Explainable Machine Learning for COVID-19 Detection in Chest X-Rays

Subhajit Mondal, Subhankar Patra
M.Sc. Data and Knowledge Engineering
Matriculation Number: {229590},{229798}
Course: Scientific Team Project
University of Magdeburg, Germany
{subhajit.mondal},{subhankar.patra}@st.ovgu.de

Jalaj Vora, Shivam Singh
M.Sc. Digital Engineering
Matriculation Number: {221510},{229819}
Course: Interdisciplinary Team Project
University of Magdeburg, Germany
{jalaj.vora},{shivam.singh}@st.ovgu.de

***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**

I. Introduction

From late December 2019, a novel corona-virus (SARS-CoV-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 reported[1] 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-COVID-19 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 radiological-level feature. We aimed to achieve similar or better performance compared to the existing deep learning Networks along with explaination.

This paper further discusses the motivation, ground truth and review of existing deep learning architectures in Section II. Section III proposes a machine-learning based architecture. Section IV describes implementation details where as Section V discusses the explainibility aspects of the models. Section VI evaluates the results and we discuss the results in Section VII. In Section IX we draw conclusions from the study.

[1] (https://www.worldometers.info/coronavirus/)

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 X-Rays 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.

A. 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 X-ray 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 X-Ray 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 Non-Interpreable classifiers or black-box models. We used black-box 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 Accuracy, AUC and F_1-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.

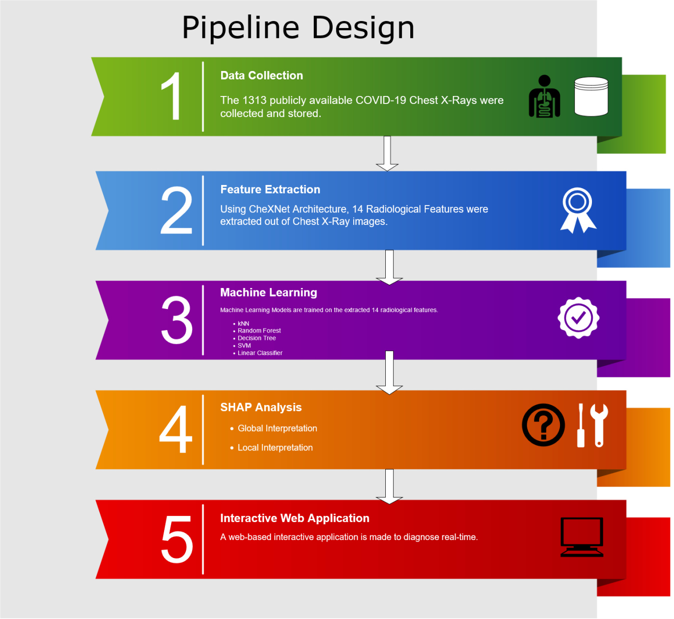

Pipeline Design
1
Data Collection
The 1313 publicly available COVID-19 Chest X-Rays were collected and stored.
2
Feature Extraction
Using CheXNet Architecture, 14 Radiological Features were extracted out of Chest X-Ray images.
3
Machine Learning
Machine Learning Models are trained on the extracted 14 radiological features.
• kNN
• Random Forest
• Decision Tree
• SVM
• Linear Classifier
4
SHAP Analysis
• Global Interpretation
• Local Interpretation
5
Interactive Web Application
A web-based interactive application is made to diagnose real-time.

Fig. 1. Overview of Explainable Machine Learning Pipeline

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 F_1-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.

F_1-score: F_1-score is the harmonic mean between precision and recall. The range for F_1-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. That view connects LIME and Shapley Values. SHAP specifies the explanation as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

here g is the explanation model, $z' \in \{0,1\}^M$ is the coalition vector, M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature attribution for a feature j, the Shapley values. What we call ”coalition vector” is called ”simplified features” in [50].

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 gold-standard for us and also portray satisfying results when evaluated in terms of Accuracy and F_1-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 competitive in terms of quality in regards with CT-Scans but [42] suggests CXR to be sufficient and comparable to CT-Scans 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.

Here, Figure 2 shows the all pairwise feature-feature correlation among 14 features of the data-set with Spearman Correlation as the measure. The figure indicates a high correlational density among Atelectasis, Cardiomegaly, Effusion and second high correlational density among Emphysema, Fibrosis and Pleural Thickening. The figure also suggests Consolidation, Pneumonia and Infiltration with more number of correlation with other features.

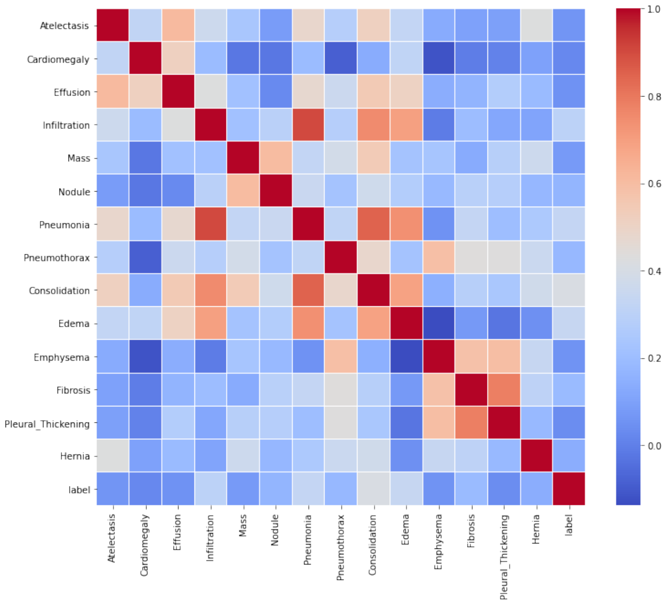


Fig. 2. Feature-Feature correlation Heatmap

B. 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].

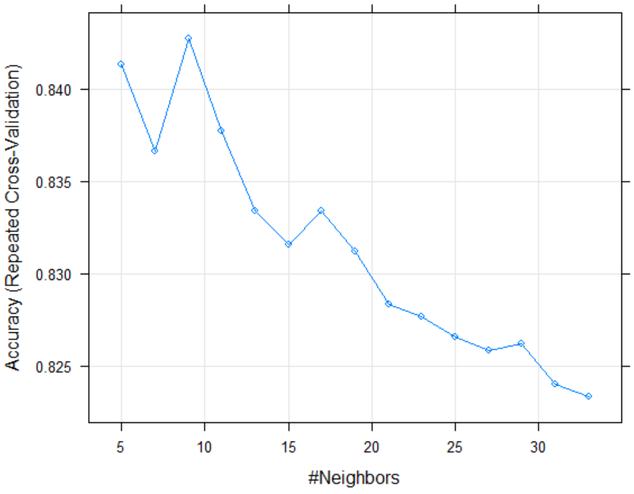


Fig. 3. Choice of k vs. Accuracy Plot for Knn

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.

The learned model weight parameters when visualised as importance plot looks like Fig. 4.

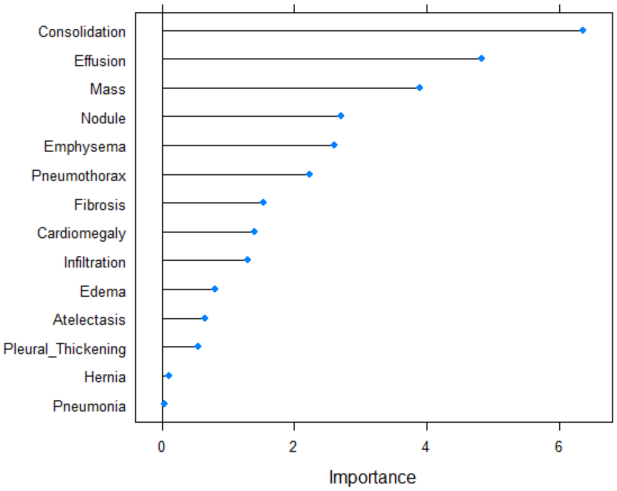


Fig. 4. Linear classifier Weights

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.

V. Explainability

There exists no specific mathematical definition of Explainability [48]. As per [46], Explainability could be defined in non-mathematical form as a degree to which humans can understand and interpret the reason of a decision made. The higher the degree, higher it is easier for any human being to comprehend grounds of the decision.

The need for interpretability also arises from incomplete formulation of the problem statement as in certain cases, prediction is not enough [47]. Rather a model needs to explain how it came to a conclusion and why such a decision was made. The importance of explanation of a model depends on various factors and scenarios. Specifically in this case, it is important to know the significance of model on the final outcome. It is not adequate for a model to detect whether a patient is COVID-19 positive or negative. The questions how and why was such a behaviour noticed are consequential as well. If these questions are not asked, it can have serious consequences on the patient as well as on the radiologist. The issue also moreover lies in the use of single evaluation metric. Although, using multiple evaluation metrics might give a higher confidence on the working of a specific model but it doesn't give exploration of a model.

The experiment chooses SHAP as two way interpretation; namely Global Interpretation and Local Interpretation. Global Interpretation suggests global perspective of the explaination, i.e., it answers the question how would black-box model perform on the ground truth and how would does surrogate model suggest about the behaviour of the black-box model. Similarly, Local Interpretation means local perspective of the explaination, i.e., at every instance of a Chest X-Ray image, an individual prediction is been explained using SHAP. .

a) Global Interpretation: For Global Interpretation, we took Random Forest as our internal baseline black-box model. This black-box model was trained and tested on the ground truth. The reason for this is to understand average prediction of the ground truth. To explain the black-box model, a Global Surrogate Model (white-box) model, i.e., a simpler interpretable model decision trees was used. Surrogate Model replicates the behaviour of black-box model and similar to black-box, it was trained and tested on the ground truth. To evaluate the similarity of the surrogate model w.r.t the black-box model, R^2 value was used a the measure. The higher the R^2 value, the higher it replicates the behaviour of the black-box model.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_*^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{\hat{y}})^2}$$

where $\hat{y}_*^{(i)}$ is the prediction for the i$^{\text{th}}$ instance of the surrogate model, $\hat{y}^{(i)}$ the prediction of the black box model and $\bar{\hat{y}}$ the mean of the black box model predictions. SSE stands for sum of squares error and SST for sum of squares total. The R-squared measure can be interpreted as the percentage of variance that is captured by the surrogate model. If R-squared is close to 1 (= low SSE), then the interpretable model approximates the behavior of the black box model very well. If the interpretable model is very close, you might want to replace the complex model with the interpretable model. If the R-squared is close to 0 (= high SSE), then the interpretable model fails to explain the black box model.

b) Local Interpretation: In Local Approach, we try to explain each image in real-time using SHAP plots. These plots consists of feature dependence, so on. All of these plots are generated from the respective intrinsically interpretable classifiers such as Logistic Regression, Naive Bayes, Decision Trees and k-nearest neighbors.

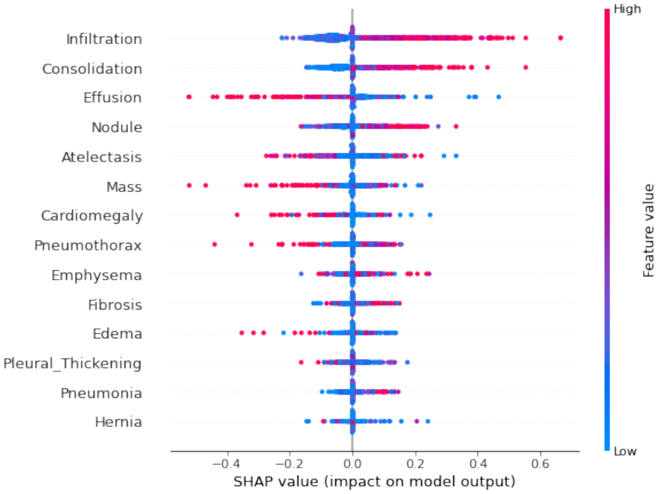


Fig. 5. Patient-individual SHAP value

From the attributions of the 14 features as shown in Fig. 5, The Infiltration was identified as most important, with an average absolute SHAP value magnitude (change in log odds) of 0.329. The Consolidation and Effusion found as second and third most relevant. Fig. 6 depicts the patient-individual SHAP values for each feature as points where color represents the actual feature value. The high attribution of Infiltration is emphasized by the wide spread in the value distribution. For this feature, high feature values correspond to an increased probability of specific feature contributing more towards COVID-19 positive detection. However, this trend is not monotone, since small values (light orange) are associated with a SHAP value just slightly below 0.

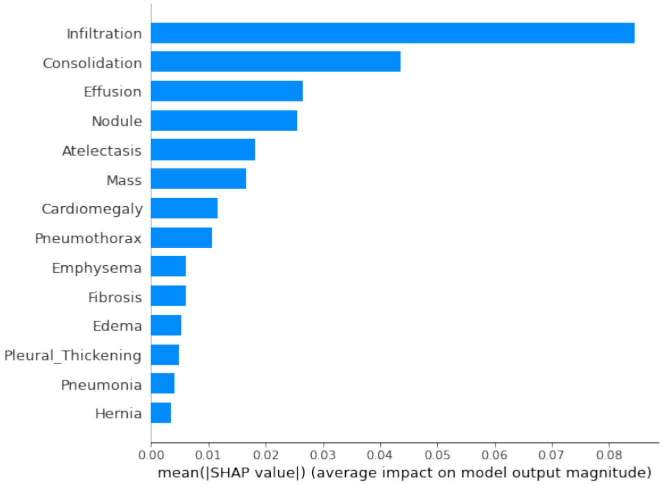


Fig. 6. Mean SHAP value

c) Clustering Approach: With the goal of finding groups of instances that are similar in nature, clustering of shapley values was employed. Among the 14 features sorted by importance, as shown in 6, the top six features with highest average shapley value magnitude and a combined rest(hence 7) was selected.

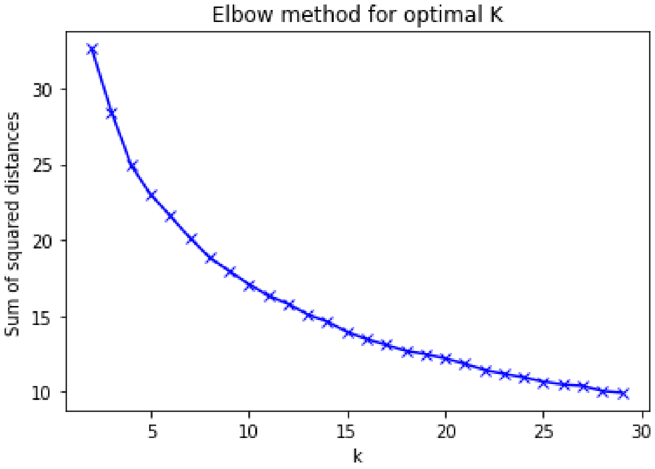
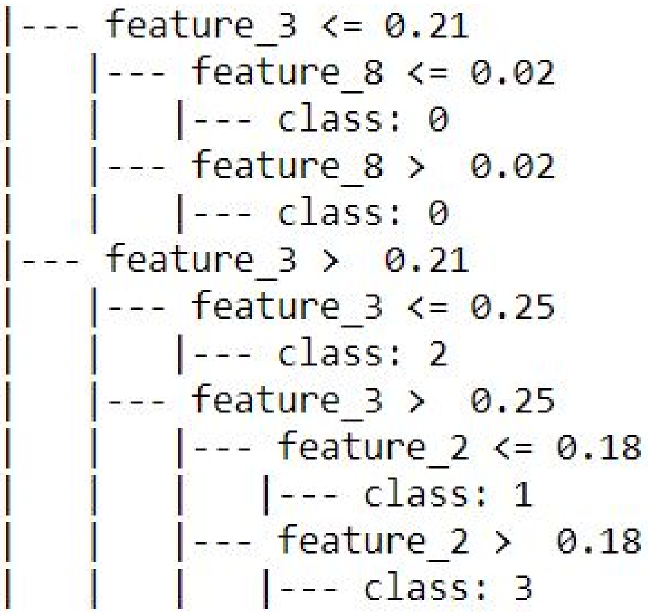


Fig. 7. Elbow Plot to Find Optimal k for k-Means

kMeans clustering was applied, while k was set to 4. The choice of k was empercically chosen by using the elbow method(as shown in 7). Hence, four clusters were identified with similar explanation similarity [43].

In order to interpret the clusters, decision tree classifier was used to find a set of production rules over the original features that were achieved with the image, but now using the cluster labels. In order to avoid overfitting, minimum sample leaf size is set to 50 and pruning factor(ccp_alpha is set to 0.01.

The original order of the features were as follows: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule,



```
|--- feature_3 <= 0.21
|   |--- feature_8 <= 0.02
|   |   |--- class: 0
|   |--- feature_8 >  0.02
|   |   |--- class: 0
|--- feature_3 >  0.21
|   |--- feature_3 <= 0.25
|   |   |--- class: 2
|   |--- feature_3 >  0.25
|   |   |--- feature_2 <= 0.18
|   |   |   |--- class: 1
|   |   |--- feature_2 >  0.18
|   |   |   |--- class: 3
```

Fig. 8. Decision Tree From Shapley Value Cluster

Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural_Thickening, Hernia. The decision tree encodes this data as feature_0 which means Atelectasis and so on.

Fig 8 shows the structure of the tree. We can see the only the most important features as indicated by shapley values as is shown in fig. 6 are used by the tree to classify points. Hence, the underlying clusters identified the most important features to identify similarity.

VI. Results

For the purpose of this study, we have used 2 datasets, one for our study, and the second one for to compare with other results as a benchmark dataset (more about benchmark dataset on section VIII). Here we present the results obtained for both the data set.

1) On our Dataset: As we mentioned earlier, this dataset contains, 313 COVID-19 positive images, where as 1000 COVID-19 negative images. The dataset is imbalanced.

Fig 10 compares the accuracy of our models. The RoC curves for the models are given in fig. 9. Fig 11 shows the F1 scores for the models with our dataset.

The information in the plots are summarized in Table I.

Measures		Models				
		kNN	SVM	LR	DT	RF
Original dataset	Accuracy(%)	80.62	83.98	80.60	79.33	85.41
	AuC(%)	65.1	65.1	65.1	66.7	62.14
	F_1-score(%)	87.91	88.17	88.17	89.91	91.6

TABLE I
Evaluation of the Models on Ground Truth

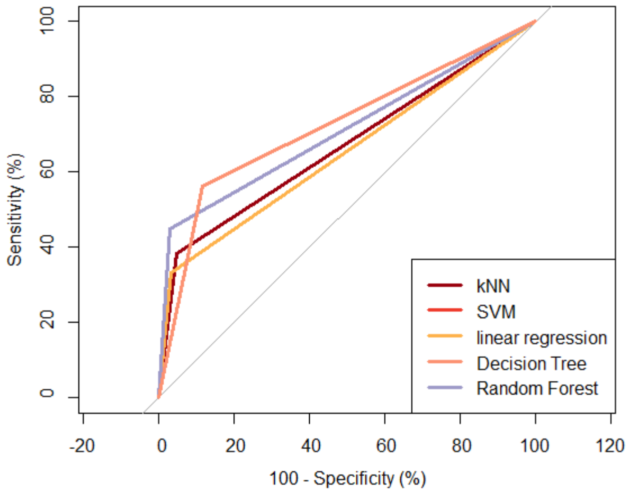


Fig. 9. AUC for Models (SVM and Linear Regression Overlapped)

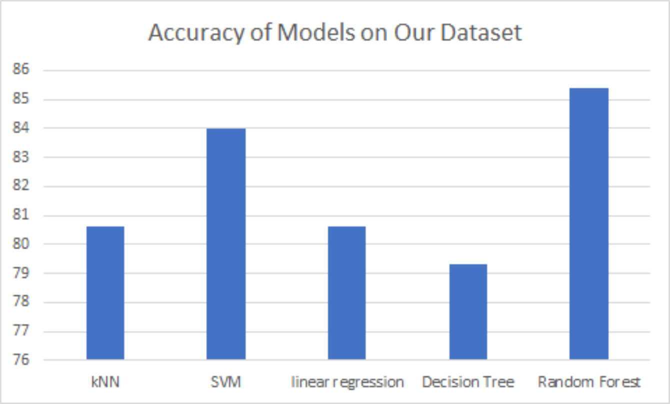


Fig. 10. Accuracy of the Models On Our Dataset

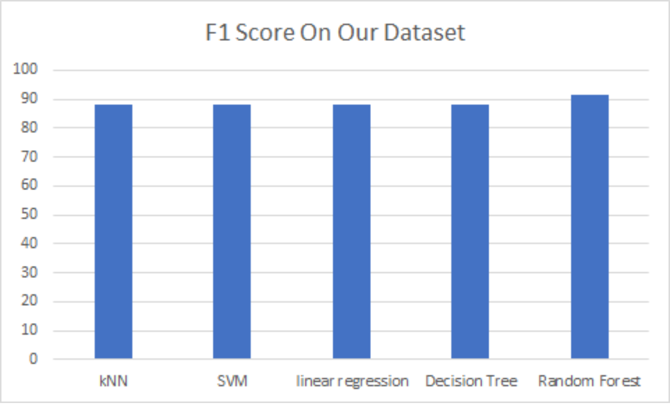


Fig. 11. F_1-score of the Models On Our Dataset

2) On Benchmark Dataset: This dataset contains, 201 COVID-19 positive images, where as 999 COVID-19 negative images. The dataset is imbalanced. More information on the dataset can be found in VIII section.

Fig. 14 compares the accuracy of our models on the benchmark dataset. The AUC curves for the models are given in Fig. 12. Fig. 13 shows the F_1-scores for the models with our dataset.

The information in the plots are summarized in Table II.

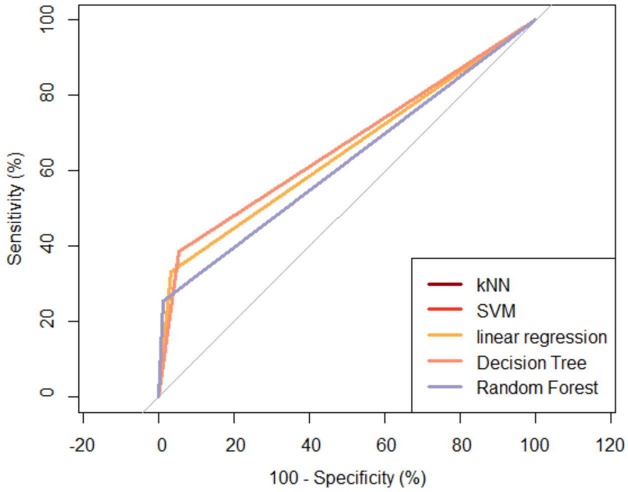


Fig. 12. AUC Plots for classifiers (SVM and Linear Regression Overlapped)

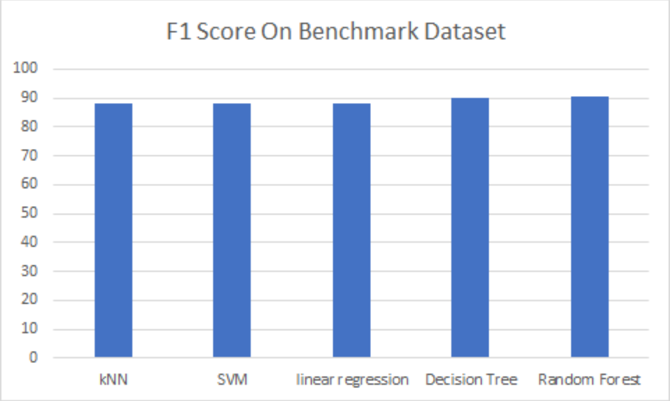


Fig. 13. F_1-score of classifiers on ground truth

Measures		Models				
		kNN	SVM	LR	DT	RF
Benchmark dataset	Accuracy(%)	85.19	85.75	83.76	81.67	84.67
	AuC(%)	66.76	65.1	65.1	71	72.35
	F_1-score(%)	88.15	88.15	88.15	89.85	90.52

TABLE II
EVALUATION OF THE MODELS ON BENCHMARK DATA-SET

VII. DISCUSSION

The study found that simpler ensemble models like random forest or kNN could not out perform the state of the art deep learning based models, but has quite high accuracy given its simplistic and interpretable nature.

This study finds through shapley value that infiltration is the most important feature for detection of COVID-19. This result is line with [44], where the authors claim that, ground glass infiltration is the typical appearance of COVID-19. Others features like consolidation(that is the second most important feature from our study) appears gradually at later stage of the disease.

Also, the study shows that, when shapley values are subject to clustering, where the aim is to find instances with explanation similarity, the important features plays a crucial role in the formation of the cluster, as we saw using the rules generated by decision tree.

VIII. BENCHMARK EVALUATION

In [45], the authors presented a study of the state of the art models for COVID-19 detection using chest X-ray images. To compare the results, the authors used accuracy as the metric, and hence so did we. Table III presents our results compared to results from [45] where as fig 14 reports the accuracy between the models we used. The details of the data set is as follows.

The database of images contains specimens of COVID-19 images from different angles. We used frontal chest X-ray images, in accordance to [45]. The database contains 201 COVID-19 positive X-ray images as of now, while the database is still being upgraded. However, there are no images that represent normal lungs X-ray, hence we have used the images from our original database that represent normal lung X-ray.

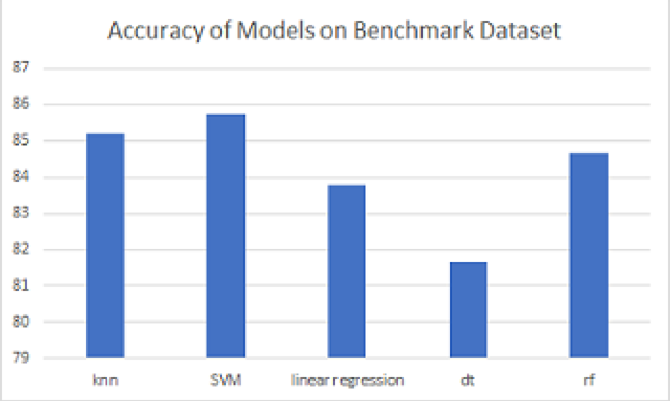


Fig. 14. Accuracy of classifiers on Benchmark Dataset

Study	**No. Of Cases**	**Method Used and Settings**	**Accuracy**
Sethy and Behra	25 COVID-19(+) 25 COVID-19 ()	ResNet50+ SVM	95.38
Hemdan et al.	25 COVID-19(+) 25 Normal	COVIDX-Net	90
Narin et al.	50 COVID-19(+) 50 COVID-19 ()	Deep CNN ResNet-50	98
Ying et al.	777 COVID-19(+) 708 Healthy	DRE-Net	86
Wang et al.	195 COVID-19(+) 258 COVID-19()	M-Inception	82.9
Zheng et al.	313 COVID-19(+) 229 COVID-19()	UNet+ 3D Deep Network	90.8
Our Study	999 COVID-19(-) 201 COVID-19(+)	Decision Tree	81.67
Our Study	999 COVID-19(-) 201 COVID-19(+)	Random Forest	84.67
Our Study	999 COVID-19(-) 201 COVID-19(+)	kNN	85.19
Our Study	999 COVID-19(-) 201 COVID-19(+)	Linear Regression	83.76
Our Study	999 COVID-19(-) 201 COVID-19(+)	SVM	85.75

TABLE III
COMPARISON OF DIFFERENT MODELS

IX. CONCLUSION

In this study, we aimed to build a system which can be a visual aid to the medical personal dealing with diagnosis of COVID-19. We hypothesize to achieve a comparable performance to the deep learning based architecture and our internal baseline model by using simpler and more interpretable machine learning models. We conducted the experiment where we also came up with the explanation of the classification using SHAP graphs. The graphs explains the decision based on features used by the model stating values to be higher the greater.

The section VI shows that the experiment didn't out perform the baseline MLP but the models were more explicit and interpretable. Therefore, we would like to study more introspection to the models to better tune them to achieve higher performance than the internal baseline MLP and the related deep learning based architectures.

Future work can be done by fine tuning the ensemble framework and models and also by comparing with existing deep learning architectures. Also, a user study can be conducted with the help of radiologists to evaluate more objectively.

A prototype application is created as a part of the project, that can be used to classify a chest X-ray image as COVID-19 positive or negative. Among the models chosen, kNN and SVM had the highest accuracy and hence the system gives choice between kNN and SVM model for classifying an image.

The application provides three plots. The first plot describes the extracted features. The second plot gives a local explanation in terms of shapley value of the features extracted from the image. The third plot gives a global explanation. The global explanation is created such that it considers nearest (training) points that were calculated using clustering process. Then provides the average of the shapley values of the points obtained.

SUPPORTING INFORMATION

Appendix shows the working snippet of the application and further feature-feature correlation analysis plots.

ACKNOWLEDGMENT

We would especially like to thank our mentors M.Sc. Uli Niemann and Dr.-Ing. Monique Meuschke for their incredible support and guidance throughout the project. This project has been a journey of an experience and it has taught us tremendously.

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-at-the-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 (COVID-19): 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 COVID-19: 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 Sejdić. "Radiological images and machine learning: trends, perspectives, and prospects." Computers in biology and medicine 108 (2019): 354-370.

[17] Sánchez-Oro, Raquel, Julio Torres Nuez, and Gloria Martínez-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.

[19] Cleverley, Joanne, James Piper, and Melvyn M. Jones. ”The role of chest radiography in confirming covid-19 pneumonia.” bmj 370 (2020).

[20] Cozzi, Diletta, et al. ”Chest X-ray in new Coronavirus Disease 2019 (COVID-19) infection: findings and correlation with clinical outcome.” La radiologia medica 125 (2020): 730-737.

[21] Rajpurkar, Pranav, et al. ”Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning.” arXiv preprint arXiv:1711.05225 (2017).

[22] Santosh, K. C. ”AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data.” Journal of medical systems 44.5 (2020): 1-5.

[23] Sethy, Prabira Kumar, et al. ”Detection of coronavirus disease (COVID-19) based on deep features and support vector machine.” (2020).

[24] Wang, Linda, Zhong Qiu Lin, and Alexander Wong. ”Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images.” Scientific Reports 10.1 (2020): 1-12.

[25] Hemdan, Ezz El-Din, Marwa A. Shouman, and Mohamed Esmail Karar. ”Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images.” arXiv preprint arXiv:2003.11055 (2020).

[26] Apostolopoulos, Ioannis D., and Tzani A. Mpesiana. ”Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks.” Physical and Engineering Sciences in Medicine 43.2 (2020): 635-640.

[27] Hasan, Ali M., et al. ”Classification of covid-19 coronavirus, pneumonia and healthy lungs in ct scans using q-deformed entropy and deep learning features.” Entropy 22.5 (2020): 517.

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

[29] Wang, Linda, Zhong Qiu Lin, and Alexander Wong. ”Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images.” Scientific Reports 10.1 (2020): 1-12.

[30] Wu, Yu-Huan, et al. ”Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation.” IEEE Transactions on Image Processing (2021).

[31] 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).

[32] Chen, Yuanfang, et al. ”An interpretable machine learning framework for accurate severe vs non-severe covid-19 clinical type classification.” Available at SSRN 3638427 (2020).

[33] https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26

[34] Gongde Guo1, Hui Wang , David Bell , Yaxin Bi, and Kieran Greer School of Computing and Mathematics, University of Ulster Newtownabbey, BT37 0QB, Northern Ireland, UK G.Guo,H.Wang,Krc.Greer@ulst.ac.uk 2 School of Computer Science, Queen’s University Belfast, Belfast, BT7 1NN, UK DA.Bell,Y.Bi@qub.ac.uk (kNN)

[35] https://scikit-learn.org/stable/modules/svm.html

[36] https://www.sciencedirect.com/topics/engineering/random-forest

[37] https://machinelearningmastery.com/neural-networks-crash-course/

[38] https://machinelearningmastery.com/crash-course-convolutional-neural-networks/

[39] Chest X-Ray Images (Pneumonia) by PaulcMooney https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

[40] RSNA Pneumonia Detection Challenge https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

[41] COVID-19 Image Data Collection: Prospective Predictions Are the Future Joseph Paul Cohen and Paul Morrison and Lan Dao and Karsten Roth and Tim Q Duong and Marzyeh Ghassemi arXiv:2006.11988, https://github.com/ieee8023/covid-chestxray-dataset, 2020

[42] Zhao, Y. Zhang, X. He, P. Xie, COVID-ct-dataset: A ct scan dataset about covid-19, ArXiv. abs/2003.13865 (2020).

[43] Niemann U, Boecking B, Brueggemann P, Mebus W, Mazurek B, Spiliopoulou M. Tinnitus-related distress after multimodal treatment can be characterized using a key subset of baseline variables. PLoS One. 2020 Jan 30;15(1):e0228037. doi: 10.1371/journal.pone.0228037. PMID: 31999776; PMCID: PMC6991951.

[44] Hefeda, M.M. CT chest findings in patients infected with COVID-19: review of literature. Egypt J Radiol Nucl Med 51, 239 (2020). https://doi.org/10.1186/s43055-020-00355-3

[45] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U.R. Acharya, Automated detection of covid-19 cases using deep neural networks with x-ray images, Computers in Biology and Medicine. (2020) 103792.

[46] Miller, Tim. ”Explanation in artificial intelligence: Insights from the social sciences.” Artificial intelligence 267 (2019): 1-38.

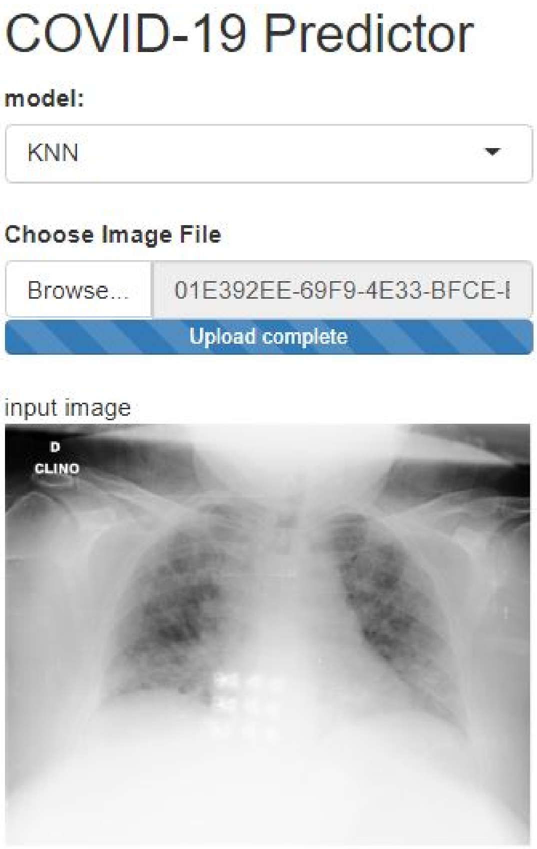
[47] Doshi-Velez, Finale, and Been Kim. ”Towards a rigorous science of interpretable machine learning.” arXiv preprint arXiv:1702.08608 (2017).

[48] Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.

[49] R.M. Pereira, D. Bertolini, L.O. Teixeira, C.N. Silla Jr, Y.M. Costa, COVID-19 identification in chest x-ray images on flat and hierarchical classification scenarios, Computer Methods and Programs in Biomedicine. (2020) 105532.

[50] Lundberg, Scott, and Su-In Lee. ”A unified approach to interpreting model predictions.” arXiv preprint arXiv:1705.07874 (2017).

APPENDIX



COVID-19 Predictor

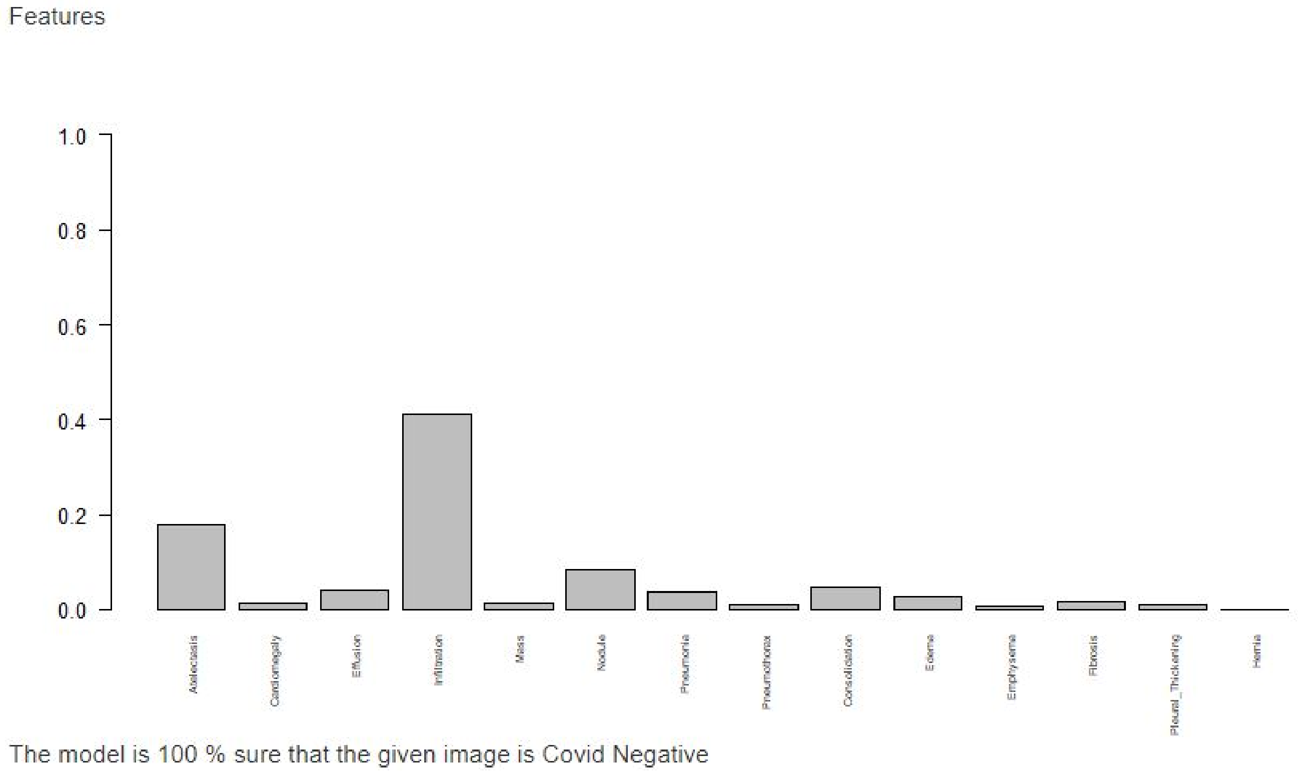
model: KNN

Choose Image File: 01E392EE-69F9-4E33-BFCE-I

Upload complete

input image

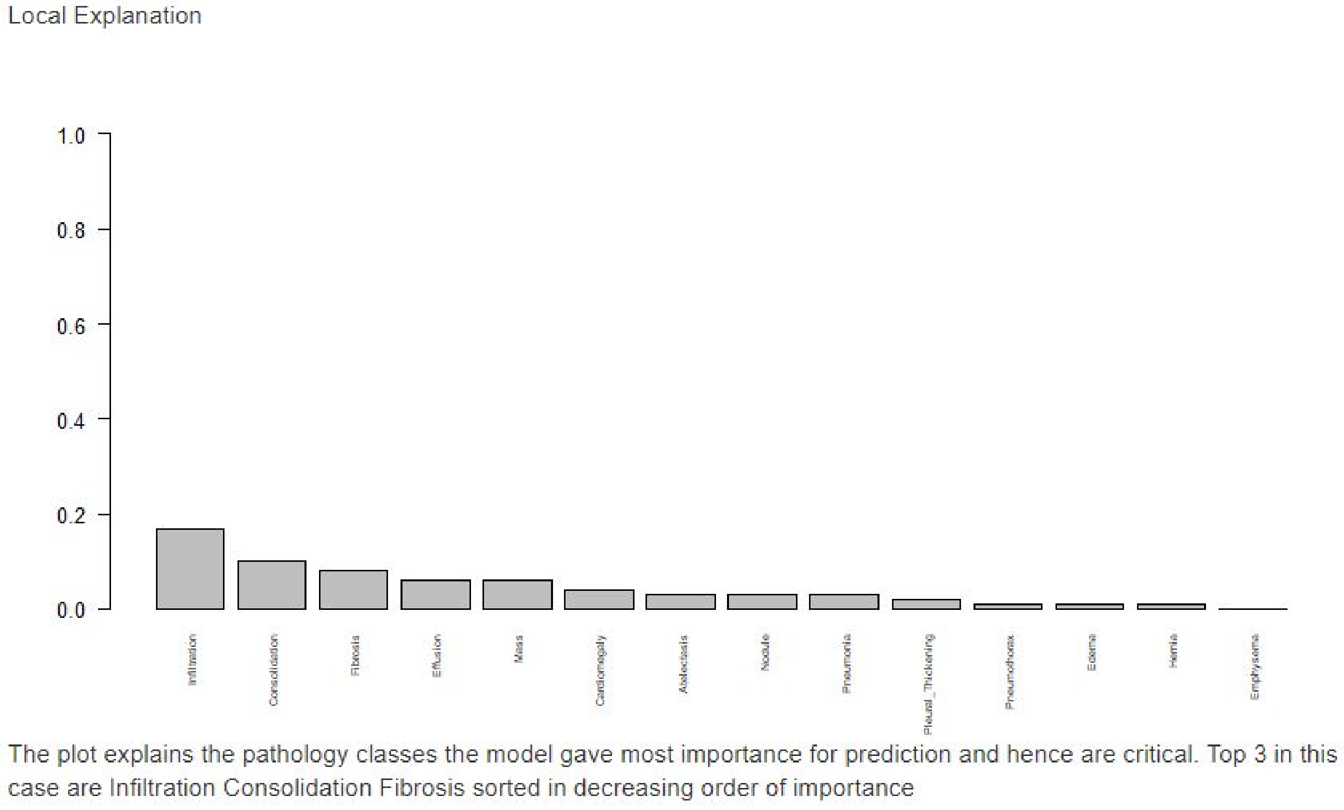
Fig. 15. Input Image Page



Features

The model is 100 % sure that the given image is Covid Negative

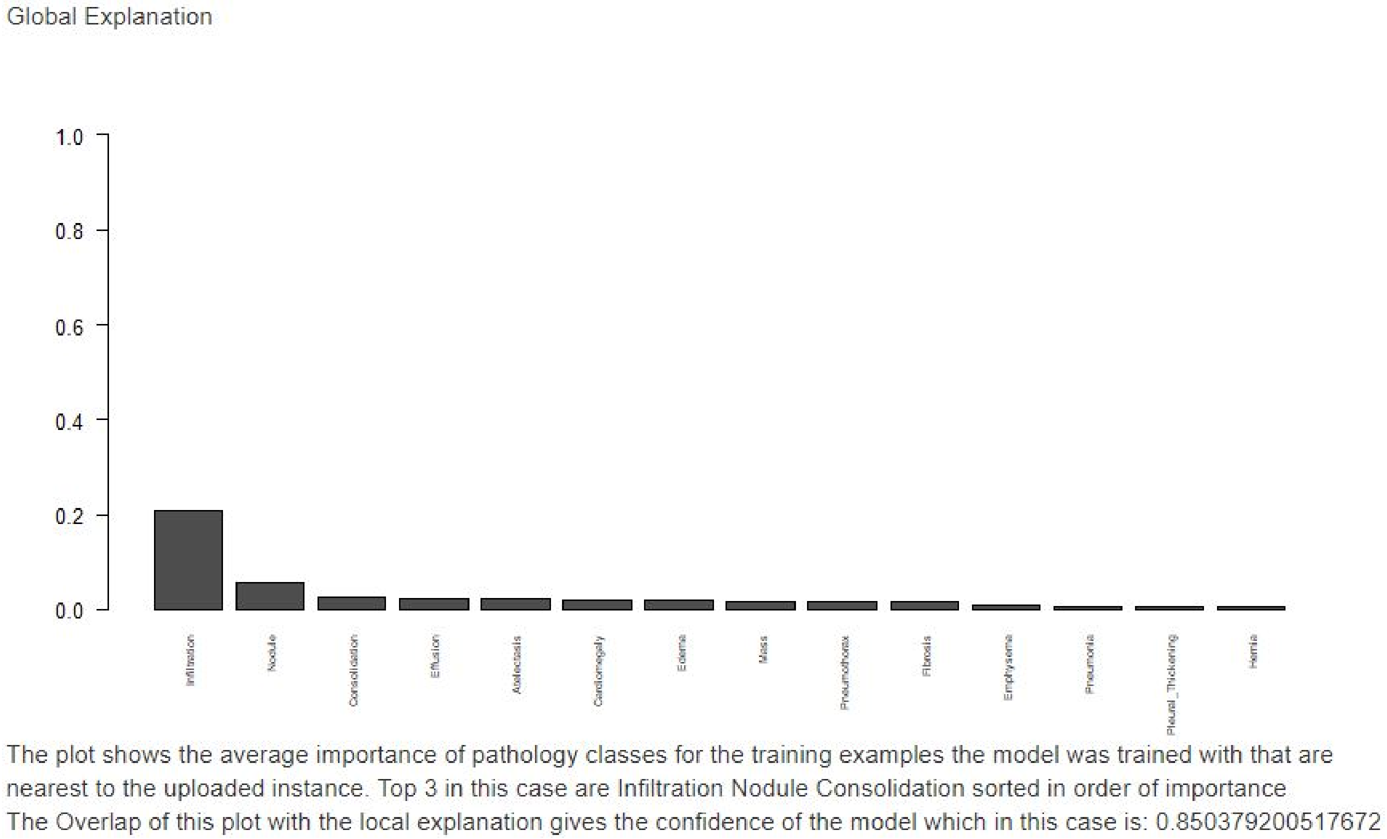
Fig. 16. Feature Extraction



Local Explanation

The plot explains the pathology classes the model gave most importance for prediction and hence are critical. Top 3 in this case are Infiltration Consolidation Fibrosis sorted in decreasing order of importance

Fig. 17. Local Explanation

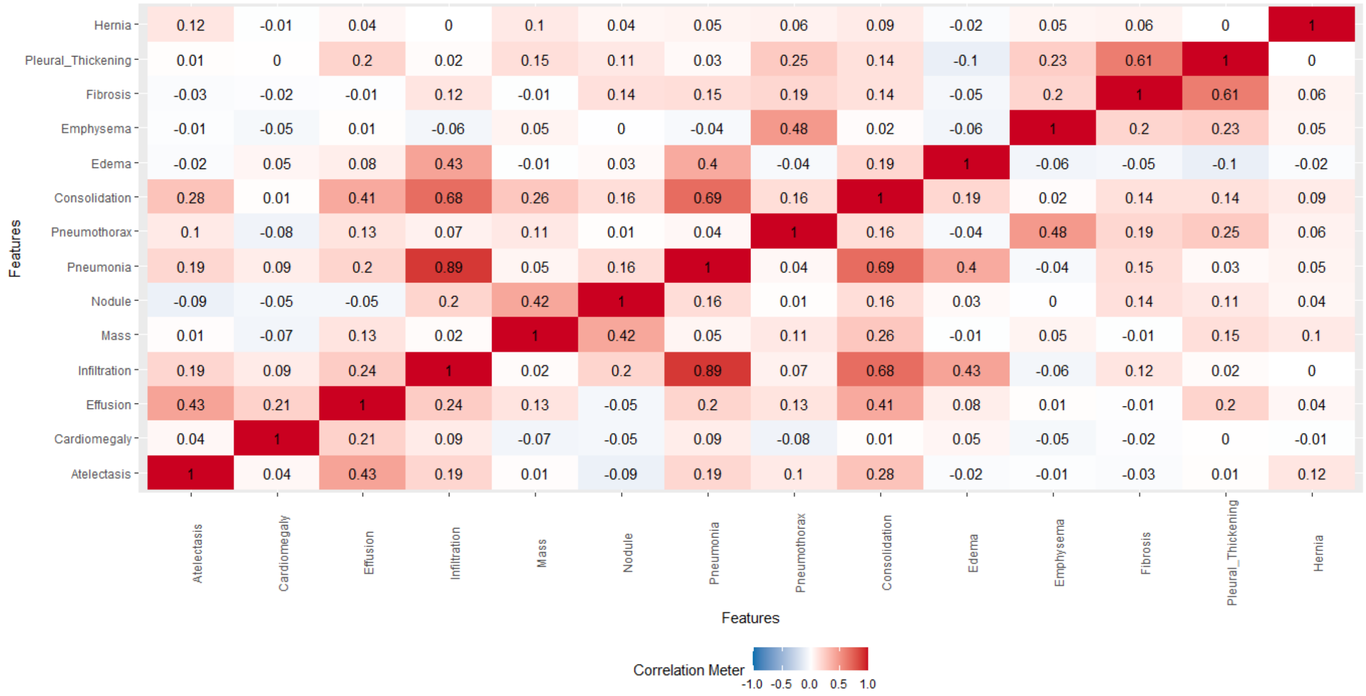


Global Explanation

The plot shows the average importance of pathology classes for the training examples the model was trained with that are nearest to the uploaded instance. Top 3 in this case are Infiltration Nodule Consolidation sorted in order of importance

The Overlap of this plot with the local explanation gives the confidence of the model which in this case is: 0.850379200517672

Fig. 18. Global Explanation



	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Consolidation	Edema	Emphysema	Fibrosis	Pleural_Thickening	Hernia
Hernia	0.12	-0.01	0.04	0	0.1	0.04	0.05	0.06	0.09	-0.02	0.05	0.06	0	1
Pleural_Thickening	0.01	0	0.2	0.02	0.15	0.11	0.03	0.25	0.14	-0.1	0.23	0.61	1	0
Fibrosis	-0.03	-0.02	-0.01	0.12	-0.01	0.14	0.15	0.19	0.14	-0.05	0.2	1	0.61	0.06
Emphysema	-0.01	-0.05	0.01	-0.06	0.05	0	-0.04	0.48	0.02	-0.06	1	0.2	0.23	0.05
Edema	-0.02	0.05	0.08	0.43	-0.01	0.03	0.4	-0.04	0.19	1	-0.06	-0.05	-0.1	-0.02
Consolidation	0.28	0.01	0.41	0.68	0.26	0.16	0.69	0.16	1	0.19	0.02	0.14	0.14	0.09
Pneumothorax	0.1	-0.08	0.13	0.07	0.11	0.01	0.04	1	0.16	-0.04	0.48	0.19	0.25	0.06
Pneumonia	0.19	0.09	0.2	0.89	0.05	0.16	1	0.04	0.69	0.4	-0.04	0.15	0.03	0.05
Nodule	-0.09	-0.05	-0.05	0.2	0.42	1	0.16	0.01	0.16	0.03	0	0.14	0.11	0.04
Mass	0.01	-0.07	0.13	0.02	1	0.42	0.05	0.11	0.26	-0.01	0.05	-0.01	0.15	0.1
Infiltration	0.19	0.09	0.24	1	0.02	0.2	0.89	0.07	0.68	0.43	-0.06	0.12	0.02	0
Effusion	0.43	0.21	1	0.24	0.13	-0.05	0.2	0.13	0.41	0.08	0.01	-0.01	0.2	0.04
Cardiomegaly	0.04	1	0.21	0.09	-0.07	-0.05	0.09	-0.08	0.01	0.05	-0.05	-0.02	0	-0.01
Atelectasis	1	0.04	0.43	0.19	0.01	-0.09	0.19	0.1	0.28	-0.02	-0.01	-0.03	0.01	0.12

Fig. 19. Pairwise Feature-Feature Correlation Analysis

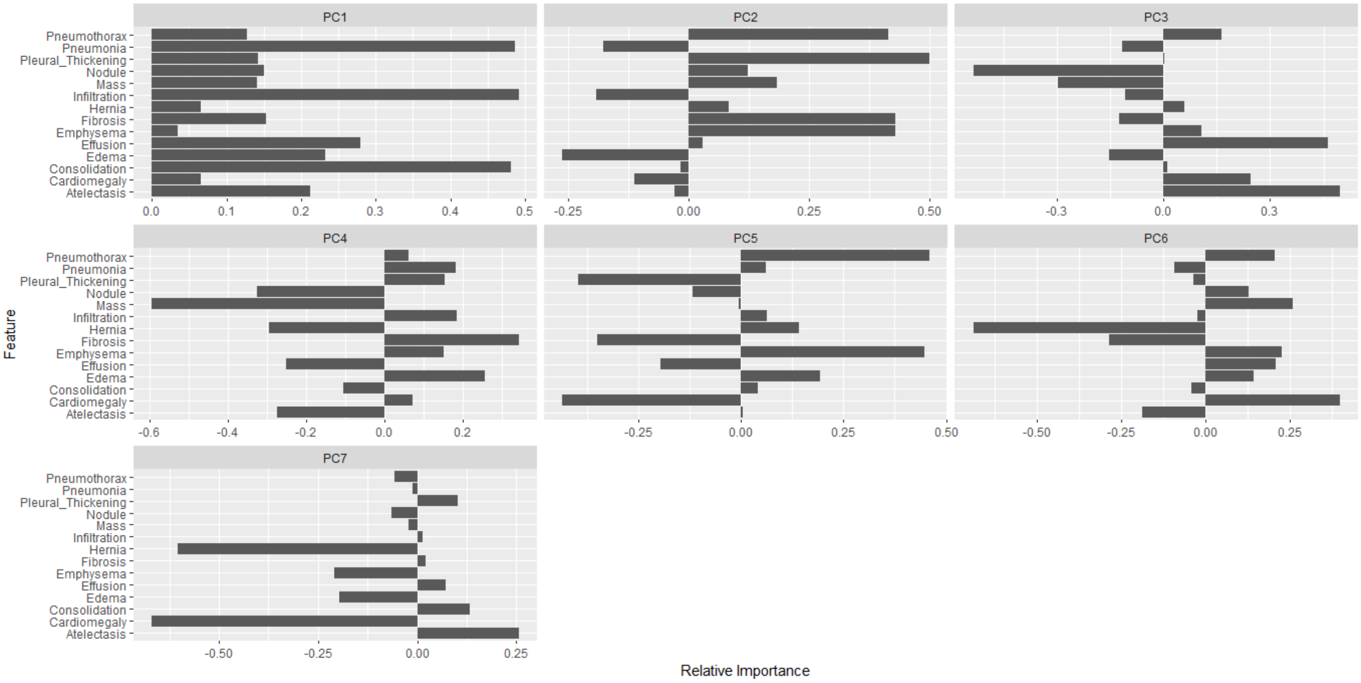


Fig. 20. Principal Component Analysis for feature importance

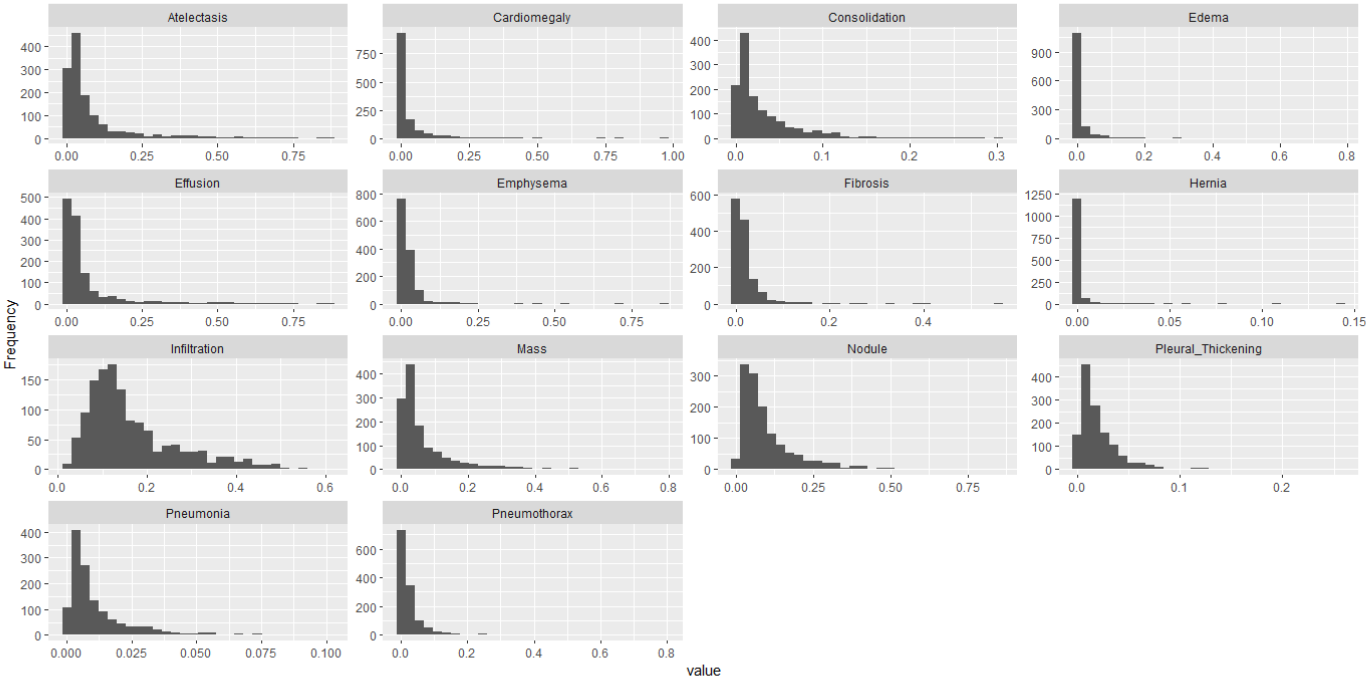


Fig. 21. Univariate Distribution of the features