Data Science with R Project Proposal

Team: COVID-19 Predictor

20.05.2020

Project Title: COVID-19 Prediction using Explainable Machine Learning

Background and Motivation:

The COVID-19 or the SARS-CoV-2 originated from the district of Wuhan, China has transpired to be a pandemic worldwide [1]. Research on the COVID-19 is a hot topic among the Artificial Intelligence community recently. Due to shortage and limited efficiency of current testing mechanism of COVID-19 tests, i.e. through RT-PCR kits [2] which usually takes upto 4-6 hours to reproduce the results which is not very optimal way to move forward as the rate of COVID-19 patients registered grows exponentially. With this problem in scientific community, it motivated to aim of Data Science Methodology be brought to be a part in helping flattening the curve. So, this lead to a possibility of building classifiers which can diagonise patients as COVID-19 negative or positive based on their respective X-Ray images [citation of some similar experiments]. As this approach is can be less time and resource consuming and hoped to achieve more streamlined performance compared to RT-PCR kits. Also in addition to a good prediction, we needed reasons that could justify what could be the features that are responsible in the diagonistic process [3].

With this idea and motivation in hand, our work tries to experiment in building classifier with CXR (Chest X-Rays) as Ground Truth predicts whether an X-Ray image belongs to COVID-19 negative or positive. Along with, we try to come up with features that contributes to the detection of an image and also with an explaination delineating why was such a behaviour observed.

Project Objective:

From the motivation to help flattening the curve of COVID-19,

> Can we use a Data Science for COVID-19 diagnosis?

To answer the this question, we aim to answer few sub-questions:

- How well could classifiers perform on Chest X-Rays?
- Although [2] and [3] extensively works with Neural Networks (Black-Box Model) to classify, Can simple and intrinsically explainable classifiers achieve a base Accuracy, F_1-Score and AUC of 85% using CXR?
- How does different features of CXR contribute to the model prediction and Can we come up with few number of feature w.r.t their importance?
- Which flavour of ~~multilabel transform~~ algorithm perform best among all, the one which considers label correlation or the one which does not?
- Can we come up with explaination of our model's decsision and prediction?

Ground Truth and Technology Stack:

Technology Stack

The project will be built in `R` with usage of API's like `magick`, `opencv` for image processing and `tidyverse` packages like `dplyr` and `tidyr` for data manipulation, `ggplot2` for data visualization, `rmarkdown` and `knitr` for reproducible and automated reporting, `shiny` for interactive web applications and `tidymodels` for inferential and predictive modeling.

Dataset

Our Dataset 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 of [3][1], Kaggle CXR Pneumonia dataset by Paul Mooney,[2] CXR images of adult subjects from the RSNA Pneumonia Detection Challenge,[3] original and augmented versions of COVID-19 examples[4] from [4].

According to [2,3,5,6] CT-Scan data would be gold-standard for us and also potray pretty good results 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 competible in terms of quality w.r.t CT-Scans but [7] suggests CXR to be sufficient and comparable to CT-Scans in order to diagnose COVID-19 patients.

In particular we will use the COVID-19 Dataset Repository as our Ground Truth.

Github

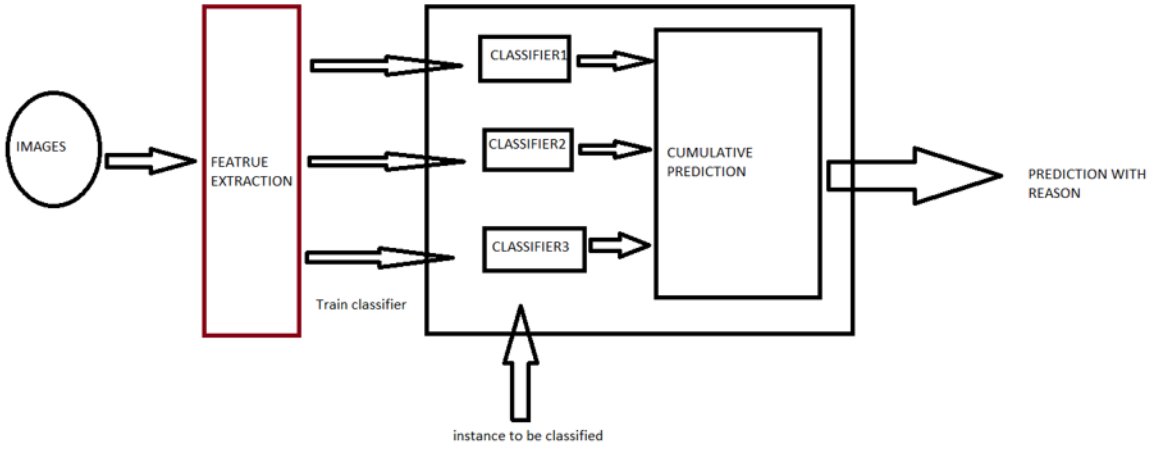
The `R` scripts, process notebook and other resources have been stored at the repository.

Design overview (Algorithms and Methods):

- **Pre-Processing**
  - We would be following a typical Data Science pipeline starting with Pre-Processing of the Dataset, Feature Extraction and Selection and then feeding descriptors (Trainable Vectors) to different classifiers to train and test and then finally evaluation would be done based on predictor's results. The details are delineated in the following sections.
  - For the part of feature extraction, we will try texture-based descriptors.
  - There exist several texture-based vision algorithms. We will try to combine features before training and train our model on a combined feature set. Or else we can train models on individual features, and then combine prediction results might be combined and thus one feature might only not be selected but multiple features can be selected.
  - Literature survey tells us Local Binary Patterns shall be a good choice for texture-based descriptor. We will also try to use pretrained networks to gain texture descriptors or vision API's for the extraction part.
  - Moreover, there are several Neural-Nets we faced in literature survey, that are carefully curated for the purpose of the COVID-19, which requires the image to directly fed to the net, and thereby auto-encodes the parameters.

[1] https://github.com/rezacsedu/DeepCOVIDExplainer
[2] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
[3] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
[4] https://github.com/ieee8023/covid-chestxray-dataset

- **Classification**
  - The problem in hand is a classification problem where we would be classifying an image being COVID-19 positive or negative.
  - Here, we would like to emphasize that the model won't predict presence or absence or pneumonia, which is a result not only of COVID-19 but other kind of reasons also affect this.
  - We intend to work on two kinds of algorithms:
    1. One that can be trained based on features to be extracted such as:
    * Clustering Algorithms
      · k-Means
      · kNN
    * Support Vector Machines
    * Binary Classifiers
      · Decision Trees
      · Naive Bayes
    * Mostly because the models are intrinsically explainable.
    2. Neural Network based approaches, where the model is a black box model and we will use tools like saliency maps for description.



- The above figure suggests the overall overview of an architecture of the system we would like to develop.
- We will have images as data and we would pre-proccess by cropping of images and extract features from them.
- Then after, the features or preprocessed images are fed to the classifier for training.
- Once trained unknown instance is supplied for classification.
- We will use late fusion; hence we accumulate the prediction of each of the classifier with certain confidence.
- The confidence shall be extracted by calculating the MAP score [8].
- The Map score can be calculated on a test set that shall be segregated from the overall dataset before the training phase begins.
- The ratio of the MAP score shall give the confidence contribution of each of the systems.

- **Evaluation**
  - Evaluation Metrics are used to calculate the performance of the model.
  - We have different type of evaluation methods but selecting a metric is an important step in the project.
  - Most commonly used metrics are Precision and Recall.
  - Precision and Recall are also used with other metrics like Accuracy, F1-Score, Area under ROC curve, MAP Score.
  - The higher the metric value the better the performance.
  - *Accuracy:* Best and mostly used metric. Easily suited for binary as well as multiclass classification problem.
  - *Precision:* It is a best choice when we want to be very sure of our prediction.
  - *Recall:* Captures as many positives as possible
  - *F_1-Score:* It is a harmonic mean between Precision and Recall.
  - *MAP score:* Quantifies how good our model at performing the query. First we calculate the average of the precision for each query and then the mean of all these AP scores. Q – No.of queries in the set and AveP(q) – Average precision for a each query q
  - *AUC & ROC:* Indicates how well the probabilities from the positive classes are separated from the negative classes. Mostly used to check or visualize the performance of the multi-class classification models. ROC is the probability curve and AUC represents degree or measure of separability.
- **Visualization**
  - Visualization is a computer generated image using a computer representation of data as primary source and a human as its primary targets.
  - It's an abstract of information.
  - Box Plot easily display data and we can see outliers as well.
  - It graphically displays the data in five statistics Minimum quartile, 25^{th} or lower quartile, 50^{th} or median quartile, 75^{th} or upper quartile and the maximum quartile, which summarizes the distribution of dataset.
  - Our plan is to visualize our output in Box-plot diagram, for instance, we are gonna display probability of having COVID in quartiles.
  - X-ray's without COVID as minimum quartile, X-ray's which has very less probability of getting COVID as the lower quartile and continues till X-ray's which has very strong probability of getting COVID as Maximum quartile.

Time Plan:

2 Meetings per week. Tuesday and Friday at 17:30 Sharp via Zoom[5]. Every Friday, Weekly Achievements would be discussed.

Team:

- **Jalaj, Vora** *M.Sc. Digital Engineering*
- **Shivam, Singh** *M.Sc. Digital Engineering*
- **Subhankar, Patra** *M.Sc. Data and Knowledge Engineering*

[5]https://zoom.us/

- **Subhajit, Mondal** *M.Sc. Data and Knowledge Engineering*
- **Roshmitha, Thummala** *M.Sc. Data and Knowledge Engineering*

Supervised by: M.Sc. Uli Niemann

References

[1] World Health Organisation, Novel Coronavirus – China 2020, (2020). https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/.

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

[3] M. Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, O. Beyan, others, Deepcovid-explainer: Explainable covid-19 predictions based on chest x-ray images, arXiv Preprint arXiv:2004.04582. (2020).

[4] J.P. Cohen, P. Morrison, L. Dao, COVID-19 image data collection, arXiv 2003.11597. (2020). https://github.com/ieee8023/covid-chestxray-dataset.

[5] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, others, A deep learning algorithm using ct images to screen for corona virus disease (covid-19), MedRxiv. (2020).

[6] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, others, Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct, Radiology. (2020) 200905.

[7] D.S. Kermany, M. Goldbaum, W. Cai, C.C.S. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M.K. Prasadha, J. Pei, M.Y.L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V.A.N. Huu, C. Wen, E.D. Zhang, C.L. Zhang, O. Li, X. Wang, M.A. Singer, X. Sun, J. Xu, A. Tafreshi, M.A. Lewis, H. Xia, K. Zhang, Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell. 172 (2018) 1122–1131.e9. https://doi.org/https://doi.org/10.1016/j.cell.2018.02.010.

[8] C.D. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Cambridge University Press, USA, 2008.