A Best Effort Classification Model For Sars-Cov-2 Carriers Using Random Forest

Shrabani Mallick[1†], **Ashish Kumar Verma** [2††], **Dharmender Singh Kushwaha**[3††]

Dr. B. R. Ambedkar Institute of Technology, Port Blair, A & N Islands, India

Kendriya Vidyalaya-2, Port Blair, A & N Islands, India

Motilal Nehru National Institute of Technology Allahabad, Prayagraj, Uttar Pradesh, India

Summary
The whole world now is dealing with Coronavirus, and it has turned to be one of the most widespread and long-lived pandemics of our times. Reports reveal that the infectious disease has taken toll of the almost 80% of the world's population. Amidst a lot of research going on with regards to the prediction on growth and transmission through Symptomatic carriers of the virus, it can't be ignored that pre-symptomatic and asymptomatic carriers also play a crucial role in spreading the reach of the virus. Classification Algorithm has been widely used to classify different types of COVID-19 carriers ranging from simple feature-based classification to Convolutional Neural Networks (CNNs).

This research paper aims to present a novel technique using a Random Forest Machine learning algorithm with hyper-parameter tuning to classify different types COVID-19-carriers such that these carriers can be accurately characterized and hence dealt timely to contain the spread of the virus. The main idea for selecting Random Forest is that it works on the powerful concept of "the wisdom of crowd" which produces ensemble prediction. The results are quite convincing and the model records an accuracy score of 99.72 %. The results have been compared with the same dataset being subjected to K-Nearest Neighbour, logistic regression, support vector machine (SVM), and Decision Tree algorithms where the accuracy score has been recorded as 78.58%, 70.11%, 70.385,99% respectively, thus establishing the concreteness and suitability of our approach.

Key words:
SARS-CoV2, symptomatic, pre-symptomatic, Random Forests, Decision Tree, SVM.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing coronavirus disease 2019 (COVID-19) has reached a pandemic level. The general characteristics of people infected with the COVID-19 virus are mild to moderate respiratory illness and recover without requiring special treatment. Aged adults and those with critical medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness. This novel coronavirus is bizarre for many reasons, making its spread unpredictable and hard to control. Based on the visibility and severity of symptoms the carriers are classified into, Fig 1-

- Asymptomatic
- Pre-Symptomatic and
- Symptomatic

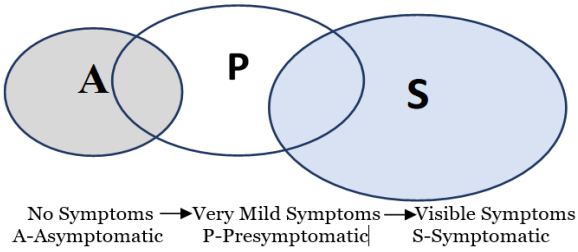

Fig 1: Venn Diagram of Corona affected Population (General Scenario)

Global Research shows that asymptomatic cases are those who do not have symptoms but are infected with a virus and may not develop any symptoms at all in future. However, pre-symptomatic is the phase when an individual is infected and maybe shedding virus but hasn't yet developed substantially visible symptoms, whereas symptomatic cases have readily visible symptoms and have been potentially tested positive. Essentially, the term asymptomatic is not associated with time, while pre-symptomatic is, which means if a COVID-19 test comes back positive and the patient does not have symptoms, the patient need not remain asymptomatic; the patient may develop symptoms in due course.

The Centers for Disease Control and Prevention (CDC) estimates that 35% of all people with COVID-19 are asymptomatic during that time, but says that those people are just as infectious as those with symptoms. The CDC also estimates that 40% of transmission happens before people feel sick. This forms the basis of our research work.
Cases who have acquired the virus infection and are pre-symptomatic with very mild to mild symptoms but are

Manuscript received January 5, 2021
Manuscript revised January 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.1.5

ignorant and have not been tested or exposed to medical intervention early may work as sleeper cells to spread the infection to a large population. The study, published in the journal Science, finds that about 4 in 5 people with confirmed coronavirus in China were likely infected by people who didn't know they had it.

Thus, various studies show that silent carriers of this killer virus have played a vital role in spreading the infection chain in which identification of such subjects is key to limit the spreading wave.

Machine Learning algorithms have been widely researched for the classification of COVID-19 carriers using feature-based classification algorithms like logistic regression, naïve bayes classifier, support vector machine, etc.

The key issue with these approaches is that these classifiers work by placing the data points on either side of the classifying hyperplane. There is no probabilistic explanation for the classification. Random forests are among the most popular machine learning methods due to their relatively good accuracy, robustness, and ease of use. The next section discusses some of the related research works.

2. Related Works

Unfortunately, there is not much research that differentiates asymptomatic COVID-19 cases versus pre-symptomatic cases. This paper presents a best-effort approach using Random Forest Classifier to classify different types of potential COVID carriers so that the infectivity potential of such carriers can be broken by isolating them.

Lalmuanawma et al. (2020) authors present a survey on recent studies that are based on AI and ML for screening, predicting, forecasting, contact tracing and drug development for COVID-19. Randhawa et al. [2020] identify the genomic signature of COVID-19 virus. They combine the supervised ML model with the digital signal processing (MLDSP) for genome analysis and Spearman's rank correlation coefficient analysis for result validation. This paper reports 100% accuracy in classification of COVID-19 virus sequence. In another work Elaziz et al. (2020), authors proposed a new ML-based method for classifying the chest x-ray images as COVID-19 patient or as non-COVID-19 person. It uses new Fractional Multichannel Exponent Moments (FrMEMs) for feature extraction from chest x-ray images and a modified Manta-Ray Foraging Optimization for most significant feature selection. The proposed method achives the 96.09% and 98.09% accuracy in two different datasets. The first dataset in Cohen et.al. [2020] is prepared by Joseph Paul Cohen and Paul Morrison and Lan Dao and second is COVID-19 dataset Andrea [2020]. The performance of the proposed approach is compared with CNN.

In another work Malki et al. (2020), authors are tried to explore the impact of weather/seasonal pattern on spreading COVID-19 virus for that they use various ML models to extract the relationship between various factors and spreading rate of COVID-19 virus. They conclude that temperature and humidity are important features for predicting the mortality rate of COVID-19 virus. It also establishes that higher temperature implies the lower number of infected cases. In Wang et. al. (2020), authors are tried to predict the trend of global epidemic of five countries i.e., Brazil, Russia, India, Peru and Indonesia and find out the global outbreak peak at the end of October with 14.12 million infected persons. This proposed forecasting methods is a hybrid of Logistic and Prophet model.

In Debnath et al. (2020), authors present a ML-based assist model for shared clinical decision making for COVID-19 patients. The model works between all medical personnel like, doctors and nurses; patients and their families. In Yadav et al. (2020) authors propose a support vector regression method for the analysis of five tasks related to the virus spread of COVID-19. These are (I) Predicting the spread of coronavirus across regions. (II) Analyzing the growth rates and the types of mitigation across countries. (III) Predicting how the epidemic will end. (IV) Analyzing the transmission rate of the virus. (V) Correlating the coronavirus and weather conditions. The aim of this work is to predict the level of spread, so that government and citizens can make a proper plan to handle the pandemic situation.

Another study Barstugan et al. (2020) presents early phase COVID-19 detection based on the abdominal CT images. It tries to identify the anomalies is CT images based on features specified by clinical experts. In Loey et al. (2020), authors present a hybrid of deep transfer learning model and classical ML methods for face mask detection for COVID-19 because wearing the facemask at public place is one of the effective and protective method from COVID-19 virus. The proposed model uses Resnet50 for feature extraction and decision tree, SVM and ensemble algorithm for classification. In this, three dataset are used for the experimental evaluation and it achives more then 99% accuracy.

Ardabili et al. (2020) presents a analysis of ML and soft computing models to predict the COVID outbreak. The analysis concludes that multi-layered perception (MLP) and

adaptive network-based fuzzy inference system (ANFIS) are showed their promising performance among the other comparable ones. This paper also provides an initail benchmarking to demonstrate the potential of ML for future research. Nemati et al. (2020), proposed a predictive model for predicting patients' stay in hospital because this time is very crucial for decision makers because it uses for planning and preparing the required infrastructure for COVID-19 hospitals. This work presents statistical methods and ML-based methods to predict the patient stay and discharge time in/from hospital. These statistical methods are Kaplan-Meier, CoxPH, Coxnet, Accelerated Failure Time; and ML methods are SVM, Stage wise Gradient Boosting and Component wise Gradient Boosting.

Tian et al. (2020) presents the performance comparison of 3 ML models i.e., hidden Markov chain model (HMM), hierarchical Bayes model, and long-short-term-memory model (LSTM) using the root-mean-square error (RMSE) and comprises that LSTM model has smallest prediction error rates. In Khanday et al. (2020), authors are presented ML-based approaches for COVID-19 detection based on clinical text data. It comprises that Logistic regression and Multinomial Naive Bayes algorithms has excellent performance than other ML algorithm with 94% precision, 96% recall, 95% f1 score and 96.2% accuracy.

The next section presents the proposed work that uses a best effort mechanism using Random Forest Based Classification powered by hyper-parameter tuning for classification of COVID-19 carriers.

3. Proposed Work

COVID-19 has been a bizarre since its onset. Believed to be originated from bats, several symptoms ranging from very mild to severe have been recorded in the human carriers of this virus. Some of them are fever, sore throat, cough, diahorrea, rashes, taste and visibility change, breathing difficulty and many more. Some carriers show a pattern matching behavior regarding the visibility of symptoms when the infection can be easily predicted, whereas in many cases, symptoms simply un-noticed who become the potential subset of a vulnerable population to spread the virus. Hence simple feature-based classification doesn't work well as it works by placing the data points on either side of the classifying hyperplane following a pattern of the feature values belonging to the domain of classes and hence there is no probabilistic explanation for the classification.

To overcome this problem, the proposed work presents an early phase classification of subjects using Random Forest Classifier. A Random Forest classifier works by creating a large no. of induvial decision trees that operate as an ensemble. Each individual tree in the random forest splits out a class prediction and the class with the most votes becomes our model's prediction as shown in Fig 2.

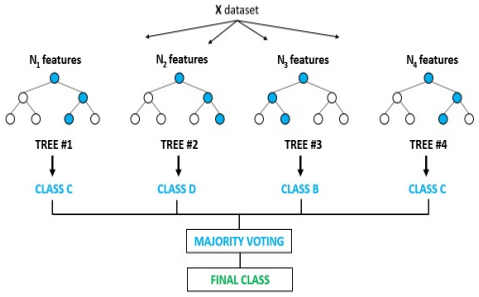

Fig. 2. Individual Classification Tree

In the present work, a sample population survey with 1080 subjects is considered having symptoms ranging from very mild to severe. Open access freely available datasets 'kaggle' has been used for the study. The various steps performed are Fig 3, shows the workflow of the proposed work:

i. Feature Extraction and Calculation of Feature Importance
ii. Hyperparameter tuning.
iii. Applying Random Forest Classifier for early phase classification of COVID-19 carriers given a potential suspected population-Asymptomatic, Pre-symptomatic and Symptomatic
iv. Comparing the results with other classifications algorithms – Logistic Regression and Support Vector Machine.
v.

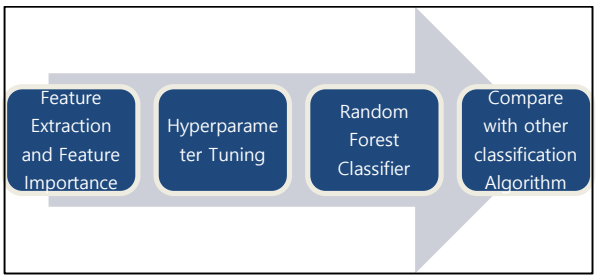

Fig. 3. Workflow of Proposed work

The subsequent subsections discuss each contribution in detail.

3.1 Feature Extraction and Calculation of Feature Importance

As per medical researchers, the specific symptoms of COVID-19 along with their occurrence percentage are in fever (98%) and cough (76%) in addition to other non-specific symptoms such as fatigue (44%), headache (8%), and dyspnea (3%) [7, 8, 9].

The Fig. 4, shows the histogram plot of the different symptoms seen in the subjects under study from the dataset. It is evident from the plot that almost 30% of the population is not showing any significant symptom which could be misleading such that they are either asymptomatic or pre-symptomatic.

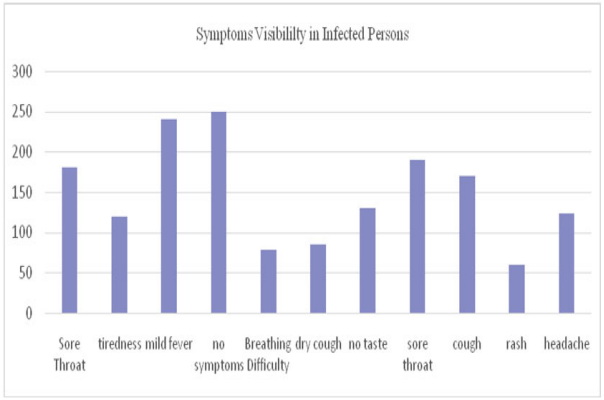

Fig. 4. Histogram plot Symptoms

Random Forests classifier works by creating multiple trees and choosing the class based on ensemble voting. Once the features are extracted, the calculation of Feature Importance is crucial Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The higher the value, the more important the feature. Assuming a binary tree with two child nodes, node importance for each decision tree is calculated using the formula:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (1)$$

ni_j=the importance of node j
w_j=weighted number of samples reaching node j
C_j=the impurity value of node j
$C_{left(j)}$=Child node from left split on node j
$C_{right(j)}$=Child node from right split on node j

The importance for each feature on a decision tree is then calculated as:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \quad (2)$$

fi_i= the importance of feature i
ni_j= the importance of node j

These values are then normalized to a value between 0 and 1 by dividing by the sum of all feature importance values. The final feature importance, at the Random Forest level, is its average over all the trees and is calculated by the formula:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \quad (3)$$

$RFfi_i$= the importance of feature i calculated from all trees in the Random Forest model
$Normfi_{ij}$= the normalized feature importance for i in tree j
T = total number of trees

3.2 Hyperparameter tuning.

The next step Optimizing hyperparameters for machine learning models is a key step in making accurate predictions. Hyperparameters define characteristics of the model that can impact model accuracy and computational efficiency. The following Hyperparameters have been considered for tuning.

n_estimators = number of trees in the foreset.
max_features = max number of features considered for splitting a node.
max_depth = max number of levels in each decision tree.
min_samples_split = min number of data points placed in a node before the node is split.

Cross-validation is used to determine the optimal values for hyperparameters so that it works best on untrained data. The optimum RF hyperparameters are determined to be n_estimator = 945, max_depth = 16, min_samp_split=2, min_samp_leaf=1, max_features=0.649.

3.3 Implementing Random Forest Classifier for classification

Machine Learning Algorithms presents a range of classification algorithms such as logistic regression, support vector machine, naive Bayes classifier, and decision trees. But one of the most competent and top classification models is the random forest classifier.

The main reason why random forest model works so well is -A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. The low correlation between

models is the key. The pre-requisites for a random forest to be accurate are –

There needs to be some actual signal in our features so that models built using those features do better than random guessing. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

In the instant classification problem, the correlation between the symptoms for each classification is very less, but the Feature matching and signalling in the classification domain is quite high hence model building is accurate. For example – Consider P(P1, P2, P3, P4, P6) as set of Individuals in the domain of classification and S (S1, S2, S3, S4, S5) is set of symptoms (feature variables) used for Classification, C (C1, C2, C3) be the set of classes:

Table 1. The planning and control components

S.No	Candidate from P	Symptoms seen	Class
1.	P1	S1, S2	C1
2.	P2	S1, S3	C2
3.	P3	S3, S4	C3
4.	P4	S1, S4,S5	C2
5.	P5	S2, S4	C3
6.	P6	S1, S5	C3

As evident from the above table, there is no strict correlation between the Symptoms seen each time for a classification. But if a large number of classifications are considered, then there is certainly a signal about the class and feature, as C3 generally has a symptom S4. S0 when the no. of trees is substantially large, the majority voting aspect levels out the effect of error in the predictions made by the individual trees.

With the usual classification algorithm like logistic regression, a strong correlation between the feature variables becomes very crucial for accurate results. On the other hand, using the Random Forest classification method, multiple classification trees are generated having minimum correlation, and then the final class is selected using majority voting among trees. This helps to refine the boundaries of the classification domain.

We have used a dataset of size 1283x26. The subset of data of size 1283x8, which comprises the symptom parameters and travel history is taken for the classification. The swarm-scatter plot in Fig 5 shows the distributions of the dataset population over various ages. It can be seen that a comprehensive selection of subjects belonging to all ages have been taken for the study to understand the problem in a realistic way.

The algorithm for the classification algorithm is as follows:

Feature Set:{mild fever , tiredness , no taste, sore throat, no symptoms, fever, dry cough , diff breath , sore throat, travel, rash, headache}
Class: {PRESYMTOMATIC = 0 FOR ASYMPTOMATIC = 1 FOR SYMPTOMATIC = 2}

```
a = 0
for i inrange(0, len(data)):
S = data["symptom"][i].split()
travel = data["Travel"][i]
   if ('0' inS):
      if ('6' inS) or ('8' in S) or ('4' in S):
         if travel == 1:
data["result"][i] = 0
eliftravel == 0 and ('4' in S):
data["result"][i] = 1
elif ('5' inS):
      if ('6' inS):
data["result"][i] = 2
   else:
symptoms = [0] * 10 + [2] * 4
choice = random.choice(symptoms)
data["result"][i] = choice
sym = []
pre = []
asym = []
for i in data['result']:
   if i == 0:
pre.append(1)
   if i == 1:
asym.append(1)
   if i == 2:
sym.append(1)
print(len(sym), len(pre), len(asym))
```

Fig 6 shows the Radom forest Classification Plot for the subjects in our dataset. The result shows that 33.6% of the population were pre-symptomatic, 56.4% were asymptomatic yet infected and 10% were clearly symptomatic in the considered dataset.

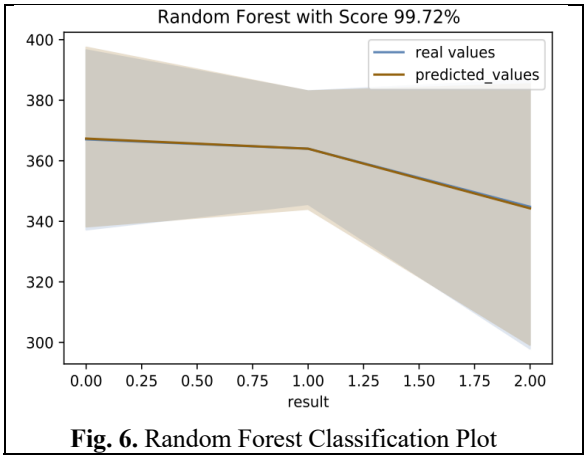

Fig. 6. Random Forest Classification Plot

The Fig. 7 is the confusion matrix plot which shows the confidence level of the algorithm. The model is established with an accuracy score 99.72%

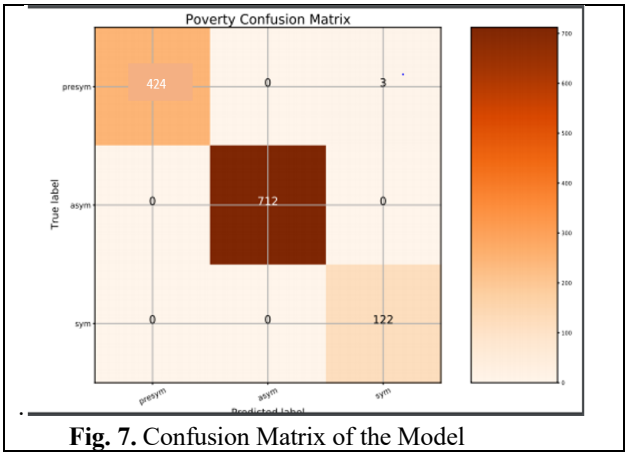

Fig. 7. Confusion Matrix of the Model

3.4 Comparing the results with other classifications algorithms.

The performance the proposed classifier has been compared with other algorithms. Fig 8 shows the performance scores of KNN, Logistic Regression, Support Vector Machines and Decision Tree. It is evident that decision tree score is very close to the Random Forest classifier, which This is because Random Forest Classifier is based on the decision tree-based classification.

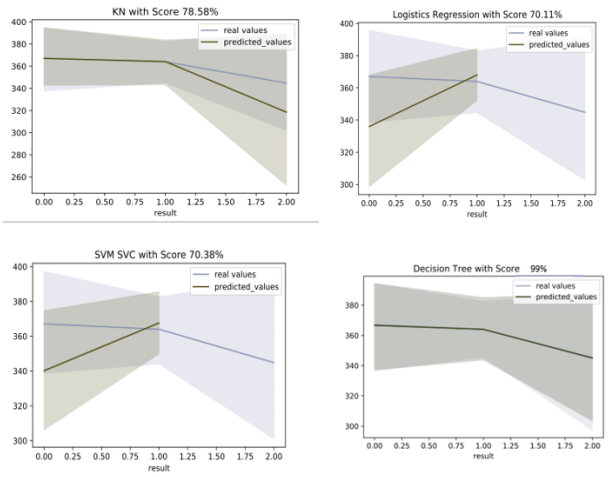

Fig. 8. Performance scores of KNN, Logistic Regression, Support Vector Machines and Decision Tree

7. Conclusion

Clinical rsearch and online reports on the ongoing pandemic reveal that more than 50% of the infection is spread by carriers who actually do not know that they are affected.

The aim of this research paper is to establish an accurate classification model to classify potential population under consideration to see that there are latent carriers in the population.

Although various machine learning approaches have been published to classify COVID-19 carriers but differentiating the pre-symptomatic from asymptomatic ones is a very judicious task because the identification boundary is not clear. In the present work, Random Forest Classification approach has been established as the most suitable model for classification as it tries to obtain a majority voting mechanism before finally assigning a subject to a particular class. The results have been compared with the two most widely used models, namely Logistic regression and Support Vector Machine. The results clearly show that Random Forest Model outweighs the other two approaches in terms for accuracy. Random Forest achieves an accuracy score of 99.6 % whereas Logistic Regression and Support Vector Machine based classification models record an accuracy score of 95.2% and 94.8%, respectively.

References

[1] Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*, 110059.

[2] Randhawa, G. S., Soltysiak, M. P., El Roz, H., de Souza, C. P., Hill, K. A., & Kari, L. (2020). Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *Plos one*, *15*(4), e0232391.

[3] Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., & Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of COVID-19. *Plos one*, *15*(6), e0235187.

[4] Cohen J. P., Morrison P., and Dao L., "COVID-19 image data collection," *arXiv preprint ar X iv*:2003.11597, 2020.

[5] D. A. L. Izzo Andrea. (2020, April-11-2020). Radiology. (2020). COVID-19 Database. Available: https://www.sirm.org/category/senza-categoria/covid-19/

[6] Malki, Z., Atlam, E. S., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., & Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos, Solitons & Fractals*, *138*, 110137.

[7] Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons & Fractals*, *139*, 110058.

[8] Debnath, S., Barnaby, D. P., Coppa, K., Makhnevich, A., Kim, E. J., Chatterjee, S., ... & Hirsch, J. S. (2020). Machine learning to assist clinical decision-making during the COVID-19 pandemic. *Bioelectronic medicine*, *6*(1), 1-8.

[9] Yadav, M., Perumal, M., & Srinivas, M. (2020). Analysis on novel coronavirus (COVID-19) using machine learning methods. *Chaos, Solitons & Fractals*, *139*, 110050.

[10] Barstugan, M., Ozkaya, U., & Ozturk, S. (2020). Coronavirus (covid-19) classification using ct images by machine learning methods. *arXiv preprint arXiv:2003.09424*.

[11] Loey, M., Manogaran, G., Taha, M. H. N., & Khalifa, N. E. M. (2020). A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. *Measurement*, *167*, 108288.

[12] Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., ... & Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. *Available at SSRN 3580188*.

[13] Nemati, M., Ansary, J., & Nemati, N. (2020). Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data. *Patterns*, *1*(5), 100074.

[14] Tian, Y., Luthra, I., & Zhang, X. (2020). Forecasting COVID-19 cases using machine learning models. *medRxiv*.

[15] Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Din, M. M. U. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology*, *12*(3), 731-739.

[16] https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html

[17] https://en.wikipedia.org/wiki/Coronavirus

[18] https://www.who.int/emergencies/diseases/novel-coronavirussdsd-2019