COSC 7362 - Advanced Machine Learning

Spring-2022

Covid 19 Prediction using TensorFlow and Ensemble Model

Team

SAI ASISH SAMINENI (2088595)

VENKATESH NALLURI (2048174)

1.ABSTRACT

Covid-19 is affecting the people's life in both financially as well as mental health. Machine Learning and Deep Learning Technology can certainly help health care workers. This project illustrates the implementation of Logistic Regression, Dense Neural Network, Random Forest, SVM, K- Means clustering, XGboost using the tensor flow in which the results are compared to ensemble model in detecting the corona virus. Classification from Supervised Learning approaches are used in our approach to solving this issue. Scaling and developing automated systems that can predict the likelihood of Covid-19. Using the information in the dataset, we can accurately predict that a patient has Covid-19 with an accuracy of 87.0 to 90.6 percent and a degree of confidence of 90 percent.

Keywords: COVID-19, Machine learning, Prediction, Random Forest classifier, Logistic Regression, Dense Neural Network, Random Forest, SVM, K- Means clustering, XGboost, Ensemble Model, TensorFlow

2.INTRODUCTION

Machine Learning, Data Science and automation are fast becoming an integral part of the health care system. These technologies help medical professionals to better manage time and prioritize patients. Machine Learning is quite capable of detecting disease, which will help in diagnosing patients. This help in avoiding the critical stages of the disease. Machine Learning is widely used in automate the task and detect the covid-19.

Covid-19 detection is usually done through rapid testing and RT-PCR. These tests require time to understand the severity. Usually when a patient was tested after the results arrived the doctor must investigate it and decide the patient status. This process will sometimes make the patient situation worse. These types of cases the severity can be reduced by just introducing the Machine learning models to predict the covid-19 virus

In this research we built several algorithms like logistic regression, K-means Clustering. on the tensor flow models. The models built on the TensorFlow with help of tuning parameters will provide best results. Usually, the TensorFlow supports both CPU and GPU for the faster

compilation. The TensorFlow in many instances has proved that it has faster compilation than other libraries like keras, torch...etc.

Ensemble methods usually combines many base models to carry on the task. The Principe of the ensemble learning is weak model alone is weak, but when combined with other models it will be a strong model. Our aim is to predict the covid -19 and look for the best model that will provide the desired results.

3.PREVIOUS WORK

Several research works were done using the TensorFlow and ensemble libraries. These algorithms are best suited for certain data sets and certain tasks. In detecting the chronic kidney disease [1] extreme gradient boost has given the best result due to the set-theory based rule. The active Learning [2] was stable and consistent in detecting the covid -19 virus in lungs. This was done by scanning the lungs pictures.

Artificial Intelligence [3] was used to help in suggesting the health care workers, this played a key role in understanding, development and suggesting. As the Artificial Intelligence best mimic's, the human intelligence. One more example in Machine Learning being using multivariate predictive analysis [4] is successful in predicting the mortality rate of the hospital mortality in patients.

Few models are successful in predicting the like hood of spreading using the neural network and random forest [8]. This one is successful in using the symptoms to detect the like hood spread. As covid have fewer samples, machine learning is used to enhance and increase samples [6] and it was successful in it. Combining two methods yields good results when predicting two methods. e-registration slips and chest CT-scan[5] combination will yield best results.

4.METHODOLOGY

Data set

Hospital Israelita Albert Einstein, in So Paulo, Brazil, has given the patient data utilized in this project for research purposes. The information was gathered from the 28th of March to the 3rd of April 2020. Patients who came to the hospital with a suspected Covid infection had their blood and clinical samples taken. In a standard and normalized form, the anonymized data was made accessible to the researchers. Standard deviation is one unit and mean are zero. The D-Dimer characteristic is missing from this data, and the potassium concentrations are lower than what has been determined to be important in connection to Covid. This information was only discovered much later in the research process when the significance of these traits had yet to be fully understood.

CBC, liver function, renal analysis, salt tests, blood gas analysis (arterial and venous), and influenza tests are some of the aspects that make up diverse clinical parameters.

A total of 5644 records and 111 characteristics makes up the raw data. A key variable, 73 object variables, and 37 numeric variables make up the model's 111 characteristics.

There are a lot of blanks in the raw data. The data has been cleaned using exploratory data analysis techniques. Based on how thorough the cleaned data is, it has been separated into two groups. To preserve the integrity of the data, no attempt has been made to impute it.

Cleaning of data

Open datasets were sifted through for quality and consistency concerns before they were used in this study records marked as "null" had the value "not done” in the field. The names of attributes were trimmed to reduce unnecessary whitespace. Binary zeros and ones were used to represent categorical variables stated in various ways. Negative and positive target classes were substituted with 0 and 1 for covid negative and positive respectively. The flu variables that were modified from 'detected' and 'not detected' to 'present' and 'absent' were also subjected to the same procedures.

Attributes with missing values of greater than 98 percent were discarded. Pre-processing indicated that instead of looking at all the records, we should be looking at them as two logical categories that would assist us solve the difficulty of completeness column and row wise.... Imputations were avoided since they might affect the results, and any records with more than 10% missing column values were removed from the datasets.

As a result of this, we were able to construct two separate data bases from which to run the modelling process.

The first subset was constructed using features linked to the full blood count, age quantile, and three categorical variables reflecting the hospitalization and the target variable ('sars cov2')... RBC, WBC, and platelets make up three of the clinical spectrum's independent properties. No missing data have been found in 598 of the 20 characteristics. No imputations were made on the few records that had missing values.

Target variables for the two tasks have been predicted using binary and multi-class classification. Analyzing the Analytics Base Table, we tested several different classifier algorithms and found that Random Forest classifier was the most effective for the job at hand. A Random Forest Classifier is used to calculate the class level metrics presented.

The Gini coefficient is used as the default (the method of this study). The increase in prediction error if the values of a variable are permuted throughout the out-of-bag observations is the metric for that variable. This metric is calculated for each individual tree, then averaged over the entire group and divided by the group's overall standard deviation. In other words, a variable's importance rises in direct proportion to its Gini score (which goes from 1 to 100)..

Logistic regression predicts output which is present in categorical form. It can be either Yes or No, 0 or 1, true or False, etc., it gives a value between 0 and 1. When applying logistic regression, the output is in categorical form. Logistic regression gives a value between 0 and 1 that cannot go above this limit, forming S like the curve. The S-form curve is called the Sigmoid function or the logistic function. The probability of a record belonging to the positive class given features predicts by the Logistic regression.

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3 \ldots\ldots\ldots\ldots+\beta_n X_n)}} \quad (1)$$

The other algorithm used in this study is KNN. Using KNN, the quick identification of the category or class is achieved [7]. Working of the KNN algorithm is following:

- Selecting the number K of the neighbors
- Euclidean distance of K number of neighbors is calculated.
- The K nearest neighbors from calculated Euclidean distance.
- Counting the number of the data points from k neighbors in each category.
- For which the number of the neighbor is maximumly Assigning them new data point.

Distances functions used to calculate the distance from the nearest neighbors are Euclidean (2), Manhattan (3), and Minkowski (4).

$$Euclidean = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (2)$$

$$Manhattan = \sum_{i=1}^{k}|x_i - y_i| \quad (3)$$

$$Minkoswski = \left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \quad (4)$$

Another algorithm used in this study is Naïve Bayes. In naïve bayes algorithm occurrence of a certain feature is independent of the occurrence of other features.

Working of naïve bayes algorithm is as following

1. Converting of the given dataset into frequency tables.

2. Likelihood table generation by finding the probabilities of given features.
3. For calculating the posterior probability use Bayes theorem.

LightGBM is used to extend the gradient boosting of an algorithm, focusing on boosting examples with more significant gradients. Predictive performance can be improved by speeding up the training process, which is accomplished through automatic feature selection. The LightGBM algorithm uses two novel techniques called Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB); these techniques make the algorithm run faster while maintaining a high level of accuracy.

If ***Y*** *is the prediction and* ***X*** *is the feature vector:*

$$\boldsymbol{Y} = Base\_tree(\boldsymbol{X}) - lr * Tree1(\boldsymbol{X}) - lr * Tree2(\boldsymbol{X}) - lr * Tree3(\boldsymbol{X})$$

TensorFlow isn't limited to building neural networks. It is a framework for performing fast mathematical operations at scale using tensors, which are simply arrays. Tensors can represent scalar values (0-dimensional tensors), vectors (1D tensors), matrices (2D tensors), and so on. A neural network is basically a workflow for transforming tensors. The 3-layer perceptron takes a 1D tensor containing two values as input, transforms it into a 1D tensor containing three values, and produces a 0D tensor as output [8]. TensorFlow lets you define directed graphs that in turn define how tensors are computed. And unlike Scikit, it supports GPUs.

The Decision tree algorithm is the most straightforward and efficient supervised learning algorithm. In the decision tree, algorithm data points are split continuously according to some parameters, and the algorithm tries to solve the problem [6]. Other names of Decision trees are classification and regression trees. Decision trees follow a top-down approach. The tree leaves represent the outcomes of the decision tree. Decision trees are called Divide and Conquer means (recursive partitioning). The function of Entropy is an information theory metric that measures the impurity or uncertainty using a group of observations built using a heuristic given in (5). In the meantime, when selected randomly, Gini calculates the probability of a specified feature classified incorrectly as in (6).

$$Entropy = \sum_{i=1}^{C} -p_i * log_2(pi) \quad (5)$$

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \quad (6)$$

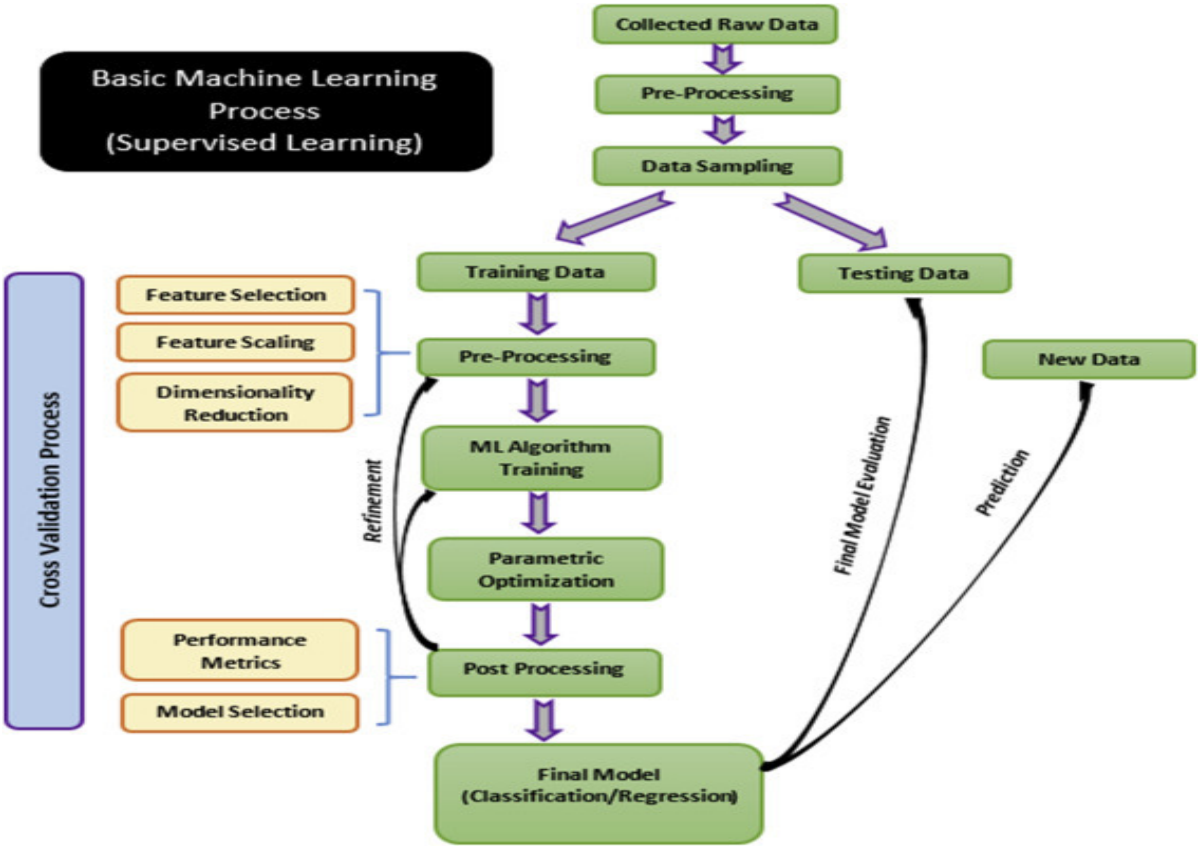

Figure1: Flowchart

5.EXPERIMENTS

Highest accuracies were obtained using the tensor flow library and accuracy we achieved is about 90.6%

Accuracies comparison figure is given in figure 2.

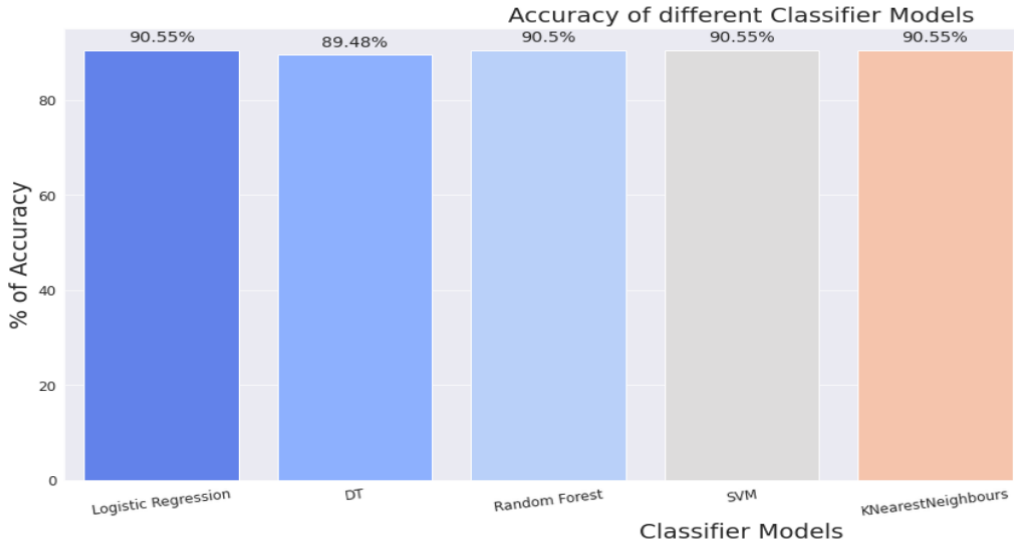

Figure2: Accuracy Comparison

Cross validation score of different models.

Figure3: Cross validation Accuracy

6.CONCLUSION

When it comes to novelty, ensemble modeling technique is used to compare the working of ensemble model with other algorithms. Ensemble model consists of XG Boost, Neural Networks, and K Nearest Neighbor. The Ensemble model is compared with the algorithms trained using the neural network. The results have proved that TensorFlow accuracy is much better when the parameters were tuned.

More data and characteristics relevant to coagulation (D-Dimer), salts (Potassium), and gender information should be used to train the model to acquire the best-fit model for our class of interest, which would then be used to aid a broader population. Prioritization and future medical decision-making can be aided by the model's output. SARS-CoV-2 infection would then be detected using the model's binary indicator, likelihood measurement, and accuracy. A hospital's policy and input parameters can be adjusted on a regular basis, depending on the state of the healthcare system. It would be necessary to go back and re-train the model using the newly acquired data as well.

7.REFERENCES

[1] A. A. Ogunleye and W. Qing-Guo, "XGBoost model for chronic kidney disease diagnosis," IEEE/ACM transactions on computational biology and bioinformatics, 2019.

[2] K. Santosh, "AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross Population Train/Test Models on Multitudinal/Multimodal Data," Journal of Medical Systems, vol. 44, pp. 1-5, 2020.

[3] Vaishya R, et al. Artifcial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr. 2020;14(4):337–9.

[4] Banoei MM, Dinparastisaleh R, Zadeh AV, et al. Machine-learning-based COVID-19 mortality prediction model and identifcation of patients at low and high risk of dying. Crit Care. 2021;25:328. https://doi.org/10.1186/ s13054-021-03749-5.

[5]Hassam Tahir; Annas Iftikhar; Mustehsan Mumraiz(2021),Forecasting COVID-19 via Registration Slips of Patients using ResNet-101 and Performance Analysis and Comparison of Prediction for COVID-19 using Faster R-CNN, Mask R-CNN, and ResNet-50

[6]Prerak Mann; Sahaj Jain; Saurabh Mittal; Aruna Bhat , 2021,Generation of COVID-19 Chest CT Scan Images using Generative Adversarial Networks

[7]Ryan Yixiang Wang; Tim Qinsong Guo; Leo Guanhua Li; Julia Yutian Jiao; Lena Yiqi Wang, (2020),Predictions of COVID-19 Infection Severity Based on Co-associations between the SNPs of Co-morbid Diseases and COVID-19 through Machine Learning of Genetic Data