Detection and Diagnosis of COVID-19 via SVM-based Analyses of X-Ray Images and Their Embeddings

Karan S Soin
The Shri Ram School Moulsari, DLF Campus
Gurugram, India

Abstract:- COVID-19 has led to a worldwide surge of patients with acute respiratory distress syndrome (ARDS) in intensive care units. Milder cases of the virus that do not reach the ARDS stage are still often characterized by inflammation of the lung, causing shortness of breath. A salient step in fighting COVID-19 is the ability to detect infected patients early enough to be able to put them under special care. Detecting the virus from radiology images can be a much-needed, expeditious method to diagnose patients. Given this, we propose a method to detect COVID-19 using chest X-ray images as well as embeddings generated from such images. The model used to train this COVID-19 data is a Support Vector Machine (SVM) Classifier. We achieved an accuracy of 55% on raw image data and 63% on embeddings of X-ray images generated using Resnet. Further refinement is possible by training a larger image data set, extra pre-processing steps and data image refining techniques, and more sophisticated modelling to improve accuracy.

Keywords:- *SVM, COVID-19, X-ray, Machine Learning.*

I. INTRODUCTION

To date, the COVID-19 pandemic has infected 37 million people, and claimed just over 1 million lives globally [1]. The disease, caused by the *Severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), spreads via respiratory droplets from coughing, sneezing, or talking as well as common contact with contaminated surfaces. There is a growing consensus now that the virus can also spread via airborne particles [2]. In the majority of the cases, COVID-19 is either asymptomatic or a mild respiratory infection with fever and cough. However, a significant minority require hospitalization with breathlessness, pneumonia and other serious complications. Nearly a fifth of those hospitalized need intensive care and ventilation. Due to the sheer numbers, healthcare facilities have been stretched in an unprecedented manner, and intensive care units across the globe have been inundated with patients needing specialized respiratory care for acute respiratory distress syndrome (ARDS). Milder cases of the virus that do not reach ARDS are still often characterized by pneumonia and/or inflammation of the lung, causing shortness of breath.

Currently there is no vaccine or drug against COVID-19 approved for widespread use. In light of the above, a salient step in fighting COVID-19 will continue to be our ability to detect infected patients in a timely manner so as to be able to assist with physical isolation to avoid infection spread and progression, and/or provide special care if need be.

Studies have shown specific abnormalities in the chest radiographs of patients infected with COVID-19 [3]. Detecting this disease from radiological imaging presents an opportunity for quicker and more cost-effective disease detection and prognostication. With this aim, we propose a method to detect COVID-19 using chest X-ray images as well as embeddings generated from such images. While embeddings have been generated using the Resnet-50 CNN which was pre trained on ImageNet dataset, both datasets comprising raw image data as well as embeddings data are trained using a Support Vector Machine (SVM) Classifier.

II. BRIEF LITERATURE REVIEW

There has been a boom of COVID-19-related imaging data and AI resources coming from both academic and industry settings. Take for instance the “COVID-19 + Imaging AI Resources" portal by the Center for Artificial Intelligence in Medicine Imaging at Stanford University that actually amalgamates many such resources [4].

Despite the recency of the COVID-19 crisis, there exists notable work in literature that helps lay the foundation for this paper. At the outset of the pandemic, Panacea Medical Technologies, India used 30 chest X-ray images for COVID-19 classification, where they extracted GLCM (Gray-Level Co-Occurrence Matrix) features to train COVID-19 images on an SVM classifier [5]. With a limited dataset and 50-50 class distribution between COVID/non-COVID images, they were able to achieve an accuracy of 57.1%. Sethy, Behera, Ratha and Biswas in the International Journal of Mathematical, Engineering and Management Sciences developed a hybrid approach between extracting deep features from a fully connected layer of a Convolutional Neural Network (CNN) model and feed them into an SVM for final classification [6]. With 127 images in their study, they were able to demonstrate a viable proof of concept that feature embeddings generated from neural network model layers combined with a support vector machine (SVM) can indeed help with quick detection of COVID infected patients when using their X-ray images to appreciable levels of accuracy. In fact, their approach towards combining the best of both worlds between Neural Network architecture and traditional supervised machine learning techniques is a source of motivation for a similar methodology adopted in this paper that is talked about in the next section.

Mucahid Barstugan et.al developed and classified Coronavirus (COVID-19) using CT images by Machine Learning Methods [7]. GLCM, LDP, GLRLM, GLSZM were used as feature extraction and support vector machines for classification. Additionally, Togacar, Ergen and Comert in their research aim to detect the disease with deep learning models using COVID-19, normal, and pneumonia chest data [8]. By utilizing techniques like social mimic optimization to extract feature sets from CNN models, they are able to classify between the aforementioned three classes with a success rate of 99.27% using an SVM.

III. METHOD

Two experiments were carried out in this research project. While both adopt an SVM classifier, the first directly feeds features generated from raw image data into an SVM while the second approach uses embeddings generated from the Resnet-50 CNN as feature vectors. Both methods ultimately utilized 5-fold cross validation using Grid Search CV functionality in scikit-learn along with a random 80-20 split between training and evaluation data.

A. Data

1) Images

For the first task, we used BIMCV-COVID19+ [9]:

This is a large dataset with chest X-ray images CXR (CR, DX) and computed tomography (CT) imaging of COVID-19 patients along with their radiographic findings, pathologies, polymerase chain reaction (PCR), immunoglobulin G (IgG) and immunoglobulin M (IgM) diagnostic antibody tests and radiographic reports from Medical Imaging Databank in Valencian Region Medical Image Bank (BIMCV). The first iteration of the database, which we worked on, includes 1380 CX, 885 DX and 163 CT studies. For the image-based SVM task, CT studies and lateral views of X-ray images are omitted, leaving us with 721 COVID, 262 normal and 1138 other (non-COVID but abnormal) images. This sets us up for a multi-class classification task.

2) Image Embeddings

For this task also we considered BIMCV-COVID19+ images and instead applying SVM on images we used feature vectors generated a well-known deep convolutional neural network, Resnet [10]. We considered 2048 dimension feature vectors, also called embeddings of images, which are obtained by removing the final few layers from the standard Resnet architecture and taking the output. We used Resnet pretrained on the ImageNet Dataset to generate the desired embeddings. Subsequently used SVM to perform multiclass classification based on the generated embeddings.

B. Background on ResNet

Resnet (or residual network) is a convolutional neural network (CNN) based architecture that was proposed in 2015 by the team at Microsoft Research [10]. This architecture came about in order to solve the problem of the vanishing/exploding gradient that has been extremely common in neural network computations. In this network we used a technique called skip connections. The skip connection skips training from a few layers and connects directly to the output. The advantage of adding this approach is that any layer hurting the performance of the architecture will then be skipped by regularization. In this case, the embeddings were trained on the ImageNet dataset. The ImageNet dataset is a large collection of human annotated photographs designed by academics for developing computer vision algorithms. It contains over 14 million images [9].

C. SVM-based classification

A Support Vector Machine (SVM) Classifier was utilized for both tasks i.e., training on imaging data as well as on Resnet-generated embeddings. Since we used a classifier other than a deep neural network that was expected to discover features automatically, there was a fair amount of image pre-processing involved.

1) SVM Background

SVM is a supervised machine learning algorithm that is used mostly for classification problems. In the SVM algorithm, we plotted each data item as a point in n-dimensional space (where n is number of features available) with the value of each feature being the value of a particular coordinate. Then, we performed classification by finding the hyper-plane that differentiates the two classes. For multi-class classification, which is the approach followed in this paper, the same principle was utilized. The multiclass problem was broken down to multiple binary classification cases, which is also known as one-vs-one, or one-vs-rest which divides the data points in class x and rest. Table 1 displays the particular configurations of hyperparameters for the experiments in this paper. Some reasons we chose an SVM include, but are not limited to:

SVMs work well with unstructured and semi-structured data like text, images and trees [11]. Its kernel trick along with an appropriate function enables us to solve complex problems. SVMs scale relatively well to high dimensional data, which is something that came up in our experiment. SVM models also show good generalization in practice, making them less susceptible to overfitting.

Experimental Flows

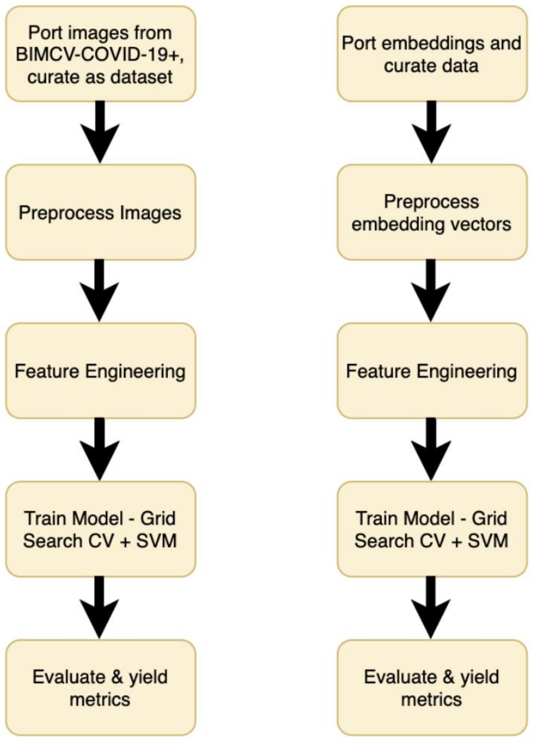
Port images from BIMCV-COVID-19+, curate as dataset
Preprocess Images
Feature Engineering
Train Model - Grid Search CV + SVM
Evaluate & yield metrics
Port embeddings and curate data
Preprocess embedding vectors
Feature Engineering
Train Model - Grid Search CV + SVM
Evaluate & yield metrics

Fig. 1: Image-based task (left) and embeddings task (right)

2) Pre-processing and Feature Engineering:

i) Experiment 1: For this experiment, raw image data served as input. Image data is represented as a matrix, where the depth is the number of channels. An RGB image has three channels (red, green, and blue) whereas the returned grayscale image has only one channel.

We applied a slate of image pre-processing techniques:

- Read in image as grayscale.
- Image resizing (500, 500) and cropping to the center
- Enhance image using histogram equalization, a technique that improves contrast in images by spreading out the most frequent intensity values, i.e. stretching out the intensity range of the image.
- Begin feature engineering, apply a standard scaler so that features have properties of a standard normal distribution
- Transform the data using Principal Components Analysis (PCA): A linear transformation such that most of the information in the data is contained within a smaller number of features called components.

ii) Experiment 2: For this experiment, embeddings generated using Resnet and trained using ImageNet form the basis of the feature set for the classifier. The pre-processing and feature extraction here is as follows:

- Read in embeddings that are originally stored as pickle files, convert to readable arrays
- Concatenate the feature vectors into a feature matrix
- Apply standard scaling to prepare for training

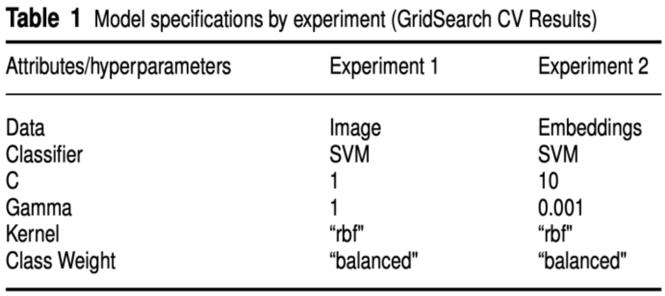

Table 1 Model specifications by experiment (GridSearch CV Results)

Attributes/hyperparameters	Experiment 1	Experiment 2
Data	Image	Embeddings
Classifier	SVM	SVM
C	1	10
Gamma	1	0.001
Kernel	"rbf"	"rbf"
Class Weight	"balanced"	"balanced"

IV. RESULTS

Using a grid search over the hyperparameter space with cv = 5, we first found optimal hyperparameters before evaluating the model, as listed in the table above.

For the experiment where an SVM was fed features directly obtained from raw image data, the SVM classifier yielded an accuracy of 55%. In the embeddings-based task, we obtain 63% accuracy in predicting between COVID, normal and non-COVID but abnormal X-ray images.

Furthermore, for this experiment, we also compute ROC metrics, which are indicative of the model's ability to discriminate between cases (positive examples) and non-cases (negative examples). Specifically, we obtain 72% One-vs-One and 71% One-vs-Rest AUROC scores.

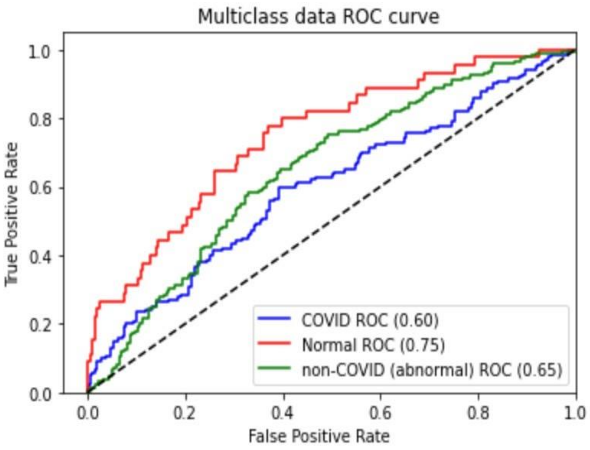

Fig. 2: AUROC Analysis - COVID vs Non-COVID Abnormal vs Normal X-rays

GitHub link to the source code: **https://github.com/karansoin/COVID-19-X-RAY-SVM-Project**

V. DISCUSSION

Two experiments with SVM classifiers were conducted, one of which fed raw image data into the model while the other used pregenerated embeddings of X-ray images from a Resnet model. Varying image pre-processing and feature engineering approaches were also utilized. Finally, data was standardized (in both experiments) and used to train an SVM classifier that was able to predict on the test set with scores of between 55-63% depending on the experiment.

The marked improvement in performance with embeddings as feature vectors shows that a dual approach whereby a neural network is used in part to generate a feature set (embeddings) and then an SVM classifier is used to train those very features is likely to yield more favorable results for COVID-19 detection. With this approach, we were able to improve upon the results of R Reddy et al. who with a 50-50 class distribution between COVID/non-COVID images, attained an accuracy of 57.1% [5]. However, our model fell short of the accuracy obtained by Togacar et al. who used social mimic optimization to extract feature sets from CNN models, to obtain an accuracy of 99.27% in classifying COVID/other pneumonia/normal images [8]. This was possible since they converted images to JPG before classification, something that was not possible in my experiment due to restrictions on storage. Additionally, the original dataset was refined through the use of the Fuzzy Color technique which removed the image noise. To create a better data quality image, they also used the stacking techniques. Finally, they trained the datasets using the MobileNetV2 and SqueezeNet deep learning models and classified the models by the SVM method.

Needless to say, my work is a basic stepping stone to more sophisticated models, and serves to demonstrate proof of concept of an idea that is gaining steam in the quest to find quicker and more economical solutions towards COVID-19 detection.

A. Limitations

Though we worked with image sample sizes of 1000+ (significantly higher than some studies cited in the literature review), an ideal next step would be to train models on even more images. Further, the images in my experiment used images resized to 500 by 500 pixels, hence potentially hindering the overall accuracy.

A limitation of the application of an improved, more accurate model would be the close involvement of a physician to allow its real-world manifestation.

VI. CONCLUSIONS

In our first experiment where an SVM was fed features directly obtained from raw image RGB data which returned grayscale images, the SVM classifier yielded an accuracy of 55%. In the second, we used Resnet CNN embeddings, and obtained an improved 63% accuracy in predicting between COVID, normal and non-COVID but abnormal X-ray images.

VII. THE FUTURE

Future work must use larger data sets, extra pre-processing steps and data image refining techniques, and more deep learning models and other sophisticated methods to improve accuracy to distinguish COVID-19 from multiple other pathologies. The models must also root themselves in strong ethical foundations (protecting patient privacy, minimizing model biases etc.). Such systems, if deployed off-the-shelf and in a clinical setting, will potentially automate the diagnosis of COVID via a simple X-ray in the majority of the cases, hence enhancing the reach of a remotely placed physician. The physician's expertise however, will still be needed to validate this, and for complex diagnoses.

ACKNOWLEDGMENT

I would like to thank my mentor Dr. Deepak Mishra, Assistant Professor at IIT Jodhpur for giving me the invaluable opportunity to conduct this research, and for his constant feedback and encouragement along the way.

REFERENCES

[1]. Worldometers COVID-19 dashboard, https://www.worldometers.info/coronavirus/

[2]. McCallum K.:' 'What Does Coronavirus Do to the Lungs?', https://www.houstonmethodist.org/blog/articles/2020/jul/what-doescoronavirus-do-to-the-lungs/

[3]. Minaee S., Kafieh R., Sonka M., Yazdani S., Soufi, G.: 'Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning', Medical Image Analysis Volume 65, October 2020, 101794. https://doi.org/10.1016/j.media.2020.101794

[4]. COVID-19 + Imaging AI Resources. Stanford University: https://aimi.stanford.edu/resources/covid19

[5]. Reddy R., Panacea Medical Technologies: 'COVID-19 Detection using SVM Classifier'. International Journal of Engineering Science and Computing, Volume 10 Issue No.4.

[6]. Sethy P., Behera S., Ratha P., Biswas P.: 'Detection of Coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine'. International Journal of Mathematical, Engineering and Management Sciences , Vol. 5, No. 4, 643-651, 2020. https://doi.org/10.33889/IJMEMS.2020.5.4.052 643

[7]. Barstugan M., Ozkaya1 U., Ozturk S.: Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods

[8]. Togacar M., Ergen B., Comert Z.: 'COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches'. Computers in Biology and Medicine Volume 121, June 2020, 103805. https://doi.org/10.1016/j.compbiomed.2020.103805

[9]. BIMCV-COVID19+ data. https://osf.io/nh7g8/

[10]. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778)

[11]. Mitosis Tech. https://www.mitosistech.com/support-vector-machine/