model is described in the following section. Data collection, pre-processing, feature extraction and classification are the different stages in the classifier implementation as illustrated in Figure 1. In the following subsections, each of this stage is elaborately explained.

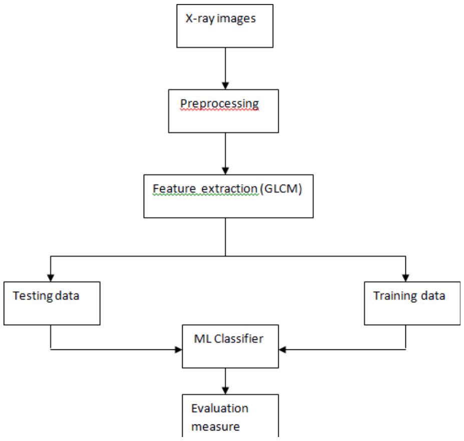


Figure 1. Proposed model.

IV. Dataset

A. Data collection

Data collection is one of the biggest barriers we faced during our study. Since we have no public Chest X-ray of Covid-19 patients. X-ray images for the study were collected from two different sources. Covid-19 X-ray images were collected from Cohen JP image from various open sources. From this we took 267 Covid19 X-ray images and discard all other class images. From Kaggle Chest X-ray8 dataset, 234 normal X-ray images and 390 Pneumonia Chest X-ray images were collected. Thus, a collection of 891 Chest X-ray images are used for our study. The dataset was divided into 67% for training the model and 33% for evaluation of the classification performance.

B. Preprocessing

Chest X-ray images of Covid-19 patients collected from internet highly varies, owing to diversity. X-ray images obtained from the internet are of different dimensions. To make this dataset uniform pre-processing was applied on it. Then the X-ray images are scaled to a common dimension of 480 x 580 pixels and normalize the data values.

C. Feature extraction

For accurate analysis of the image we have to extract proper features from the image. Feature extraction plays a major role in image processing.

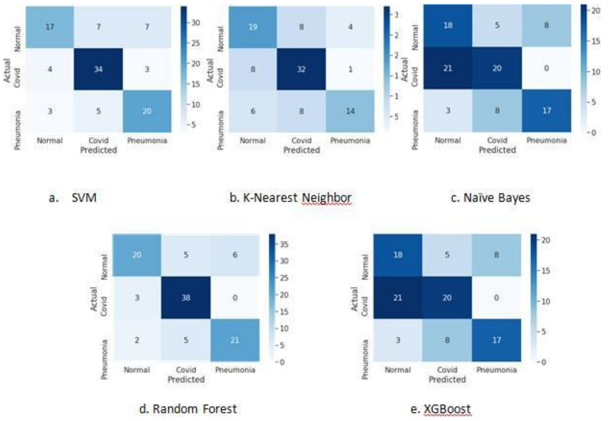
These features contain information about colour, shape, texture context etc. Spatial features, Transform features, Edge and boundary features, colour features, shape features, Texture features are some of the commonly used features in image processing. In this work we are taking the texture features from the X-ray image. Texture features give a significant contribution in image classification. It has information regarding size, density, shape, arrangement of its elementary parts.

A second order statistical texture feature, GLCM is used as feature extractor in this study. Special dependencies in an image are calculated using Gray Level Co-occurrence Matrix (GLCM). Fourteen textural features are extracted and saved as comma separated value (csv) file. Texture features like contrast, correlation, homogeneity, energy etc are used to capture the visual content of images. Angular Second Moment or energy is the sum of squares of entries. Entropy measures loss of information and also measures the image information and correlation measures the dependency of pixels with the neighbouring pixels.

D. Classifiers. Different classifiers like Support Vector Machine (SVM), K Nearest Neighbour (KNN), Gaussian Naïve Bayes and the Ensemble methods like Random Forest (RF) and XGBoost are used in this study.

V. Experiment and Result Analysis

Implementation of the proposed work was done with Python programming language with its supporting libraries by Anaconda distributions like glob for retrieve files/pathnames, OpenCV for computer vision, PIL for image processing, matplotlib for visualization and other packages like NumPy for scientific computations. The experiment was conducted on a PC with the following configuration: Intel ® CoreTM i7-7700k CPU@ 4.20 GHz x 8, GeForce GTX 1050 Ti/ PCle/ SSE2 and 31.4 GB RAM.



a. SVM b. K-Nearest Neighbor c. Naïve Bayes

d. Random Forest e. XGBoost

Figure 2. Confusion matrix of different classifiers.

In this work, the first step is the collection of data. Then these raw images are processed as discussed earlier. Next step is the extraction of features from the X-ray images. After the feature-extraction was completed, five classifier models namely SVM, KNN, Gaussian Naive Bayes classifier, Random Forest and XGBoost were used for the classification of the data. An analysis of the results obtained from each of the classifier models is discussed below. For the analysis, the experiment was carried out for five times, by randomly splitting the dataset in the same ratio. Figure 2 illustrates the Confusion matrix of KNN classifier, SVM classifier, Naive Bayes classifier, Random Forest and XGBoost respectively.

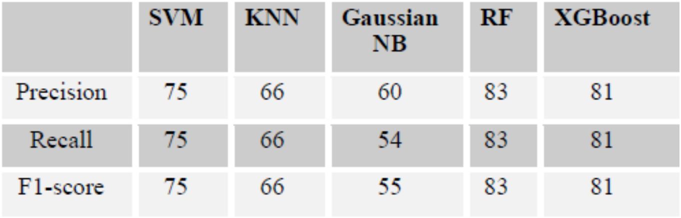
Accuracy is an important parameter to validate the model. The proposed model achieves an accuracy of 66% on K-Nearest neighbor, 54% on Naive-Bayes Classifier, 75% on Support Vector machine, 83% on Random forest and 81% on XGBoost.

Table 1. accuracy measure of different classifiers.

SVM	**KNN**	**Gaussian NB**	**RF**	**XGBoost**
75%	66%	54%	83%	81%

The accuracy obtained for the five models are shown in Table 1. From the table, it is clear that Random forest classifier gives better accuracy of 83%. Some other evaluation metrics for evaluating the classifier like precision, recall and f1-score are also calculated.

Table 2. Precision, recall and f1 score of different classifiers.



	SVM	KNN	Gaussian NB	RF	XGBoost
Precision	75	66	60	83	81
Recall	75	66	54	83	81
F1-score	75	66	55	83	81

Within total predicted actual class observations precision gives a measure of correctly predicted actual class observations. Where recall gives the information of correctly predicted actual class observation in the total actual class observation. F1 score is the harmonic mean of precision and recall. Classification summary for different models are shown in Table 2. The table compares the overall performance of different classifiers. From the results we can found that ensemble methods can be used for Chest X-ray image classification. From the results it can be inferred that Random forest classifier with GLCM features gives the highest overall accuracy than the traditional machine learning methods.

VI. Conclusion

During the last few years, machine learning has a tremendous growth in our daily life. In recent years machine learning can assist the humans in medical image diagnosis. Covid-19 can be detected from chest X-ray and CT images since it is a respiratory disease. The proposed model was implemented as an application of Image Processing, Computer Vision, and Machine Learning. In this chapter, we highlighted different machine learning algorithms in medical image analysis. Though, it is not the complete list, it gives an overview of some machine learning algorithms. We tested the model with different classifiers like Gaussian Naive Bayes, *K*-Nearest Neighbor, Support Vector Machine, Random Forest, XGBoost and observed that Random Forest provide the best accuracy (83%) for Covid-19, Normal and Pneumonia classification. Our model has less computation time and it requires less memory (cost effective). There are several challenges that affect the growth of the machine learning. One of the big challenges is medical image varies from one person to another. Analysing these images with more features may give better performance for classification. In our future work, we are planning to modify the existing model by adding a greater number of other significant features.

References

[1] T. Singhal, A review of coronavirus disease-2019(COVID-19), Indian J. Pediatr. 87(2020) 281-286.

[2] A. Bernheim, X. Mei, et al., Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection, Radiology (2020).

[3] C. Long and H. Xu, et al., Diagnosis of the Coronavirus disease (Covid-19): rRT-PCR or CT.

[4] M.de Bruijne, "Machine learning approaches in Medical image analysis: From detection to diagnosis", Vol.33 94-97.

[5] Lu, Dengsheng, and Qihao Weng A survey of image classification methods and techniques for improving classification performance, International journal of Remote sensing 28.5 (2007), 823-870.

[6] P. Mohanaiah, P. Sathyanarayana and L. Guru Kumar, Image texture feature extraction using GLCM approach, International journal of scientific and research publications 3.5 (2013), 1-5.

[7] Öztürk, Şaban, and Bayram Akdemir, Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA, Procedia computer science 132 (2018), 40-46.

[8] Bhende, G. Prachi, and A. N. Cheeran. A Novel Feature Extraction Scheme for Medical X-ray Images, Prachi. G. Bhende Int. Journal of Engineering Research and Applications www.ijera.com ISSN: 2248- 9622, (Part -6) 6(2) (2016), 53-60.

[9] L. Nanni, A. Lumini and S. Brahnam, Local binary patterns variants as texture descriptors for medical image analysis, Artificial intelligence in medicine 49(2) (2010), 117-125.

[10] Tan, Jiaxing, et al., GLCM-CNN: gray level co-occurrence matrix-based CNN model for polyp diagnosis, 2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2019.

[11] Saad, Mohd Nizam, et al. Multiclass Classification Application using SVM Kernel to Classify Chest X-ray Images Based on Nodule Location in Lung Zones, Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 9.1-2 (2017), 19-23.

[12] Chan, Yuan-Hao, et al. Effective pneumothorax detection for chest X-ray images using local binary pattern and support vector machine, Journal of healthcare engineering (2018).

[13] Camlica, Zehra, R. Hamid Tizhoosh, and Farzad Khalvati. Medical image classification via svm using lbp features from saliency-based folded data. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). IEEE, 2015.

[14] Alqahtani, Naif, et al., Machine learning for predicting properties of porous media from 2d X-ray images, Journal of Petroleum Science and Engineering 184 (2020), 106-514.

[15] Kassani, Sara Hosseinzadeh, et al. Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning-Based Approach, arXiv preprint arXiv:2004.10641 (2020).

[16] Kumar, Rahul, et al., Accurate Prediction of COVID-19 using Chest X-ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers" medRxiv (2020).

[17] Sayeed, Mohammed Azam, et al., Accelerated Diagnosis and Reporting of Patients using Analysis of Bulk Chest X-ray Images to Aid Impacted Healthcare System during Covid19.

[18] Singh, Vibhav Prakash, et al., Mammogram classification using selected GLCM features and random forest classifier, International Journal of Computer Science and Information Security 14.6 (2016), 82.

[19] Gornale, Shivanand S., Pooja U. Patravali, and Ramesh R. Manza, Detection of osteoarthritis using knee X-ray image analyses: a machine vision-based approach." International Journal of Computer Applications 145.1 (2016).

[20] Alarabeyyat, Abdulsalam and Mohannad Alhanahnah Breast cancer detection using k-nearest neighbor machine learning algorithm, 2016 9th International Conference on Developments in eSystems Engineering (DeSE). IEEE, 2016.

[21] G. Karthick, and R. Harikumar, Comparative Performance Analysis of Naive bayes and SVM classifier for Oral X-ray images, 2017 4th International Conference on Electronics and Communication Systems (ICECS). IEEE, (2017).

[22] Abdolshah, Majid, Mehdi Teimouri and Rohallah Rahmani. Classification of X-ray images of shipping containers, Expert Systems with Applications 77 (2017), 57-65.

[23] Brahim, Abdelbasset, et al., A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from the OsteoArthritis Initiative. Computerized Medical Imaging and Graphics 73 (2019), 11-18.

[24] Ahmad, Wan Siti Halimatul Munirah Wan, et al., Classification of infection and fluid regions in chest X-ray images, International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, (2016).

[25] Barstugan, Mucahid, Umut Ozkaya and Saban Ozturk, Coronavirus (covid-19) classification using ct images by machine learning methods, arXiv preprint arXiv:2003.09424 (2020).

[26] Yan and Li, et al., Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan, MedRxiv (2020).