**Explainable Machine Learning For COVID-19 Detection In Chest X-Rays**

**ABSTRACT**

Starting from late December 2019, COVID-19 (SARS-CoV-2) has been declared a pandemic and a global emergency. The utmost priority is therefore an early detection and hospitalization of infected persons. In detecting the virus, RT-PCR kits are used which takes from hours to days for a diagnosis. This motivated the idea of building automated COVID-19 detector using artificial intelligence. There have been various deep learning based approaches built with high diagnostic accuracy, however, these architectures are not interpretable, i.e. a human cannot consistently predict the model results. Therefore, this study reviews such deep learning based approaches and proposes an interpretable architecture to diagnose COVID-19 by achieving comparable performance to the existing black-box approaches.

Index Terms—COVID-19 detection, Interpretable machine learning, Chest X-Rays, Radiology Level features, Explainability, SHAP.

**Chapter 1**

**INTRODUCTION**

From late December 2019, a novel corona-virus (SARSCoV-2) has spread all around the globe originating from Wuhan district of China [1], [2]. As of April 06, 2021 more than 130 million confirmed cases, and more than 2 million deaths were reported1 worldwide. Due to unavailability or difficult reach for immediate vaccination, early diagnosis is highly critical. It provides the opportunity for immediate isolation of the suspected person and decreases the chance of multiplying infection to healthy population. Reverse transcription polymerase chain reaction (RT-PCR) is used as main diagnosing method for COVID-19 [3], though it can be considered as a time-consuming test, as it takes typically hours or days to get the results and also, it suffers from false negative cases [4]. Chest radiography imaging (X-ray or computed tomography (CT)) is used as a routine tool for pneumonia diagnosis and is easy to perform with fast diagnosis [5], [7]. Chest CT has a high sensitivity for diagnosis of COVID-19 and X-ray images show visual indexes correlated with COVID-19 [6], [8].

The rapid use of chest X-rays (CXRs) were used by the radiology departments in Italy and the U.K. to sort non-COVID19 patients with pneumonia to allocate hospital resources efficiently [9]. However, there exists similarity among chest radiography images of COVID-19 and pneumonia caused by other viral infections such as common flu (Influenza-A). This similarity makes a differential diagnosis of COVID-19 cases by expert radiologists challenging [10], [11]. An explainable automated algorithm for classification of COVID-19 on CXR images can speed up the triage process of COVID-19 case detection and maximize the allocation of hospital resources.

Considering the huge rate of infected people and limited number of training kits and trained radiologists, machine learning methods for identification of such subtle abnormalities contribute to an automated, objective diagnosis and increase the rate of early diagnosis with high accuracy. Machine learning based solutions could be potentially powerful tools for solving such problems. Such an approach and initiative has already been shown by researchers especially using deep learning based models more specifically using convolutional neural networks (CNNs). These architectures have been shown to outperform the classical AI approaches in most of computer vision and and medical image analysis tasks in recent years, but this approach is considered black-box due to complexity and inability to explain its decisions [12]–[14]. It is tough to analyse for medical expert or any individual, why a system responded in the manner it did and raises the question of interpretability of the tool. Explainability and reliability are the crucial factors in medical visual analytics. Therefore, we hypothesized that CXR images of COVID-19 patients can be distinguished from other forms of pneumonia using an interpretable machine learning based classifiers using radiologicallevel feature. We aimed to achieve similar or better performance compared to the existing deep learning Networks along with explaination.

II. BACKGROUND The need to expedite the process of diagnosis or developing the diagnostic tool is greater than ever for COVID-19 patients. Typically, the results from RT-PCR kits take up-to approx. 6- 8 hours to diagnose a patient being COVID-19 positive [14]. This motivates researchers to use Chest Radiography Imaging especially, Chest X-Rays for diagnosis, as the Chest XRays are non-invasive tool to monitor progression of disease. Although, Chest CT Scan are considered high quality imaging but our experiment deals with Chest X-Rays due to its high public availability.

1. Related Work There are numerous experiments and studies built in order to apply machine learning and deep learning to assist diagnosis process of COVID-19. Most of the studies are deep learning based architectures. According to [16], there exists numerous studies which uses Statistical based feature extraction for COVID-19 detection. Though, these experiments also achieve comparable performance to the deep learning architecture, but these architectures cannot be interpreted easily. especially for any medical individual and/or radiologist. Also, study from [15]–[20] suggest that clinical and radiological-level features are crucial in identifying COVID-19 from Chest X-Rays. Among the deep learning architectures used, Hemdan et al. [25] used deep learning models to diagnose COVID-19 in Xray images and proposed a COVIDX-Net model comprising seven CNN models. Wang and Wong [24] proposed a deep model for COVID19 detection (COVID-Net), which obtained 92.4% accuracy in classifying normal, non-COVID pneumonia, and COVID-19 classes. Ioannis et al. [26] developed the deep learning model using 224 confirmed COVID-19 images. Their model achieved 98.75% and 93.48 % success rates for two and three classes, respectively. Narin et al. [22] achieved a 98% COVID-19 detection accuracy using chest X-ray images coupled with the ResNet50 model. Sethy and Behera [23] classified the features obtained from various convolutional neural network (CNN) models with support vector machine (SVM) classifier using X-ray images. Their study states that the ResNet50 model with SVM classifier provided the best performance. Finally, there are also several recent studies on COVID-19 detection that employed various deep learning models with CT images [27]–[32]. In this study, an ensemble framework is proposed. This framework works with CheXNet architecture as feature extractor detecting 14 radiological features from raw Chest XRay images. These features are forwarded to the interpretable model which in turn diagnosis with feature-importance based explanation.

III. PROTOTYPE DESIGN The project answers the following research questions:

A. Research Questions :

1) Does the simpler machine learning models achieves results as good as deep learning based classifiers while being better explainable with regards to input features ?

2) Are the SHAP features extracted from machine learning models are interpretable by Radiologist for COVID-19 detection ?

B. Selected Classifiers We select five classifiers namely; k-Nearest Neighbours, Support Vector Machine with Radial Basis Function kernel, Linear Classifier, Random Forest and Decision Trees. Out of these five, we chose kNN, SVM, Linear Classifier and Decision Trees as interpretable simpler classical machine learning models and we considered Random Forest as NonInterpreable classifiers or black-box models. We used blackbox model Random Forest as our internal baseline for a global perspective of the ground truth and compared the performance with the interpretable models or white-box models.

1) k-Nearest Neighbour: KNN is an algorithm that is considered both non-parametric and an example of lazy learning. kNN is a case-based learning method, which keeps all the training data for classification. However, to apply kNN we need to choose an appropriate value for k, and the success of classification is very much dependent on this value. Among the differnt ways of choosing the k value, here a simple run of the algorithm many times with different k values is performed and the one with the best performance is chosen.

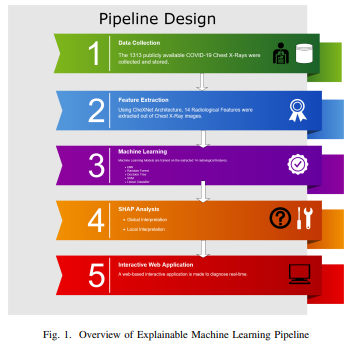
2) SVM with RBF-kernel: A support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. RBF kernel is a function whose value depends on the distance from the origin or from some point. When training an SVM with the Radial Basis Function (RBF) kernel, two parameters must be considered: C and gamma. The parameter C, common to all SVM kernels, trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly. Gamma defines how much influence a single training example has. The larger gamma is, the closer other examples must be to be affected.

3) Linear Classification: Another machine learning classification algorithm that is used to predict the probability of a categorical dependent variable. In linear classification, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). Binary linear classification requires the dependent variable to be binary. For a binary classification, the factor level 1 of the dependent variable should represent the desired outcome. Only the meaningful variables should be included. The independent variables should be independent of each other. That is, the model should have little or no multicollinearity. The independent variables are linearly related to the log odds. Logistic regression requires quite large sample sizes.

4) Decision Trees: Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

5) Random Forest: RF algorithm is one of the best algorithms for classification. RF is able for classifying large data with accuracy. It is a learning method in which number of decision trees are constructed at the time of training and outputs of the modal predicted by the individual trees. RF act as a tree predictors where every tree depends on the random vector values. The basic concept behind this is that a group of “weak learners” may come together to build a “strong learner”. RF classifier is an ensemble method that trains several decision trees in parallel with bootstrapping followed by aggregation, jointly referred as bagging. Bootstrapping indicates that several individual decision trees are trained in parallel on various subsets of the training dataset using different subsets of available features. Bootstrapping ensures that each individual decision tree in the random forest is unique, which reduces the overall variance of the RF classifier. For the final decision, RF classifier aggregates the decisions of individual trees; consequently, RF classifier exhibits good generalization. RF classifier tends to outperform most other classification methods in terms of accuracy without issues of over-fitting.

C. Explainable Machine Learning Pipeline The state-of-the-art image classification algorithms states that image is classified based upon the feature learned by the model against the class. The explanations of the deep learning model with CXR images of COVID-19 positive patients could reveal the portion from the image that highly influence the prediction task. However, this visual explanation could not explain biological features responsible for the prediction.



Therefore Fig. 1 demonstrates the overview of the pipeline. For each image in the dataset, the feature extractor generates 14 radiological features i.e. Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia. Upon these 14 features, the interpretable classifiers predicts the outcome and is evaluated based on Accuraccy, AUC and F1-score. Further analysis on the classifiers and explaination is done using SHAP. The explaination of the classifiers are given Local and Global perspective for robustness of the explaination.

D. Dataset Generation ChexNet is a deep learning based state-of-the-art pneumonia detection algorithm. It detects and localizes 14 radiological features from given chest X-ray images. As per [21], a 121-layer densely connected convolutional neural network is trained on ChestX-ray14 dataset, which contains 112,120 frontal view X-ray images from 30,805 unique patients. The result surpasses the performance of practicing radiologists. We used a pre-trained CheXNet model for generating the 14 radiological features for each image in the dataset with label as covid positive or negative.

E. Evaluation Metrics Choosing an appropriate evaluation metric is a challenge in machine learning, but is particularly difficult for imbalanced classification problems. As most of the standard metrics that are widely used assume a balanced class distribution, and because typically not all classes, and therefore, not all prediction errors, are equal for imbalanced classification. Therefore, we tried to evaluate based upon the correct classification given by model. We choose Accuracy, Area Under the ROC curve (AUC) and F1-score for evaluating the performance of each classifiers. Accuracy: It is the ratio of number of correct predictions to the total number of input samples. Area under the ROC curve: AUC is a diagnostic plot for summarizing the behavior of a model by calculating the false positive rate and true positive rate for a set of predictions by the model under different thresholds. F1-score: F1-score is the harmonic mean between precision and recall. The range for F1-score is [0, 1]. Further, we considered a benchmark dataset to evaluate performance of the models for data coming from different distribution. The database was developed by [33] using images from various sources. The database is constantly upgraded. As of now, the content comprises of around 201 COVID-19 positive X-Ray images. There were no images, which represent X-Ray of normal lungs, hence we have used thousand healthy X-Ray images from our original database as negative class, which we have used to feed the classifiers

F. SHAP (SHapley Additive exPlanations) The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. The SHAP explanation method computes Shapley values from coalitional game theory. The feature values of a data instance act as players in a coalition. A player could be an individual feature value. A player can also be a group of feature values. One innovation that SHAP brings to the table is that the Shapley value explanation is represented as an additive feature attribution method, a linear model.

IV. IMPLEMENTATION A. Ground Truth The ground truth consists of 313 Positive COVID CXR and 1000 Negative CXR collected from four different sources to make our version of the dataset to work upon. This includes COVIDx dataset from [33], Kaggle CXR Pneumonia dataset by Paul Mooney [34], CXR images of adult subjects from the RSNA Pneumonia Detection Challenge [35], original and augmented versions of COVID-19 examples from [36]. We split the data set in ratio of 70:30 and trained the models with 920 data points and 393 test points. According to [37]–[41] CT-Scan data would be goldstandard for us and also portray satisfying results when evaluated in terms of Accuracy and F1-Score. However, due to CT Scan being available in very less quantity publicly, we would like to use Chest X-rays as our dataset. Though, it won’t be that competible in terms of quality in regards with CT-Scans but [42] suggests CXR to be sufficient and comparable to CTScans in order to diagnose COVID-19 patients. Real world datasets are usually imbalanced. The ground truth here are nearly three times more negative cases than that of positive. The classification algorithms in this case tends to favor the majority class. The distribution of the classes in the dataset in reality refers to the actual class distribution of the COVID infected cases.

Model Train Settings and Hyper Parameter

1) kNN: a) Train Test Data Split: During training time, in ratio of 70:30, the data is split for train and test sets consecutively. Class proportion is also maintained in both the sets with respect to original set. b) Data Preprocessing: Standard scaling and centering of data is done. c) Train Setting: For purpose of training, in order to tackle class imbalance problem on the train set, 10 fold cross validation is done with 3 repeats. d) Choice of K: At train time, for evaluation of K, accuracy was used to select the optimal model using the largest value. In our case, k=9 gave us the best accuracy. Fig 3 depicts the number of neighbour vs accuracy plot.

2) SVM: Support Vector Machine with Radial Basis Function(RBF) along with a) Train Test Data Split: During training time, in ratio of 70:30, the data is split for train and test sets consecutively. Class proportion is also maintained in both the sets with respect to original set. b) Train Setting: For purpose of training, 5-fold cross validation is used. c) Choice Of Kernel: Radial Basis Function is used as kernel. d) HyperParameters: The cost = 1 and gamma = 1 following a similar approach to [49].

3) Linear Classification: a) Train Test Data Split: During training time, in ratio of 70:30, the data is split for train and test sets consecutively. Class proportion is also maintained in both the sets with respect to original set. b) Train Setting: For purpose of training, in order to tackle class imbalance problem on the train set, 10 fold cross validation is done with 3 repeats.

4) Decision Tree: a) Train Test Data Split: During training time, in ratio of 70:30, the data is split for train and test sets consecutively without replacement. b) Other Settings: Using ’gini’ function to measure the quality of a split with minimum 2-samples required to split an internal node 5) Random Forest: a) Train Test Data Split: During training time, in ratio of 75:25, the data is split for train and test sets consecutively without replacement. b) Other Settings: Using ’gini’ function to measure the quality of a split with minimum 2-samples required to split an internal node. The forest consists of 100 trees.

REFERENCES [1] World Health Organization. Director-General’s opening remarks at the media briefing on COVID-19 - March 11, 2020. https://www. who.int/dg/speeches/detail/who-director-general-s-opening-remarks-atthe-media-briefing-on-covid-19—11-march-2020. [2] Wang, Chen, et al. ”A novel coronavirus outbreak of global health concern.” The lancet 395.10223 (2020): 470-473. [3] Cohen, Jon. ”Wuhan seafood market may not be source of novel virus spreading globally.” Science 10 (2020). [4] Ai, Tao, et al. ”Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases.” Radiology 296.2 (2020): E32-E40. [5] Long, Chunqin, et al. ”Diagnosis of the Coronavirus disease (COVID19): rRT-PCR or CT?.” European journal of radiology 126 (2020): 108961. [6] Zu, Zi Yue, et al. ”Coronavirus disease 2019 (COVID-19): a perspective from China.” Radiology 296.2 (2020): E15-E25. [7] Lee, Elaine YP, Ming-Yen Ng, and Pek-Lan Khong. ”COVID-19 pneumonia: what has CT taught us?.” The Lancet Infectious Diseases 20.4 (2020): 384-385. [8] Rubin, Geoffrey D., et al. ”The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society.” Chest 158.1 (2020): 106-116. [9] Castiglioni, Isabella, et al. ”Artificial intelligence applied on chest X-ray can aid in the diagnosis of COVID-19 infection: a first experience from Lombardy, Italy.” medRxiv (2020). [10] Neuman, Mark I., et al. ”Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children.” Journal of hospital medicine 7.4 (2012): 294-298. [11] Ng, Ming-Yen, et al. ”Imaging profile of the COVID-19 infection: radiologic findings and literature review.” Radiology: Cardiothoracic Imaging 2.1 (2020): e200034. [12] Ahsan, Md Manjurul, et al. ”Study of different deep learning approach with explainable ai for screening patients with COVID-19 symptoms: Using ct scan and chest x-ray image dataset.” arXiv preprint arXiv:2007.12525 (2020). [13] Akl, Elie A., et al. ”Use of chest imaging in the diagnosis and management of COVID-19: a WHO rapid advice guide.” Radiology 298.2 (2021): E63-E69. [14] World Health Organization. (2020). Use of chest imaging in COVID19: a rapid advice guide, 11 June 2020. World Health Organization. https://apps.who.int/iris/handle/10665/332336. Licencia: CC BY-NC-SA 3.0 IGO [15] Won, Joungha, et al. ”Development of a Laboratory-safe and Low-cost Detection Protocol for SARS-CoV-2 of the Coronavirus Disease 2019 (COVID-19).” Experimental neurobiology 29.2 (2020): 107. [16] Zhang, Zhenwei, and Ervin Sejdic. ”Radiological images and machine ´ learning: trends, perspectives, and prospects.” Computers in biology and medicine 108 (2019): 354-370. [17] Sanchez-Oro, Raquel, Julio Torres Nuez, and Gloria Mart ´ ´ınez-Sanz. ”Radiological findings for diagnosis of SARS-CoV-2 pneumonia (COVID-19).” Medicina clinica (English ed.) (2020). [18] Zhang, Ran, et al. ”Diagnosis of coronavirus disease 2019 pneumonia by using chest radiography: Value of artificial intelligence.” Radiology 298.2 (2021): E88-E97.