MACHINE LEARNING APPLICATIONS IN FORECASTING OF COVID-19 BASED ON PATIENTS' INDIVIDUAL SYMPTOMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the COVID-19 outbreak has been studied by many researchers in recent years. Many machine learning models have been used for the prediction of the transmission in a country or region, but few studies aim to predict whether an individual has been infected by COVID-19. However, due to the gravity of this global pandemic, prediction at an individual level is critical. The objective of this paper is to predict if an individual has COVID-19 based on the symptoms and features. The prediction results can help the government better allocate the medical resources during this pandemic. Data of this study was taken on June 18th from the Israeli Ministry of Health on COVID-19. The purpose of this study is to compare and analyze different models, which are Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayesian (NB), Decision Tree (DT), Random Forest (RF) and Neural Network (NN).
Keywords: COVID-19, Machine learning, Classification

1 INTRODUCTION

So far, the outbreak of the novel coronavirus disease (COVID-19) from the Hubei province of the People's Republic of China has threatened humans in more than 180 countries around the world since December 2019. Based on the latest statistics, more than 4,081,566 people have died from COVID-19 and more than 188,479,408 people have been diagnosed to be infected with COVID-19 (Dong et al., 2020). As the global pandemic prevails, reverse transcriptase polymerase chain reaction (RT-PCR), the main medical measure used in diagnosing COVID-19, faces the burden of insufficiency, especially in developing countries. That results in an increase of infection rates due to a failure to administer preventative measures during the time of diagnosis. Therefore, rather than using RT-PCR, the prediction of COVID-19 based on clinical symptoms will give an immediate diagnosis and thus mitigate the burden on resources to some extent. Although the results are not medical advice, they can be helpful for people to evaluate their risk of having COVID-19, and this will be beneficial for those countries that are suffering from a slow diagnosing time.

In various fields of practical medicine, AI has been proved to be effective. Reasonably, many researchers have tried plenty of AI technologies such as machine learning or deep learning in order to combat COVID-19. In related research, AI has been applied to early detection and diagnosis of infection, forecast of the spread of the infection, the development of drugs and vaccines, and the monitoring of patient treatment (Enughwure & Febaide, 2020). For instance, Ozturk et al. (2020) applied DarkCovid deep learning to test the possibility of automatically detecting COVID-19 cases using X-ray images, Zoabi et al. (2021) used the decision tree model to predict the infected cases based on clinical symptoms and demographic features in Israel and Ong et al. (2020) applied the Vaxign RV and Vaxign-ML approaches as machine learning models to predict COVID-19 protein candidates for vaccine development (Ozturk et al., 2020).

According to the exploration above, our research tries to use six models, including Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayesian (NB), Decision Tree (DT), Random Forest (RF) and Neural Network (NN), to predict the diagnosis based on five common clinical symptoms of COVID-19: cough, fever, sore throat, shortness of breath, and headache. Moreover, our models also consider additional influencing information such as gender and ages (above 60 and

below 60). This is because obvious bias on the rate of infections and mortality has been observed for this additional information.

Our experiment involves the evaluation of six machine learning models based on COVID-19 data from Israel with different symptoms. We compared the accuracies of six models test datasets, and tried to find the best model. Through comparison, we hope to further improve the model. This may make a certain contribution to alleviating the global shortage of covid19 medical resources in the field of machine learning.

2 RELATED WORK

Singh et al. (2020) use SVM to predict the number of cases, mortality, and recovery of the world wide population based on world health data. The authors claim that their model is capable of forecasting COVID-19 in individual cities if more information is available.

Guhathakurata et al. (2021) have done a comparative study with six models to forecast the severity of COVID-19 infection in patients and SVM outperformed the rest in all aspects, which is based on a dataset with 200 records with eight attributes.

Zhang et al. (2020) use univariable and multi- variable logistic regression to make a scoring system that can predict the severity of COVID-19 infection from clinical parameters based on data from 102 patients in Beijing You'an Hospital.

Hu et al. (2020). construct a severe COVID-19 risk model by multivariate logistic regression. According to clinical features and the transformation course of the virus, the authors also identify several independent early predictors from clinical features.

Zoabi et al. (2021) use gradient boosting decision tree to predict the corona results based on different kinds of clinical symptoms. Data of their research is a total of 51,831 people all from the Israeli ministry of health. Authors have addressed that if the limitation and bias of testing data is reduced, the performance of their model may be improved.

Philemon et al. (2019) use sampling and survey of epidemic forecasts based on ANN to compare with other methods reviewed. They conclude that ANN hybridized with a series of other algorithms and models, data transformation, and technology should be used for an epidemic forecast.

Mohammadi et al. (2021) build ANN and LR models to diagnose infected patients by COVID-19 based on 29 characteristics, symptoms and underlying diseases that were obtained from hospitalized patients. Data of research is from 6 provinces in Iran. Finally, the study demonstrated that ANN and LR models have a high ability in the diagnosis of COVID-19 infection.

Mansour et al. (2021)use Feature Correlated Naïve Bayes (FCNB) classification to accurately detect Covid-19 patients based on typical symptoms. But authors focus more on theories and hypothetical models, rather than using specific experimental data to verify.

Tiwari et al. (2021) apply the powerful machine learning algorithms, namely Naive Bayes, Support Vector Machine (SVM) and Linear Regression, on real time-series dataset. Naive Bayes produces promising results to predict Covid-19 future trends with smaller Mean Absolute Error (MAE) and Mean Squared Error (MSE), although the global footprint of this pandemic is still uncertain.

Prakash et al. (2020) claim that Random Forest Regressor and Random Forest Classifier outperformed other machine learning models like SVM, KNN+NCA, Decision Tree Classifier and Gaussian Naïve Bayesian Classifier. But there is a limitation in the author's experiment which is that they only consider the impact of age group differences on the incidence of COVID-19.

3 METHODOLOGY

3.1 DATA AVALIABILITY

Our data comes from the Israeli Ministry of Health on COVID-19. The data was taken on June 18th and consists of 4,704,597 individuals' data with 10 features, including test date, gender, ages, cough, fever, sore throat, shortness of breath, headache, corona result, and test indication.

3.2 DATA PREPROCESSING

The data preprocessing process is as follows. Firstly, we translate all Hebrew texts of all data features into English. Then, we delete irrelevant data features, such as test date and test indication, and 8 features are hence left. Secondly, for the gender column, we use 1 to represent male and 0 for female. Similar transformations have been made for other features, where 1 means under 60 years old and 0 is for above; Likewise, we have 1 for positive corona result, 0 for negative corona result, and 2 for other results. Next, we prepare the sample dataset by randomly selecting 5% of the samples from the population. We repeat this process to make five such sample datasets for experiments. Lastly, for each sample dataset, we randomly split them into training and testing sets by a proportion of 8: 2.

3.3 EXPERIMENT DESIGN

All models are built through Python, including Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayesian (NB), Decision Tree (DT), Random Forest (RF) and Neural Network (NN). The experiment is designed to have two main steps: tuning and optimization. For each model, we tune it to find the best parameters and fit the model to check its performance. The next step is optimization which can be divided into two sub-steps. First, we remove 1 data feature from the dataset and refit the model to check for any improvements. We repeat this process for each data feature (except the corona result), and summarize all the features that can result in improvements of accuracy after removal. Second, we remove 2 features selected from the summary of data features in the previous step. We refit the model and check for improvements after the removal of any combination of 2 features from the summary.

3.3.1 SUPPORT VECTOR MACHINE

Support vector machine, or SVM, is widely used for classification as it creates hyperplanes to separate dataset into different classes. It is very versatile since different kernels can be used as decision functions. We tune the model, and the tuned parameters are 'C' and 'gamma', where 'gamma' is the coefficient of the kernel function and 'C' is the regularization parameter. The kernel function is chosen to be 'rbf', which is the radial basis function. The best function is chosen by GridsearchCV and we use it as the baseline model for optimization after checking its performance.

3.3.2 LOGISTIC REGRESSION

Logistic regression is used since the target variable, corona result, is a categorical variable. The tuning variables are 'solver' and 'C', where 'C' is the inverse of regularization strength. The GridsearchCV then finds the best parameters, and we refit the model on the datasets as the baseline model for optimization.

3.3.3 DECISION TREE

Decision Tree, a supervised machine learning model, is commonly used for classification tasks. In this experiment, the result of tuning parameter combinations shows that as one of pruning strategies of the decision tree model, the parameter combination whose max depth is 6 and min sample split is 17 realizes the best local optimization. The set of the best parameters from GridSearchcv then builds the baseline model.

3.3.4 NEURAL NETWORKS

Neural Network (NNs) is a kind of relatively complicated deep learning through constructing artificial systems of neurons to process information without supervision. According to the result of GridSearchcv, the bag of parameters that the baseline model uses are 'relu' for activation, alpha equals 0.0001, hidden layer sizes is 300 and solver is 'lbfgs'. The set of parameters makes the baseline model achieve the local optimization.

3.3.5 NAIVE BAYES

Naive Bayes Classifier is a very simple and effective Classifier. One of the great advantages of Naive Bayes is that it is easy to tune the parameters, but the accuracy of the results is often great. In the experiment, we used Bernoulli Naive Bayes, Multinomial Naive Bayes and Gaussian Naive Bayes, among which Multinomial Naive Bayes only needs to tune the variables named alpha, Bernoulli Naive Bayes tunes alpha and binarize. During the experiment, we tried to obtain the best parameters and the highest prediction performance data by drawing the graph of the prediction performance of each model as a function of the parameters.

3.3.6 RANDOM FOREST

Random forest Classifier contains multiple decision trees, which is often used for unsupervised clustering. The main parameters of random forest include n_estimators, max_features, and max_depth. In this experiment, we build a random forest every 10 steps, set different parameter intervals, and use GridSearchcv to compare different prediction performances and output the best parameters.

4 EXPERIENCE RESULT

Performance of the models was evaluated by Precision (weighted), Recall (weighted), F1-scores and Accuracy. We fit the models on the five sample datasets and average results. The results are summarized into tables and graphs below.

Table 1: Weighted the average results of models on full datasets

	Accuracy	Precision	Recall	F1-score
SVM	0.897172	0.861554	0.897172	0.873671
Logistic Regression	0.895496	0.853599	0.895496	0.863057
Naïve Bayesian	0.893895	0.855825	0.893895	0.869076
Decision Tree	0.897058	0.860451	0.897058	0.872033
Random Forest	0.897285	0.859358	0.897285	0.870004
Neural Network	0.897225	0.862095	0.897225	0.874134

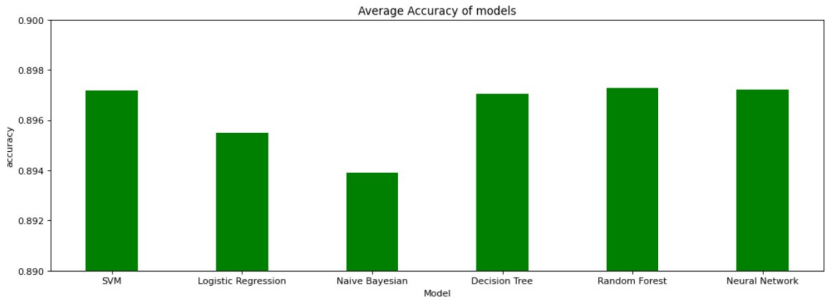

Figure 1: Accuracy results of models on full datasets

Table 1 and Figure 1 shows the results of six models on full datasets. Nonetheless, we notice that all our models fail to predict ‘other’ in our target, corona result. Thus, we decide to rerun all models based on a reduced dataset (‘other’ removed from corona_result). Below are the results.

Table 2: Weighted average results of models on reduced dataset, with 'other' removed from corona_result

	Accuracy	Precision	Recall	F1-score
SVM	0.913223	0.893833	0.913223	0.897974
Logistic Regression	0.912088	0.885824	0.912088	0.885945
Naïve Bayesian	0.913281	0.889455	0.910903	0.894439
Decision Tree	0.913281	0.893137	0.913281	0.896828
Random Forest	0.913479	0.891998	0.913479	0.894882
Neural Network	0.913899	0.893616	0.913899	0.896444

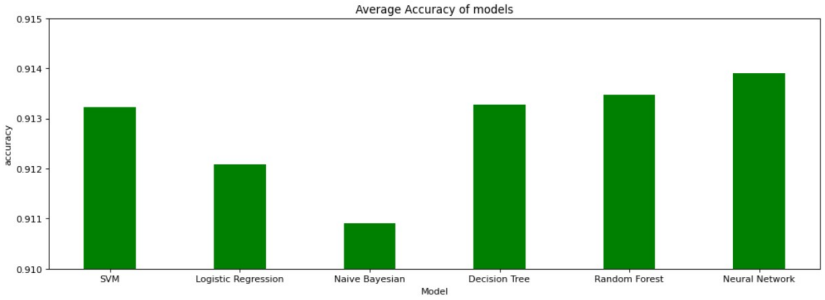

Figure 2: Accuracy results of models on reduced datasets, with 'other' removed from corona_result

As shown in Table 2 and Figure 2, all evaluation indicators improve significantly after we delete 'other' from corona_result. Therefore, we decide to use the reduced datasets in our follow-up experiments.

4.1 SUPPORT VECTOR MACHINE

Table 3: The experiment result of Support Vector Machine

C	0.1	variable	-

As shown in Table 3, we can see the optimal value regarding parameters 'gamma','kernel' and 'C', and the result of whether it is necessary to delete the variables. Its overall performances are as follows: 89.38%, 91.32%, 89.80%, 91.32% respectively. When dropping one variable, everyone gives an improvement in accuracy ranging from 0.03% to 0.1%, except for 'shortness of breath' that gives a reduction of 0.04%. So we retain the variables.

4.2 LOGISTIC REGRESSION

Table 4: The experiment result of Logistic Regression

	Best Parameters		Result
Solver	lbfgs	accuracy	91.21%
C	0.01	Drop variable	F
-	-	variable	-

As shown in Table 4, we can see the optimal value regarding parameters 'solver' and 'C', and the result of whether it is necessary to delete the variables. Its overall performances are as follows: 88.58%, 91.21%, 88.59%, 91.21% respectively. When dropping one variable, all variables give

improvements except for 'sore throat' and 'headache', although the highest improvement is only 0.1%. Removing 2 variables does not increase the improvement either, and the highest improvement we get is still 0.1%. Thus, we retain the baseline model.

4.3 DECISION TREE

Table 5: The experiment result of Decision Tree

	Best Parameters		Result
Max depth	6	accuracy	91.23%
Min samples leaf	17	Drop variable	F
-	-	variable	-

As shown in Table 5, we can see the optimal value regarding parameters 'Max_depth' and 'Min_samples_leaf', and the result of whether it is necessary to delete the variables. Its overall performances are as follows: 89.18%, 91.23%, 89.57%, 91.23% respectively. When dropping one variable, only when the variable 'sore throat' is deleted, the performance of the model does not worsen compared to the baseline model. Instead, its accuracy increases by 0.03% approximately. Therefore, deleting the variable sore throat makes the optimization better while the impact is extremely slight. As for the optimization model which is deleted 2 variables at the same time, when both 'cough' and 'sore throat' are deleted, the performance of the model will improve. Its accuracy increases by less than 0.004%. So deleting any variable combination almost makes no sense for optimizing the decision tree model.

4.4 NEURAL NETWORKS

Table 6: The experiment result of Neural Networks

	Best Parameters		Result
activation	relu	accuracy	91.32%
alpha	0.0001	Drop variable	F
hidden layer sizes	300	variable	-
solver	lbfgs	-	-

As shown in Table 6, we can see the optimal value regarding parameters 'activation', 'alpha', 'Hidden_layer_sizes' and 'solver', and the result of whether it is necessary to delete the variables. Its overall performances are as follows: 89.48%, 91.32%, 89.89%, 91.32% respectively. In this experiment, the enhancement of model performance is the most significant when the variable 'headache' is deleted. Its accuracy increases 0.05% approximately compared to the baseline model after deleting the variable. We notice that deleting both 'sore_throat' and 'headache' variables can bring the largest improvement of the model. Its accuracy increases 0.04% approximately compared with the baseline model. The extremely slight change is reasonable to be ignored in the experiment.

4.5 NAIVE BAYES

Its overall performances are as follows: 88.95%, 91.09%, 89.44%, 91.09% respectively. Another thing worthy of notice is to observe the curve of Naive Bayes' prediction performance with parameters.

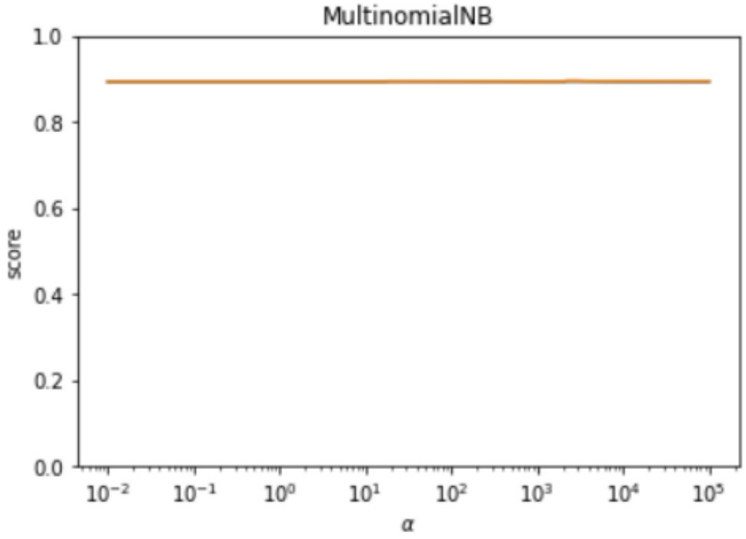

Figure 3: The influence curve of Multinomial Naive Bayes's prediction performance with the alpha parameter

Figure 3 shows that the size of the parameters does not affect the results of prediction performance.

The optimized data of different naive bayes models are not the same, and the optimized data of Gaussian Naive Bayes has the most significant change. The accuracy for the baseline model is 89.4%. When dropping two variables, cough and fever gives an improvement of 0.92%, which is even less than 1% , for accuracy. Therefore, it can be considered that deleting any variables has no effect on Naive Bayes.

4.6 RANDOM FOREST

Table 7: The experiment result of Random Forest

	Best Parameters		Result
Max depth	5	accuracy	91.35%
Max features	4	Drop variable	F
Number of estimators	50	variable	-

As shown in Table 7, we can see the optimal value regarding parameters 'Max_depth', 'Max_features' and 'Number of estimators', and the result of whether it is necessary to delete the variables. Its overall performances are as follows: 89.20%, 91.35%, 89.49%, 91.35% respectively. When optimizing the model, it was found that deleting gender and shortness_of_breath had the most significant effect, but it only increased by 0.049%. It can be concluded to maintain the baseline model.

Performance is similar for all models after tuning and optimization. The models yield similar results in terms of all scores after tuning. For accuracy of the models, we observe that 0.894 (Naive Bayesian) is the lowest and 0.897 (Random Forest) being the highest. Likewise, f1-scores are also close among the models, ranging from 0.863 (logistic regression) to 0.874 (Neural Network). In addition, we notice that all of our models fail to predict 'other' in the targets. We suspect that our models are giving results in floating numbers and hence none is classified into 'other' since they might be closer to '1' than '2'. Further experiments show that our models' predictions are either '0' and '1', and this anomaly may likely be due to the disproportionately small size of samples with 'other' for corona results. As a result, we decide to disregard samples of 'other' as corona results

and make this as a binary classification task. The binary classification results are still close among all models, and accuracies are within the range of 0.910 to 0.915 as graph 2 demonstrated above.

Also, it comes to our attention that deleting data features does not help improve the accuracy nor f1 score by a noticeable amount. The best improvement only increases the accuracy by around 0.1%. As a result, we decided to retain all the data features.

5 DISCUSSION

This research is not without shortcomings. Firstly, as we mentioned above, all models are not capable of predicting 'other' in corona results. Secondly, dataset features are not conclusive. For instance, symptoms such as chest pain and lack of smell or taste have been identified as potent predictors for the diagnosis of COVID-19. Unfortunately, they are not recorded by the Israeli Ministry of Health. Moreover, some critical information like symptoms' duration and contact with known patients is impossible to be recorded precisely. Last but not the least, this dataset's information is largely self-reported data, and this could potentially influence models' performance as well. Nevertheless, with a relatively large dataset for our research, the bias of data can be mitigated to a rather small extent.

6 CONCLUSION

From what is known so far, defeating COVID-19 completely will undoubtedly take lots of time. This will definitely put immense pressure on the medical resources of countries. Therefore, the prediction of COVID-19 based on symptoms could potentially help to save medical resources, and therefore is worthy of unremitting exploration. Seeing from the results, machine learning models such as Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayesian (NB), Decision Tree (DT), Random Forest (RF) and Neural Network (NN) are proved to be effective in the aspect of prediction. Inevitably, data from the Israeli Ministry of Health on COVID-19 is not comprehensive enough, and might be biased due to information gaps and the self-reported nature of that information. Unfortunately, 'other' in corona results cannot be predicted using those six models. However, the large enough sample size can offset some negative impacts from data bias to some extent. Moreover, all our models are robust and show excellent performance on random sample sets. With no doubt, data like ours have issues with balance since the number of 'other' in corona results is overwhelmingly small compared to the other two, and datasets on diseases like COVID-19 often suffer from this same problem. In the future, deeper exploration can be made profiting from our current results.

REFERENCES

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020. ISSN 1473-3099. doi: https://doi.org/10.1016/S1473-3099(20)30120-1. URL https://www.sciencedirect.com/science/article/pii/S1473309920301201.

Akpofure A Enughwure and Isaac C Febaide. Applications of artificial intelligence in combating covid-19: a systematic review. *Open Access Library Journal*, 7(8):1–12, 2020.

Soham Guhathakurata, Souvik Kundu, Arpita Chakraborty, and Jyoti Sekhar Banerjee. A novel approach to predict covid-19 using support vector machine. In *Data Science for COVID-19*, pp. 351–364. Elsevier, 2021.

Haifeng Hu, Hong Du, Jing Li, Yage Wang, Xiaoqing Wu, Chunfu Wang, Ye Zhang, Gufen Zhang, Yanyan Zhao, Wen Kang, et al. Early prediction and identification for severe patients during the pandemic of covid-19: a severe covid-19 risk model constructed by multivariate logistic regression analysis. *Journal of Global Health*, 10(2), 2020.

Nehal A Mansour, Ahmed I Saleh, Mahmoud Badawy, and Hesham A Ali. Accurate detection of covid-19 patients based on feature correlated naïve bayes (fcnb) classification strategy. *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–33, 2021.

Farzaneh Mohammadi, Hamidreza Pourzamani, Hossein Karimi, Maryam Mohammadi, Mohammad Mohammadi, Nahid Ardalan, Roya Khoshravesh, Hassan Pooresmaeil, Samaneh Shahabi, Mostafa Sabahi, et al. Artificial neural network and logistic regression modelling to characterize covid-19 infected patients in local areas of iran. *Biomedical journal*, 2021.

Edison Ong, Mei U Wong, Anthony Huffman, and Yongqun He. Covid-19 coronavirus vaccine design using reverse vaccinology and machine learning. *Frontiers in immunology*, 11:1581, 2020.

Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in biology and medicine*, 121:103792, 2020.

Manliura Datilo Philemon, Zuhaimy Ismail, and Jayeola Dare. A review of epidemic forecasting using artificial neural networks. *International Journal of Epidemiologic Research*, 6(3):132–143, 2019.

Kolla Bhanu Prakash, S Sagar Imambi, Mohammed Ismail, T Pavan Kumar, and YN Pawan. Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms. *International Journal*, 8(5), 2020.

Vijander Singh, Ramesh Chandra Poonia, Sandeep Kumar, Pranav Dass, Pankaj Agarwal, Vaibhav Bhatnagar, and Linesh Raja. Prediction of covid-19 corona virus pandemic based on time series data using support vector machine. *Journal of Discrete Mathematical Sciences and Cryptography*, 23(8):1583–1597, 2020.

Dimple Tiwari, Bhoopesh Singh Bhati, Fadi Al-Turjman, and Bharti Nagpal. Pandemic coronavirus disease (covid-19): World effects analysis and prediction using machine-learning techniques. *Expert Systems*, 2021.

Chi Zhang, Ling Qin, Kang Li, Qi Wang, Yan Zhao, Bin Xu, Lianchun Liang, Yanchao Dai, Yingmei Feng, Jianping Sun, et al. A novel scoring system for prediction of disease severity in covid-19. *Frontiers in cellular and infection microbiology*, 10:318, 2020.

Yazeed Zoabi, Shira Deri-Rozov, and Noam Shomron. Machine learning-based prediction of covid-19 diagnosis based on symptoms. *npj digital medicine*, 4(1):1–5, 2021.