Predicting the outcomes of COVID-19 infections using easily accessible features

Kevin Roice
Department of Computer Science
Durham University
United Kingdom
kevin.roice@durham.ac.uk

***Abstract*—One of the many challenges posed by the COVID-19 pandemic is the difficulty in protecting the most vulnerable and gauging the risk associated with a positive test. This work develops, tunes and tests three predictive models that assess whether an individual survives an infection based on their demographic, geographic and historic data. All three models were capable of predicting the outcome of an infection with 96 to 97% accuracy on infections from the first 6 months of 2020. This was accomplished without any further medical testing such as oxygen levels or platelet counts.**

***Keywords*—COVID-19, Machine Learning, Epidemiology**

I. Introduction

In December 2019, the SARS-CoV-2 coronavirus (COVID-19) originated in Wuhan, China and claimed hundreds of lives within its first month [1]. Over the course of 2020, the COVID-19 pandemic brought social and economic impediments to nations across the globe. One of the main hindrances to combatting the pandemic is the difficulty in efficiently distributing the limited medical and human resources to the vulnerable in a manner that minimizes casualties [2,3]. Besides containing transmissions, the economic impact of repeated lockdowns [4] has made many governments hesitant to keeping the majority of their population indoors to curb transmission to the vulnerable. It is evident that the identification and protection of groups of society who are most adversely affected by the virus remains a challenge to governments across the globe [5].

Studies into various correlations and risk factor identifications have confidently identified attributes such as age and sex to be linked with recovery and mortality rates [6, 7]. The possibility of building upon these identified risk factors and exploiting them for the prudent allocation of resources has become a topic of interest for governments and healthcare systems from local to international scales. Identifying and isolating the vulnerable has proven to be far more effective and less socially and economically detrimental than national lockdowns [8].

In this work, I investigate the use of various machine learning (ML) algorithms to predict the outcome of a COVID-19 infection, using readily available data about the individual. In particular, this work aims to answer the research question:

Can survival from a COVID-19 infection be accurately predicted from demographic, geographic and historic data on an individual?

Specifically, three predictive ML models are presented and analysed in terms of their performance in predicting the final outcome of an individual diagnosed with COVID-19. An accurate prediction using such readily available data would facilitate the rapid identification of vulnerable groups in society without the need for further medical tests (such as oxygen levels [9] or platelet counts [10]) in hospitals. This would allow the efficient distribution of resources to the high priority, as well as the strategic placements of shielding, lockdowns, and vaccination prioritisation.

II. Methods

This work analysed an open access epidemiological dataset [11] containing demographic, medical, geographic and travel information along with key dates on positive COVID-19 cases. This section describes the full experimental procedure used to build the outcome predictors from the dataset. To explore and identify the factors influencing outcome, the Pearson correlation coefficients of various attributes were calculated before data preparation and feature selection (Fig 1). It is worth noting that Fig 1 is just a measure of linear correlation used to initially explore the dataset.

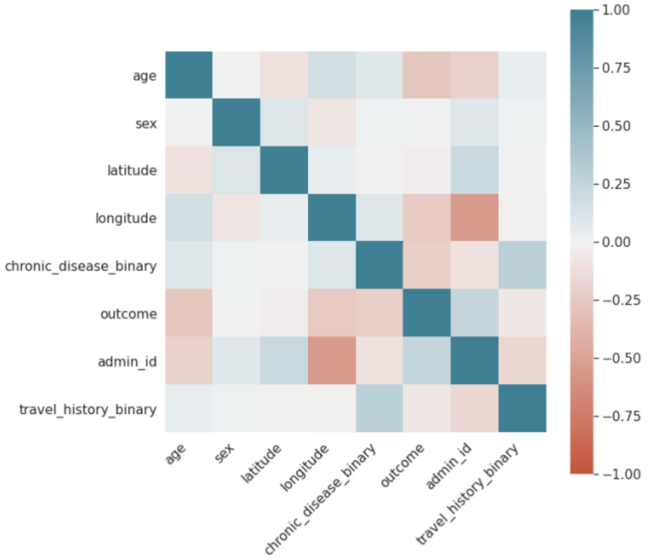

Figure 1: Pearson correlation coefficient heatmap of various features from the epidemiological dataset.

A. Data Cleaning and Preparation

Firstly, the dataset was removed of entries with no age, sex or outcome attributes, as previous works [6, 7] have shown these features play an active role in forecasting survival rates after a COVID-19 infection. Moreover, various statistics calculated in this work (outlined later) have supported the underlying relationship between age and outcome. While this reduces the dataset from 2.7 million individuals to approximately 33,300, these features were necessary attributes for the accurate and realistic use of the models. All of these entries had the location of the test (longitude and latitude) as well as whether or not the individual previously had a chronic disease. Chronic disease history is important for predicting infection outcomes since conditions like diabetes mellitus, chronic lung disease and cardiovascular disease greatly increase an individual's risk [12]. 0.36% of the remaining dataset had underlying chronic diseases (only 0.007% of the initial dataset had chronic diseases). 65% of these diseases induced a higher risk after infection according to [12].

This left a subset where 99.5% of entries were from India, Philippines, Ethiopia, Singapore and China. Since human-human contact is one of the main media for COVID-19 transmission [13]. Infection data from these densely populated countries benefited the predictive models' understanding of transmission trends to vulnerable groups [14].

Next, variables were converted to appropriate data types for the ML models. Age was converted to integers, and any entries that were age ranges were discarded. This did not hinder the accuracy of the models because age ranges made a small subset of the entries being studied (only 0.05% of the 300,000 entries with an outcome had their age entered as an age range). Outcome (label encoded for the supervised predictors), chronic disease and travel history were converted to binary features. I chose not to convert sex to a binary feature as assigning a value of 1 to one sex and 0 to another may make the model interpret sex biasedly as the presence/absence of a characteristic. Instead, a one-hot encoding was used for the sex feature in the categorical feature pipeline. This represented each sex as an unbiased group identifier, capable of accommodating multiple genders.

Similarly, the numerical feature pipeline used standardisation feature scaling for all numerical features (age, latitude and longitude). Experimentation revealed standardisation led to larger improvements in accuracy and recall compared to min-max scaling. Feature scaling is also useful in helping Gradient Descent algorithms converge [15].

B. Feature Selection

In contrast to previous ML models on COVID-19 infections, this work aims to accurately predict survival from an infection using data that is readily available to the individual, without the need for any medical check-ups or additional tests [9, 10, 16]. This includes features such as age, sex, location, travel history and disease history. However, from an algorithmic perspective, the selected features must be relevant and important to the ML model employed.

The correlation between age and outcome in Fig 1 along with existing studies between these attributes [6,7] compelled me to calculate the biserial correlation (a measure of correlation between a binary and continuous variable) between age and outcome. This was followed by a Kogolomorov-Smirnov test and a Wilcoxon test to identify and measure the distribution of age with either outcome (the dataset was too large for a Shapiro-Wilk test for normality).

With regards to geographic data, I decided to use longitude and latitude features instead of the country of testing. This is because using numerical features for location allows the model to interpolate to regions between those in the dataset. Choosing the categorical country feature would have led to large one-hot encoding matrices, and models unable to work in countries not included in the dataset. Additionally, the resolution of latitudes and longitudes gives a more accurate distribution of COVID-19 waves since different parts of a nation can be at different risks. This would allow the models' predictions to help governments decide local lockdowns.

While spatial features have been accounted for, it would have been ideal to additionally incorporate temporal features (such as the date of symptoms, case confirmation, hospital admission). The dataset did not have enough entries on these attributes. In particular, 89.8%, 97% and 99.6% of the entries in the subset had no dates of symptoms, discharge and hospitalisation respectively. Although only 0.1% of the confirmation dates were missing, the dates only spanned from 6th January 2020 to 3rd June 2020. There was not a full year's worth of data so no amount of imputation would help the models analyse risk from annual or seasonal infection waves.

C. Constructing Machine Learning models

After analysing and selecting appropriate features, several competent classification algorithms were trialled for use as predictive models. In this work, the Support Vector Machine (SVM), Random Forests (RF) and Stochastic Gradient Descent (SGD) classifiers were employed as predictive models. I was interested to see the responses from classifiers of different complexities to the same COVID-19 dataset.

SVM is a renowned classification model that aims to find an optimal hyperplane separating differently labelled data points in a feature space by using nearby data points called feature vectors [17]. This was thought to be advantageous for this project since hyperplanes would work well at separating the two outcome labels in higher dimensional data.

RF classifiers are ensemble learning models as they consist of a large, parametrised number of individual decision trees [18]. The importance of the features for estimating outcome was averaged across all the decision trees in the forest. The mean and associated standard deviations of decreases in Gini impurity for each feature is shown in Fig 2. It is interesting to note how, contrary to existing research [6, 7], sex is of relatively low importance according to the decrease in node impurity of the RF classifier.

SGD performs classifications by iterating over the parameter space of a random data point to minimise its cost function [19]. It was chosen in favour of Batch Gradient Descent due to its quick run times on large datasets.

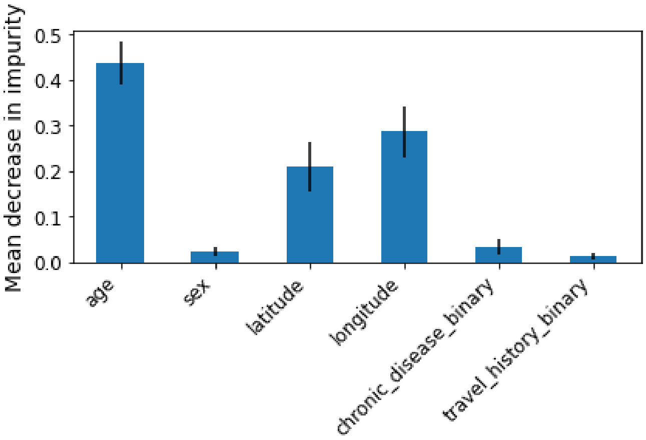

Figure 2: The mean decrease in Gini impurity for the features used by the RF classifier. Error bars indicate the standard deviation in the calculated mean node impurities.

For all three models, a randomised search 5-fold cross-validation was used to tune the hyperparameters over randomly sampled values from the distribution associated with each hyperparameter. This was found to be much quicker than grid search due to the number of hyperparameters involved. The tuned model was then compared to the default scikit-learn model [20], before testing on the test set. The three predictive models were tested on a 70% training 30% testing stratified split, and then evaluated based on accuracy, sensitivity, specificity and F1-score.

III. Results

As expected from existing literature, a point biserial correlation coefficient of $r_{pb} = -0.27$ (p-value $< 10^{-15}$) was found between age and outcome. The results of the Kogolomorov-Smirnov test strongly suggested that the age for both surviving and deceased individuals were not normally distributed (D = 0.988 and 0.987 for surviving and deceased groups respectively, with a p-value $< 10^{-15}$). The lack of normality prompted a Wilcoxon signed-rank test to identify differences in the mean ages of survivors and deceased. The enormously high Wilcoxon result of $W = 38739.5$ with p-value $= 2.2 \times 10^{-117}$, is very strong evidence that younger individuals are more likely to survive an infection than the elderly.

It is apparent that all three predictive models achieved very similar results on the testing set after hyperparameter tuning. Fig 3 illustrates this finding using 8 performance metrics: Accuracy, Recall/True Positive Rate, False Positive Rate, Specificity/True Negative Rate, Precision/Positive Predictive Value, Negative Predictive Value, F1 Score and Area Under Curve. The exact values of the key metrics are discussed in Table 1. Prediction accuracies for all models never fell below 96%, and specificity was the weakest point across all models. The most noticeable spread in performance happens at Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, and Negative Predictive Value (NPV). While the high AUC for RF can be attributed to its ensemble learning method, the higher NPV for SGD was unexpected. All three models showed near perfect recall scores.

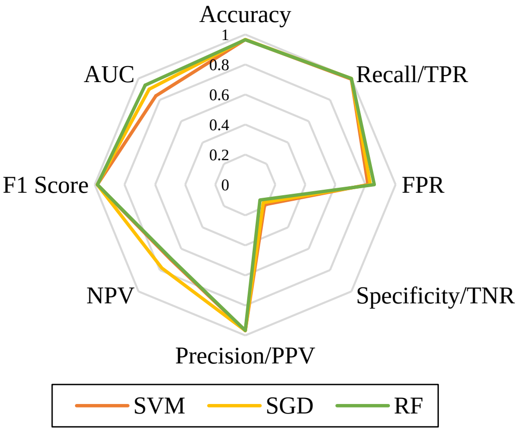

Figure 3: Performance of all 3 predictive models after tuning.

The impact of hyperparameter tuning was also measured. Table 1 compares the default scikit-learn classifiers [20] with the randomised searched 5-fold cross-validation classifiers (tuned model) after hyperparameter tuning in terms of accuracy, precision, recall and F1 scores. Tuning each model led to increases in at least one of these metrics, with the RF and SGD classifiers experiencing no reduction in performance. All classifiers showed performance improvements within the same order of magnitude.

Model	Accuracy	Precision	Recall	F1 Score
Def. SVM	96.6%	96.8%	99.8%	98.2%
Tuned SVM	96.6%	96.9%	99.7%	98.3%
Def. RF	96.6%	96.9%	99.6%	98.2%
Tuned RF	96.7%	96.9%	99.8%	98.3%
Def. SGD	96.4%	96.4%	99.9%	98.2%
Tuned SGD	96.6%	96.8%	99.9%	98.3%

Table 1: Key performance metrics from running default and tuned predictive models on the test set as percentages to 1d.p.

According to Fig 3, the three predictive models differed the most in their AUC values. Fig 4 shows the ROC curves of the models, to better understand this spread and illustrates the trade-off between recall (TPR) and specificity (1-FPR).

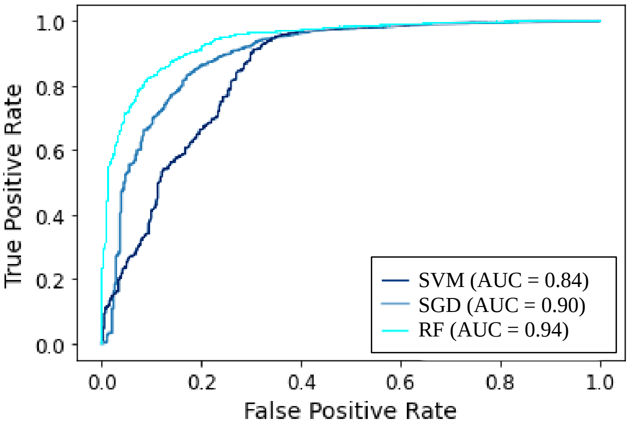

Figure 4: ROC curves for the three predictive models on the testing set after tuning hyperparameters.

IV. Discussion

The results from this work show a promising accuracy in all three models in predicting the outcome of a COVID-19 infection. The results support the claim that easily accessible features are enough to accurately identify vulnerable groups without further medical testing for individuals at the right place at the right time.

A. *Chosen Models*

Three renowned classifiers were used to independently predict the outcome of an infection. The models were deliberately chosen to differ in algorithmic complexities. Interestingly, the SGD classifier – commonly regarded as one of the computationally easier classification algorithms [21] – managed to outperform the SVM in terms of AUC and even the RF ensemble classifier in terms of NPV (Fig 3)!

The RF classifier is was found to best optimize the trade-off between recall and specificity (Fig 4), while the SVM performed the worst in this respect. This is in spite of SVM theoretically being able to work well with higher dimensional data. Furthermore, despite the low importance the RF classifier gave to sex (Fig 2), it was still able to produce accurate results with the highest AUC. The use of Gini impurity to determine feature importance is not unheard of in medical risk predictions and is even used to work with highly imbalanced data sets [22].

Overall, the significantly higher AUC and ensemble learning method of the final RF classifier puts it in the best position to make real world predictions in the countries that dominated the subset used. These high results with successful predictions were achieved without using any data on oxygen levels [9], platelet counts [10] or other medical tests typically used [23]. However, the longevity of any of the models is not guaranteed by any of the metrics in this work, due to possible mutations outlined in the limitations subsection.

B. *Experimental Procedure*

None of the models showed an accuracy less than 96%, and this can be attributed to the significance of the selected features in influencing the outcome of an infection since multiple cross-validation techniques (i.e. both in the default and tuned models) were deliberately used to prevent overfitting. The effectiveness of using features such as age and sex and disease history can be attributed to the epidemiology of COVID-19 as a respiratory virus and the general biology behind immune responses [6,7]. It is interesting to note that, although Fig 1 did not initially reveal high linear correlations between any of the features and outcome, it is evident that there is a strong non-linear correlation between the features used as revealed by statistical tests.

Although it was unexpected to measure non-gaussian age distributions for surviving and deceased individuals, this work found a clear difference in outcome between the elderly and youth (evidenced by the high Wilcoxon statistic of 38739.5).

Data pre-processing is a vital component of all three models. Since all three models were supervised learners, the outcome and its label encoding carry great importance. Similarly, the age and sex features were shown to have a relatively significant correlation with outcome (Fig 1), so it is crucial that the models receive a balanced, representative sample during training, validation and testing – which was done using a 5-fold cross-validation and stratified sampling.

The experimental procedure itself did not use many categorical features. However, the chronic disease history feature was manually sifted through to gauge the relevance of chronic disease binary. It may have been useful to incorporate or impute this feature if it had more valid entries. The feature scaling (both the standardisation and the one-hot encoding) was found to be extremely helpful since it decreased run time of randomised search cross-validation by many folds compared to the unscaled feature vectors. Additionally, the randomised searches for hyperparameter tuning were set to improve the accuracy score, which explains why accuracy never decreased for any of the models after tuning (Table 1).

Before tuning, various other features such as date of confirmation, geographic resolution and symptoms were experimented with to measure their relevance to outcome. Ultimately, the features shown in Fig 2 were chosen for two main reasons. Firstly, they were found to be particularly relevant to the RF classifier as indicated by their strong association to prediction results in their Gini impurities. Such an analysis and deduction based on prediction outcome is not unheard of in medical risk prediction [22]. Secondly, they would have caused an unjustifiable amount of data loss, giving a highly imbalanced subset of the data to work with.

The parameter space used for the randomised search cross validation was also heavily experimented with for each model. The chosen parameter space was found after trial and error to isolate the right combination of parameter ranges. Although a randomised search was implemented for this work, a combination of grid search and randomised search was used during experimentation to optimise the parameter ranges.

C. Limitations

A major limitation of all three predictive models is that they cannot cope with any time evolution of COVID-19. This means if the virus were to mutate to a variant affecting different vulnerable groups, the accuracy of the model will drop. As outlined previously, the subset with crucial information on age and sex lacked most of the key dates. Although there was a wealth of data on confirmation dates, I chose not to use them as they would hinder the generalisability of the model – the models would have only worked for infections from January to June, and extrapolation for annual/seasonal waves would have led to inaccuracies unless more data throughout the year was collected. This restricts the accuracy of all models to time periods similar to the dataset.

Although the densely populated countries used were ideal for learning trends in virus transmission [13], it would have been helpful to collect data from more countries. Currently, the classifiers only learnt from certain regions of the world (in terms of latitude and longitude). There are many locations in the world where the classifier extrapolates to make a prediction. This could lead to inaccuracies in real-world use. However, the choice of determining location using coordinates rather than country name designed all three models to be generalisable in future iterations to all locations on the globe, if provided with such location training data.

While the model prides itself in only relying on easily accessible features, this leaves it susceptible to misinterpreting the importance of features. For example, since the models only check for the presence of a chronic disease in an individual, all chronic diseases are taken to be equally lethal. In reality, only chronic conditions such as diabetes mellitus, chronic lung disease or cardiovascular disease have been known to increase the risk of fatality [12]. Likewise, the data cleaning for the outcome label stunts the generalisability of the model. At the moment, each unique entry of the outcome has to be manually interpreted as a surviving or deceased individual. This could be solved by either modifying future data collection to record the outcome as a binary feature or developing another ML model for the natural language processing to interpret keywords typed by the inputter!

V. Conclusion

This work has shown that accessible information on an individual can indeed be used to accurately predict the outcome of a COVID-19 infection. All three predictive models amplify the accuracy of risk factors from the literature and fully utilize them to predict the outcome.

We have shown that both complex and computationally light ML classifiers are capable of giving powerful predictions, indicating that prediction capability is more strongly dependent on the quality of data and epidemiology of the virus than the classification algorithm. It was also learnt that ensemble techniques are extremely powerful when working with large messy real-world datasets as they work on the consensus of multiple classifiers. I understood that other ensemble techniques such as XGBoost and Gradient Boosting Machine are worth exploring in future works.

Moreover, it was apparent that the statistical analysis of features before selection plays a huge role in the outcome. The majority of the good results achieved in this work can be attributed to prudent feature selection and correlation analysis. Furthermore, structured experimentation was found to be a much better tool at analysing underlying correlations than simply measuring linear correlation coefficients (which massively underplayed the significance of features like age that were later found to be crucial).

Therefore, the results from this work support the use of predictive models for assessing outcomes and risk in the short term localised COVID-19 waves corresponding to the data set, but the larger scale use is strongly dependent on the availability of more geographically and chronologically diverse data – which the current models are capable of generalising to by design.

References

[1] Liu Y et al., What are the underlying transmission patterns of COVID-19 outbreak? – an age-specific social contact characterization. *E-Clinical-Medicine* 22:100354 (2020)

[2] Worby C. J. and Chang H. H., Face mask use in the general population and optimal resource allocation during the COVID-19 pandemic. *Nature communications* 11.1: 1-9 (2020)

[3] Lee A., Wuhan novel coronavirus (COVID-19): why global control is challenging? *Public health* 179 (2020)

[4] Milliken A. et al., Addressing Challenges Associated with Operationalizing a Crisis Standards of Care Protocol for the Covid-19 Pandemic. *NEJM Catalyst Innovations in Care Delivery* 1.4 (2020)

[5] Inoue H. and Yasuyuki T., The propagation of economic impacts through supply chains: The case of a mega-city lockdown to prevent the spread of COVID-19. *PloS one* 15.9 (2020)

[6] Palaiodimos L. et al., Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the Bronx, New York. *Metabolism* 108, 154262 (2020)

[7] Takahashi T. et al., Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* 588.7837 pp.315-320 (2020)

[8] Karatayev V. A., Anand M. and Bauch C. T., Local lockdowns outperform global lockdown on the far side of the COVID-19 epidemic curve, *Proceedings of the National Academy of Sciences* 117.39 (2020)

[9] Tobin M. J., Basing Respiratory Management of COVID-19 on physiological principles, *American Thoracic Society Journals*, pp.1319-1320 (2020)

[10] Liu Y. et al., Association between platelet parameters and mortality in coronavirus disease 2019: retrospective cohort study, *Platelets* 31.4 pp. 490-496 (2020)

[11] Xu B., et al., Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific Data* 7.1 (2020)

[12] CDC COVID-19 Response Team et al. Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019 – United States Feb 12 – Mar 28 2020, *Morbidity and Mortality Weekly Report* 69.13 pp.382-386 (2020)

[13] Kucharski A. J. et al., Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The lancet infectious diseases* 20.5 pp. 553-558 (2020)

[14] Khakharia A. et al., Outbreak prediction of COVID-19 for dense and populated countries using machine learning. *Annals of Data Science* 8.1 pp. 1-19 (2021)

[15] Géron A., Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. *O'Reilly Media* (2019)

[16] Kang H. et al., Retest positive for SARS-CoV-2 RNA of “recovered” patients with COVID-19: Persistence, sampling issues, or re-infection? *Journal of Medical Virology* 92.11 pp. 2263-2265 (2020)

[17] Cortes C. and Vapnik V., Support vector machine. *Machine Learning* 20.3 pp. 273-297 (1995)

[18] Schapire R. E. and Freund Y., Boosting: Foundations and algorithms. *Kybernetes* (2013)

[19] Ruder S. An overview of gradient descent optimization algorithms. *arXiv preprint* arXiv:1609.04747 (2016)

[20] Pedregosa F. et al., Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830 (2011)

[21] Bottou L. and Bousquet O., The Tradeoffs of Large Scale Learning, *Optimization for Machine Learning*, MIT Press, pp.351-368 (2012)

[22] Khalilia M., Chakraborty S. and Popescu M., Predicting disease risk from highly imbalanced data using random forest. *BMC medical informatics and decision making* 11, no.1 pp. 1-13 (2011)

[23] Kermali M. et al., The role of biomarkers in diagnosis of COVID-19 – A systematic review. *Life sciences* 117788 (2020)