CLASSIFICATION MODEL FOR COVID-19 AND PULMONARY (TB) FROM X-RAY IMAGES USING HOG-PCA-LEARNING ALGORITHMS

[*1]Folorunso, Sakinat Oluwabukonla; [1]Banjo, Oluwatobi Oluwaseyi; [2]Ayo, Femi Emmanuel; [1]Ogunyinka, Peter Ibikunle; [3]Folorunso, Temitope Sariy; [4]Folorunso, Mubarak Temidayo

[1]Department of Mathematical Sciences, OlabisiOnabanjo University, Ago-Iwoye, Ogun State
[2]Department of Computer Science, McPherson University, SerikiSotayo, Ogun State
[3]Department of Medicine, Olabisi Onabanjo University, Sagamu, Ogun State
[4]Department of Pharmacy, Olabisi Onabanjo University, Sagamu, Ogun State
[*1]Corresponding Author:sakinat.folorunso@oouagoiwoye.edu.ng

ABSTRACT

CoronaVirus Disease 2019 (COVID-19) is induced by a new virus SARS-CoV2. Its incidence is unprecedented and caused a huge dent to the health care system and the whole world. The hasty spread of COVID-19 and lack of fast diagnosis drove machine learning researchers to build intelligent response system to help the healthcare delivery personnel to manage the disease and the patient. The aim of this study is to build a COVID-19/ Pulmonary Tuberculosis (PTB) classification model from Chest X-ray (CXR) images. Due to small sample size of COVID-19 CXR image available, a four-phased method is adopted involving feature extraction, selection, modelling and classification.

The CXR images of lungs infected with COVID-19, PTB and Normal were obtained from databases. Features were extracted from these images by Histogram of Oriented Gradient (HOG)descriptor. Principal Component Analysis (PCA) technique was used to extract the most relevant features to enhance classification. For this study, from 1327 CXR image samples 46,657 features were extracted. But 675 relevant and important features were selected with 95% explained variance of PCA. A number of learning algorithms such as Support Vector Machine (SVM), *k*-Nearest Neighbor (*k*-NN), Random Forest (RF), eXtreme Gradient B boosting (XGBoost) and Decision Tree (DT)classifiers was used and evaluated. The experimental results obtained showed that SVM classifier produced best results of 0.97 based on precision, accuracy, F1-Score and recall metrics when compared to other learning algorithms.

Keywords: Pulmonary Tuberculosis, COVID-19, SARS-CoV2, HOG, PCA, Multiclass classification

***Accepted Date:** 10 Oct., 2020*

Introduction

COVID-19 is a respirational disease induced by a new coronavirus SARS-CoV2 with symptoms similar to viral pneumonia. The initial discovery of this virus at Wuhan China mark the starting of its spread across the world. The current state of the COVID-19 pandemic as at April 12th, 2020 is that the ratio of confirmed cases to actual death is 1 to 10 occurring in more than 208 countries/territories as reported by WHO. The great challenge threatening the public health is the high rise in the number of mortal patients. Some common clinical signs and symptoms exhibited by COVID-19 infested patients includes fever, cough, fatigue, reduced white blood cell counts, breathlessness, muscle pain and radiographic evidence of pneumonia. Consequently, some body part defect (e.g., trauma, Acute Respiratory Distress Syndrome (ARDS), heart attack and acute renal failure) and even death can happen in critical cases. One common diagnosis method for COVID-19 is Real-Time reverse-transcription–Polymerase-

Chain-Reaction (RT-PCR) test. However, National health and health commission of china recommended Computed Tomography Scan (CT scan) or CXR as the best diagnosis method due to some errors that could be detected in samples collected by RT-PCR diagnosis method . X-rays are used to envision the internal structures of a patient. CXR is a fast-radiological imaging technique effective to identify deformities in the lung, airways, heart, ribs and diaphragm.

PTB remains a world-wide health issuewith9:1.4 ratio of new cases and deaths respectively as reported in 2011. PTB is deadly but can be cured if detected early. CXR is a vital instrument for screening for PTB disease especially where it cannot be confirmed bacteriologically. CXR popular usage is due to its advanced image quality with digital radiography. COVID-19 and TB infested patients exhibit identical symptoms like fever, breathing difficulty and cough. Another similarity between the two diseases is that they both attacks mainly the lungs. Though the biological medium transmitting for both diseases mainly via close proximity, TB has a longer gestational period from exposure to disease with slower onset. Marimuthu *et al*., (2020) guided that primary intervention measure is important for TB patients and the early diagnosis and administration is important for both COVID-19 and TB patients.

This study aims to build a classification model for COVID-19 and PTB diseases using their CXR images. In other to build a model with a good classification performance, a 4-phased methodology is proposed. Image feature extraction with HOG was employed to extract features from the images at the first phase. Next, PCA was employed for feature selection in order to enhance classification accuracy. The resultant HOG-PCA feature vector obtained is an efficient representative of the HOG features. These features effectively represent the frontal CXR images for classification. Five (5) different machine learning models deployed for the classification task are: SVM, DT, *k*-NN, RF and XG Boost classifiers.

The remaining sections of the paper is planned as follows: section 2 presents the existing work in the domain of COVID-19 and machine learning. Section 3 presents materials and methods. Section 4 highlights the results and discussion. Section 5 conclude the work

Materials and Methods

This section presents the proposed dataset, pre-processing techniques, sampling schemes, models and metrics used in this study. The dataset used in this study will be freely available for download. The methodology workflow employed by this study is presented by Figure 1.

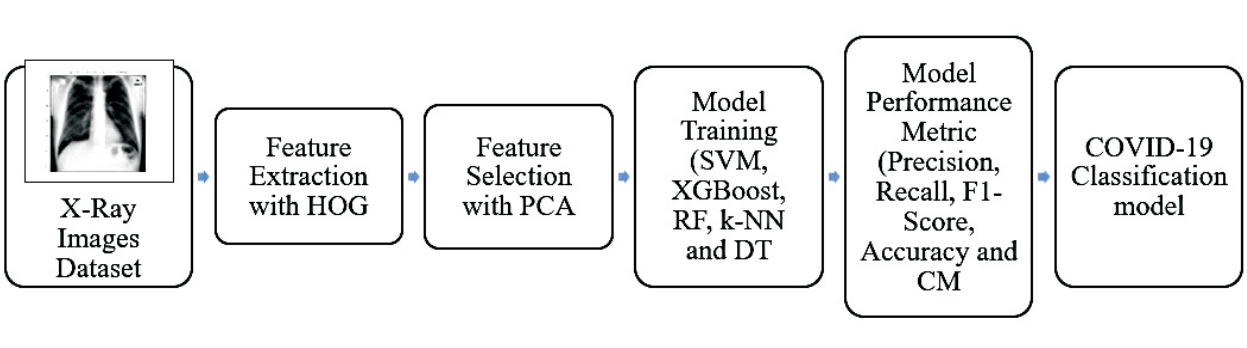


Figure 1: Framework for the HOG-PCA classification model

The dataset

The main features of the study dataset named OOU-20 and their different sizes are presented in Table 1.The dataset comprises of 1327 of frontal CXR images with 3 classes. The dataset was obtained from the following 3 different sources:

i. COVID-19 CXR images were obtained from (Cohen, *et al.*, 2020)

ii. 138 CXR images comprising of 58 pulmonary tuberculosis and 80 normal cases (Jaeger, *et al.*, 2014; Candemir, et al., 2014).

iii. 662 CXR images comprising of 336 pulmonary tuberculosis and 326 normal cases (Jaeger, *et al.,* 2014; Candemir, *et al.*, 2014)

Table 1: Class distribution of OOU-20 dataset

Class Label	Size
COVID-19	527
NORMAL	406
PTB	394
Totalnumber of Instances	1327

Pre-processing

Presenting the study as a classification problem: Let $S=\{(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)\}$ be the set of training instances of dimension d. $Y=\{y_1,y_2\cdots,y_n\}$ be the set of labels (COVID-19, PTB and Normal) where x_iis a feature with corresponding y_i label. The initial step taken in the image classification model is extraction of features. This is pertinent when the features extracted which is also the input data is extremely large and difficult to process in its raw form. Selection of important features will resolve this problem.

Extraction Method for features

The textual representative features of the study CXR image dataset was extracted by HOG descriptor. The technique is established on the ground that local object looks and shape in an image can be portrayed by the spread of intensity gradients or edge orientations. HOG method is implemented by dividing the images into cells and histogram of gradient orientations and compute the pixels within the cells. The image descriptor is represented by the combination of the resultant histograms. There are three major tasks performed to compute a HOG. The initial task is to compute the gradient values followed by computing the orientation binning of cell histogram. The final task then computes the descriptor's block and then normalize them. Dalal & Triggs, (2005) suggested the use of 1*D* mask size of [–1 0 1] when computing the gradient of image I(*x*,*y*) as shown by equations (1) and (2)

$$I_x(x,y)=I(x,y+1)-I(x,y-1) \quad (1)$$
$$I_y(x,y)=I(x-1,y)-I(x+1,y) \quad (2)$$

The magnitude |H(x,y)| and orientation θ of the gradient of image *I* can be computed from equations (3) and (4) respectively.

$$|H(x,y)| = \sqrt{I_x(x,y)^2 + I_y(x,y)^2} \quad (3)$$

$$\theta = \tan^{-1}\left(\frac{I_y(x,y)}{I_x(x,y)}\right) \quad (4)$$

The computation of the orientation bin involves creation of cell histograms channels. These histogram channels range over 0^O-180^O for signed channels. Choosing any value in the histogram channel to compute the gradient, each pixel inside the cell casts a weighted vote for an orientation-based histogram channel. Dalal & Triggs (Dalal & Triggs, 2005) experimented from his work that unsigned histogram channel of 〖20〗 ^0 giving a bin size of 9 $\left(\frac{180}{20}\right)$ histogram channels will give an optimal value and 20^0angular range. Then, the block is normalized choosing from a range of normalization methods as shown in equations (5) – (8)

$$\text{L2 - } norm: f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}} \quad (5)$$

$$\text{L2 - } Hys\ norm: f = \frac{\text{v}}{\sqrt{\|v\|_2^2 + e^2}} \text{ with max v} = 0.2 \quad (6)$$

$$\text{L1 - } norm: f = \frac{v}{\|v\|_1 + \text{e}} \quad (7)$$

$$\text{L1 - } sqrt: f = \sqrt{\frac{v}{\|v\|_1 + \text{e}}} \quad (8)$$

Where v = cooccurrence matrix and e = error term

In this study, the HOG method was implemented in Jupyter notebook with sklearn (Pedregosa, *et al.*, 2011). With the feature vectors extracted from the python code, feature selection task was directly performed on it achieving one of the study objectives of extracting feature from the CXR images. A bin size of 9 as proposed in the study of (Dalal & Triggs, 2005) was used to get a HOG feature set of size 81 each for all the frontal CXR images. The HOG parameters used in the study are shown in Table 2.

Table 2: HOG parameters used in this study

Models	Parameters
HOG	Block _ norm='L2-Hys' Cells _ per _ block = (8,8) Transform _ sqrt=True Pixels _ per _ cell = (8, 8)) Rescaled for better display in _ range = (0, 10)

Feature Selection with PCA

The aim of feature selection is to convert feature vector from a high D- dimension to a low H-dimension by removing less important, noise and redundant features. The 'curse of dimensionality' is removed by this process resulting to an improved the classification accuracy for the models. This study used PCA method by transforming feature dimensions to retain Principal Components (PC) accounting for most of the disparity in the original higher dimensional data (Hotelling, 1933). These PCs are achieved as linear combinations of the original variables.

Let $x_1, x_2, \cdots, x_n$ be the original dataset in D-dimensional space. The aim is to represent the dataset in a reduced subspace *H* with d < D. Let y_i: i=1, 2,⋯, n be the linear combinations of these variables such that $y = A^T (x - m_x)$ where $A = \alpha_{\_1} |, \alpha_{\_2} |, \cdots, \alpha_n |$ is the matrix whose columns are the eigenvectors of Σ, the covariance of the original dataset and m_x is the mean of the original data.

Classification Models, parameters and metrics

This section describes the models, their parameters and metrics used for the multi-class classification experiment. The task was performed with 5 learners: *k*-NN, SVM, DT, XG Boost; and RF. Table 2 presents the parameters of the model used in this study. Confusion matrix is a table representing the prediction performance of a model. The row and column represent the predicted and the actual class respectively as shown in Table 3. The formula for metrics computed from the confusion matrix is presented by equations (9) – (14)

Table 2: Model parameter sapplied in this study

Models	**Parameters**
SVM	C=100,probability=True, kernel = 'rbf', gamma = 0.001
DT	Criterion="gini",random_state=30
RF	n_estimators=700, max_depth=3
k-NN	Scikit learn default values
XGBoost	Learning_rate=0.05, max_depth=40, max_features=1.0, min_samples_leaf=4,n_estimators= 1000, random_state=10, subsample=0.8

Table 3: Confusion Matrix

TP	FP
TN	FN

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

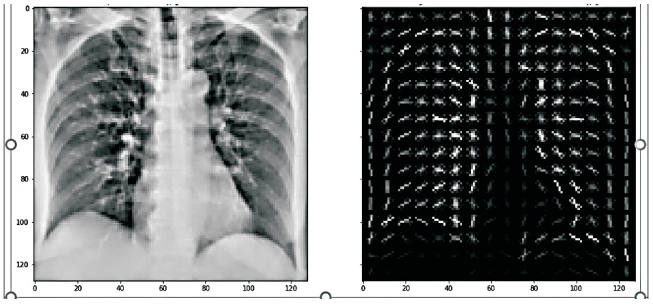
$$\text{Recall (TP Rate)} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{FP Rate)} = \frac{FP}{FP + TN} \quad (12)$$

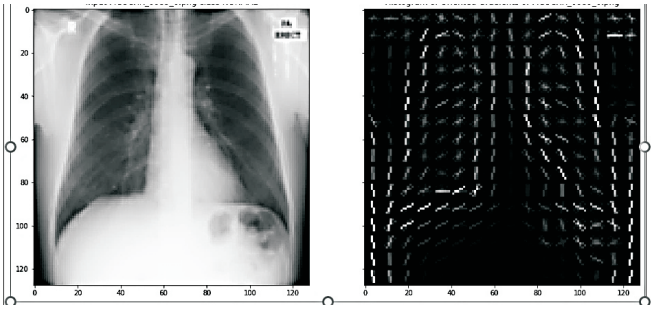
$$\text{F1 - Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

Results and discussion

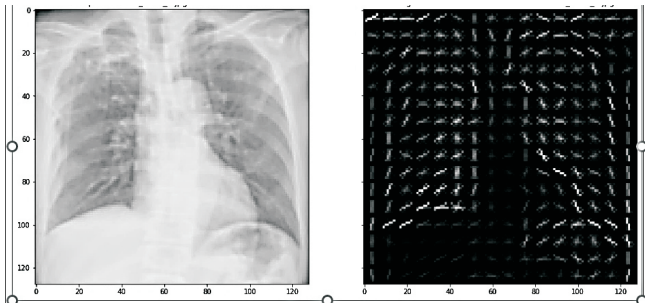
This section presents the results and discuss the discoveries in the study. The experiment was performed with Jupyter notebook with sklearn libraries on Anaconda platform. All experiments were performed on an Intel® core™ i5-7200 CPU @ 2.50GHz to 2.70 GHz Pentium Windows computer with 8GB RAM. The images were manually cropped to remove some unwanted background images and noise. Then, they were resized to 128 x 128, flattened and converted to grey scale before their features were extracted. Firstly, Figure 2 presents a sample CXR image and their corresponding transformed HOG images for the 3 different classes (COVID-19, NORMAL and PTB. The feature vectors was divided into 75: 25 train-test split ratio. As discussed n sections3.2 and 3.3, the results of the performances comparison of the 5 different machine learning models on the extracted and reduced dataset were presented. It is emphasized that PCA with explained variance of 95% captured maximum information in its components, thus achieves the best classification accuracy amongst all. The values for all metrics ranges from between 0 and 1. The closer the value of the metrics to 1, the better the model.



A. COVID-19



B. NORMAL

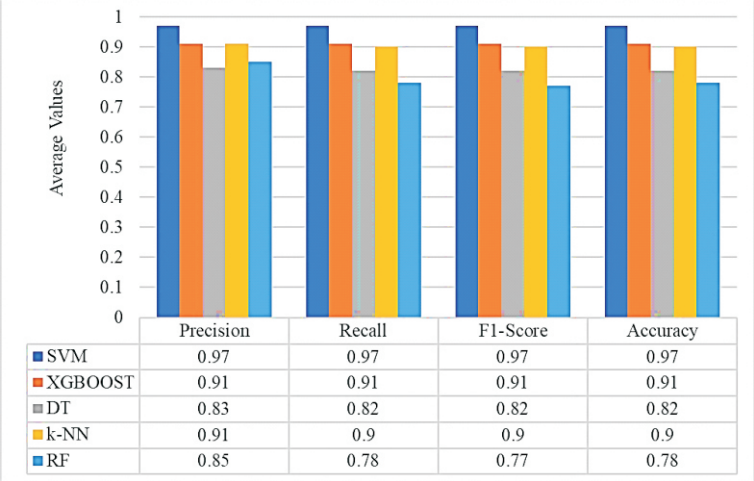


C. PTB

Figure 2: A sample of the original and the corresponding HOG image of the study dataset

Results and Discussion

Figure 3: shows the performance comparison of the 5 different classifiers based on ~~the~~ precision, recall, F1-Score and accuracy. The result is based on the test set which is 25% of the dataset. It is observed that SVM achieved the highest classification report values of 0.97 across all metrics. Its performance value is more consistent than with other models across all metrics. Using F1- score as for comparing all models, RF obtained the least value of 0.77 following closely by DT with the value of 0.82. SVM outperformed all other models with the value of 0.97. RF obtained the least recall rate and accuracy values of 0.78. Therefor making it the least performing model for the image classification task.



	Precision	Recall	F1-Score	Accuracy
SVM	0.97	0.97	0.97	0.97
XGBOOST	0.91	0.91	0.91	0.91
DT	0.83	0.82	0.82	0.82
k-NN	0.91	0.9	0.9	0.9
RF	0.85	0.78	0.77	0.78

Figure 3: Comparison of the model's performance

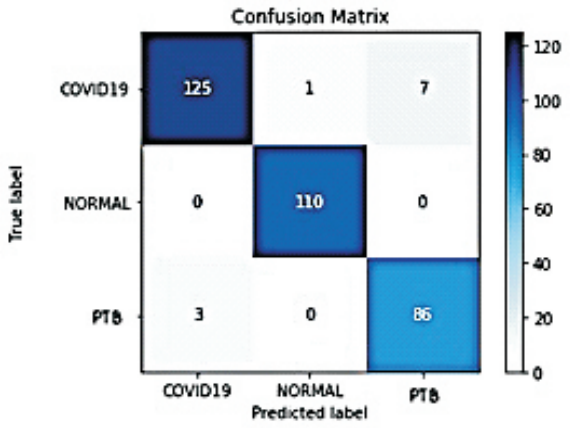
Confusion Matrix

This section further analysis the performances by all models using the confusion matrix for the test set as shown by Figure 3. It is observed that Figure 3a representing the confusion matrix for SVM gave the best recall for each class with minimal error. For SVM, out of 133 instances of COVID-19 disease, 125(94%) instances were correctly classified as COVID-19 disease, 1(1%) instance was incorrectly classified as NORMAL while 7(5%) instances were incorrectly classified as PTB disease. But there was no misclassification for the class NORMAL. The classification is 100%. Also, for the class PTB, 3(3%) instances were misclassified as COVID-19 while 86 instances (97%) were correctly classified as PTB. Comparing all models on COVID-19 prediction, RF performed best with a recall rate of 0.98. Out of 133 instances of COVID-19, 131(98%) were correctly classified as COVID-19, 2(2%) instances were misclassified as PTB. But for DT model, only 104(78%) instances of COVID-19 were correctly classified. 27(20%) instances of COVID-19 were misclassified as PTB.It could be observed that majority of the misclassification was between COVID-19 and PTB. Hence, our objective has been achieved by building a classification model for COVID-19 disease but separate from PTB.

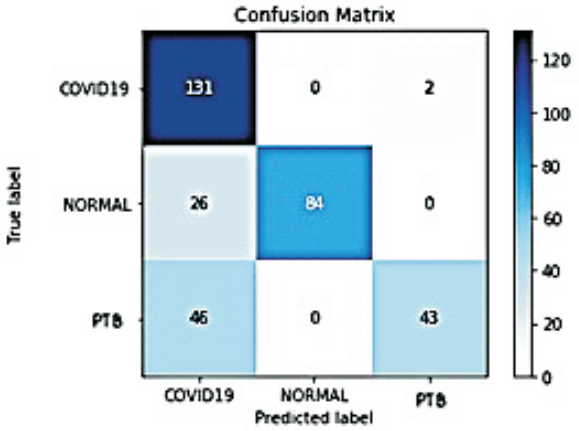
Analyzing the classification for class NORMAL, there was no misclassification for PTB. All model could distinguish between the images of NORMAL and PTB. The greatest misclassification between NORMAL and COVID-19 was from RF model where 26(24%) out of 110 instances were misclassified as COVD-19 instead of NORMAL.

Analysis of the classification of PTB also shows that there was no misclassification with the class NORMAL image. The greatest misclassification was still between COVID-19 and PTB. For RF model, 46 (53%) out of 89 instances were wrongly classified as COVID-19. SVM and *k*-NN models obtained a recall rate 0.94 showing a good detection rate.

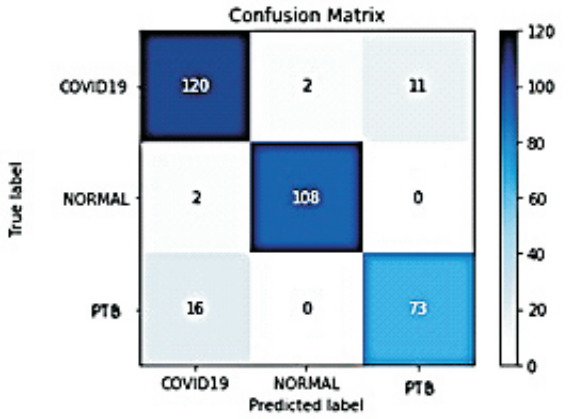
.



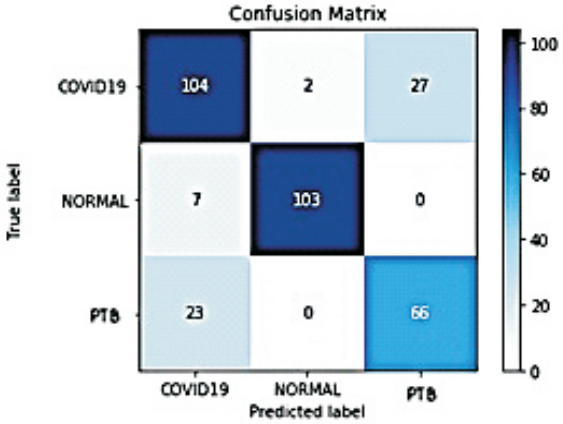
a. Confusion Matrix for SVM



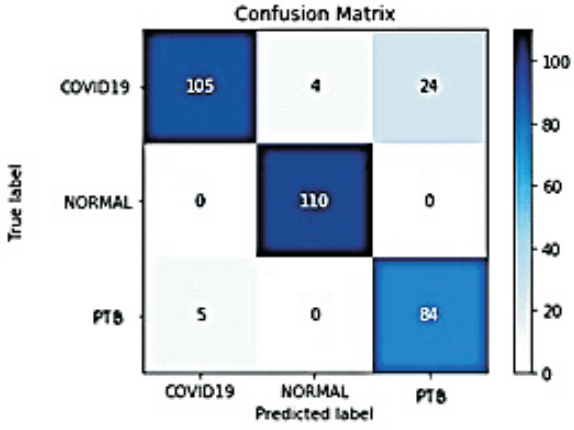
b. Confusion Matrix for RF



c. Confusion Matrix for XGBoost



d. Confusion Matrix for DT



e. Confusion Matrix for *k*-NN

Figure 3: Confusion Matrix for all models

ROC Curves

This segment explains the ROC curves for SVM model which is the trade-off between true positive and false positive rate having established that it performed best of the models as shown by Figure 4. The ROC values for all metrics are close to 1 showing a very good classification performances. The ROC value for class NORMAL is 1.00 as there are no misclassification.

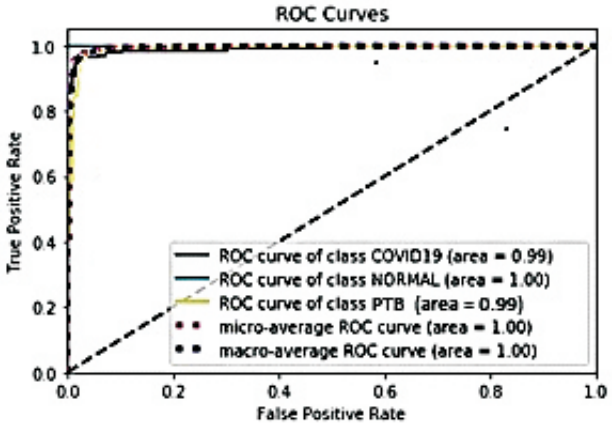


Figure 4: ROC for SVM

Conclusion

This study is aimed at building a classification model for COVID-19 and Tuberculosis diseases. The methodology adopted is a 4-phasedsystem. Firstly, texture features were extracted from the CXR images of the different classes. Then, PCA feature selection scheme was applied to the extracted features to effectively select important features to enhance classification models. Five different models (k-NN,SVM, DT, XG Boost and RF) were selected for the multi-classification task. The results obtained showed that SVM performs better with classification result of 97% across all metrics when compared to the other learning models. Finally, the developed HOG-PCA model showed a good classification results for the extraction of a rich dataset for COVD-19 and PTB diseases. In the future, the study intends to deploy more robust pre-processing and machine learning models on larger CXR and CT scan image datasets for COVID-19 and other related diseases classification.

References

Breiman, L. (2001). Random forests. *Machine Learning,, 45*(1), 5-32. Retrieved from https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

Candemir, S., Jaeger, S., Musco, J., Xue, Z., Karargyris, A., Antani, S. K., . . . Palaniappan, K. (2014). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans Med Imaging*. , 33(2), 577-590. doi:10.1109/TMI.2013.2290491. PMID: 24239990

Chen, T., & Guestrin, C. (2016). Xgboost: a scalable tree boosting system. 22nd *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785-794).

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., & Ghassemi, M. (2020, June 22). COVID-19 Image Data Collection: Prospective Predictions Are the Future. *arXiv:2006.11988v1 [q-bio.QM]*, 25. Retrieved from https://github.com/ieee8023/covid-chestxray-dataset

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)* (pp. 886 – 893). San Diego, Calif, USA, June 2005: IEEE.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology.,* 24(6), 417.

Huang, C., Wang, Y., Li, X., Zhao, J., Hu, Y., Zhang, L., . . . Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet,* 395, 497-506. doi:10.1016/S0140-6736(20)30183-5

Maduskar, P., Muyoyeta, M., Ayles, H., Hogeweg, L., Peters-Bax, L., & van Ginneken, B. (2013). Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. *The International Journal of*

Tuberculosis and Lung Disease, 17(12), 1613–1620. doi:0.5588/ijtld.13.0325

Naserghandi, A., Allameh, S. F., & Saffarpour, R. (2020). All about COVID-19 in brief. *New Microbe and New Infect, 35:100678*. doi:10.1016/j.nmni.2020.100678

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython}. *Journal of Machine Learning Research,* 12, 2825-2830.

Vapnik, V. N. (1998). Adaptive and Learning Systems for Signal Processing Communications, and control. *Statistical learning theory.*

WHO. (2012). *Global tuberculosis report , World Health Organization.* Geneva, Switzerland: WHO/HTM/TB/2012.6.

WHO. (2016). Chest radiography in tuberculosis detection- summary of current WHO recommendations and guidance on programmatic approaches. *Strategy, The End TB.*

WHO. (2020). COVID-19: Considerations for tuberculosis (TB) care. *World Health Organization (WHO) Information Note*, pp. 1-11.

Zheng, Z., Peng, F., Xu, B., Zhao, J., Liu, H., Peng, J., . . . Tang, W. (2020). Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *Journal of Infection*. doi:10.1016/j.jinf.2020.04.021

Zu, Z. Y., Jiang, M. D., Xu, P. P., Chen, W., Ni, Q. Q., Lu, G. M., & Zhang, L. G. (2020). Coronavirus Disease 2019 (COVID-19): A Perspective from China. *Radiology.* doi:10.1148/radiol.2020200490