A Review on the Paper, COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA

[Villavicencio, C.N.; Macrohon, J.J.E.; Inbaraj, X.A.; Jeng, J.-H.; Hsieh, J.-G. COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA. Algorithms 2021, 14, 201. https://doi.org/10.3390/a14070201]

Abstract

Early diagnosis is essential to prevent the development of a disease that can cause danger on human lives. COVID-19, which is an infectious disease that has mutated into several variants, has become a global pandemic that requires it to be diagnosed as soon as possible. Using technology, the available information about COVID-19 is increasing every day useful information from massive data can be extracted using data mining. Many models have been proposed which can predict the presence of coronavirus in a person and one such published paper consists of some models predicted by some machine learning algorithms and comparing which one of those algorithms produces the best model, the one having less bias and less variance. This paper is the review of the published paper which first gives an insight into the published paper, then goes on to see the loopholes in the paper and finally appreciates the paper for some noteworthy improvements over previous papers despite having some loopholes in it.

Introduction

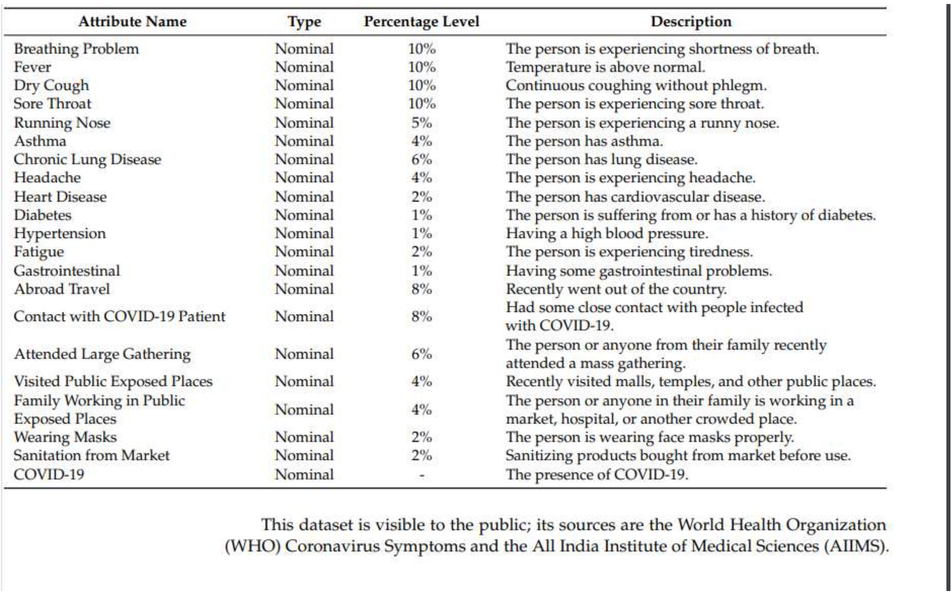
The Covid- 19 disease caused by the novel coronavirus which originated in Wuhan, China on December 2019 is a respiratory disease not limited to the respiratory system has caused havoc in the world causing thousands of deaths each day during its peak [1]. Prevention is better than cure , which means for a person to not get a disease, he/she must be away from the carrier of the disease, which are other people, any object to touch upon where the virus can survive if other people have had contact with that object, air, etc. in this case [2]. No one can be possibly ripped off the right of having to breathe air, so disease transmission from this medium cannot be prevented, people can follow sanitization measures if they get into contact with an infected surface to prevent themselves from getting this disease, and people can avoid crowds, gatherings and stay away from themselves to avoid the first carrier mentioned. It’s not always possible to avoid meeting other fellow people and we being inter-dependable upon one other have to meet and interact for our survival, which brings us to the wearing of masks by everyone which can slowdown transmission, but not all people can wear masks and masks are also not that effiecient to prevent transmission as social distancing, which brings us to the idea of separating only some group of individuals from a given environment who may be a carrier (Not all people are carriers, the virus must be present in one’s body to a significant extent for a person to become a carrier). For this we have to identify those individuals who are covid positive (who have this virus in their body) where the prediction models come handy. People who are covid positive can then be quarantined and separated from the rest of the people in the environment to slowdown and prevent the spread of this virus.

Insight onto the study

The paper first gives the statistics of covid- 19, the countries it has affected maximally, number of cases of recovery, number of cases, number of deaths, etc. Then it goes on to explain the symptoms of the disease, how the virus spreads, the parts of the body the disease affects, common prevention measures of covid- 19, how humanity is going through a failed attempt to control this disease as the number of cases everyday has been rising despite thorough human intervention through quarantine, vaccination and other measures. Then it says how prediction models help us in preventing the disease to a great extent. When enough medical support (technologies to identify the disease beforehand like blood testing, etc.) are not present, the presence of the disease is only known when the symptoms related to the disease are visible because of which early diagnosis of the disease is not possible. This is where prediction algorithms come into picture, which can be used to predict if a person has covid- 19 or not, using some attributes like the places where the person has went, if the person is present in a covid hazard zone, if the person has gone abroad, has sore throat, etc. The paper then turns its focus onto importance of machine learning models to predict the number of cases on a certain day and the various successes of these models to predict outbreaks, number of deaths, etc. in various countries, but very few models have been proposed to predict the possibility of a person getting infected, the drawback the paper intends to remove.

The machine learning algorithms J48 Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours and Naïve Bayes algorithms were applied through WEKA machine learning software. Each model was evaluated using 10 fold cross validation and compared according to major accuracy measures, correctly or incorrectly classified instances, kappa, mean absolute error, and time taken to build the model. The results show that Support Vector Machine using Pearson VII universal kernel outweighs other algorithms by attaining 98.81% accuracy and a mean absolute error of 0.012.

The attributes used for the algorithm is given below:



Attribute Name	Type	Percentage Level	Description
Breathing Problem	Nominal	10%	The person is experiencing shortness of breath.
Fever	Nominal	10%	Temperature is above normal.
Dry Cough	Nominal	10%	Continuous coughing without phlegm.
Sore Throat	Nominal	10%	The person is experiencing sore throat.
Running Nose	Nominal	5%	The person is experiencing a runny nose.
Asthma	Nominal	4%	The person has asthma.
Chronic Lung Disease	Nominal	6%	The person has lung disease.
Headache	Nominal	4%	The person is experiencing headache.
Heart Disease	Nominal	2%	The person has cardiovascular disease.
Diabetes	Nominal	1%	The person is suffering from or has a history of diabetes.
Hypertension	Nominal	1%	Having a high blood pressure.
Fatigue	Nominal	2%	The person is experiencing tiredness.
Gastrointestinal	Nominal	1%	Having some gastrointestinal problems.
Abroad Travel	Nominal	8%	Recently went out of the country.
Contact with COVID-19 Patient	Nominal	8%	Had some close contact with people infected with COVID-19.
Attended Large Gathering	Nominal	6%	The person or anyone from their family recently attended a mass gathering.
Visited Public Exposed Places	Nominal	4%	Recently visited malls, temples, and other public places.
Family Working in Public Exposed Places	Nominal	4%	The person or anyone in their family is working in a market, hospital, or another crowded place.
Wearing Masks	Nominal	2%	The person is wearing face masks properly.
Sanitation from Market	Nominal	2%	Sanitizing products bought from market before use.
COVID-19	Nominal	-	The presence of COVID-19.

This dataset is visible to the public; its sources are the World Health Organization (WHO) Coronavirus Symptoms and the All India Institute of Medical Sciences (AIIMS).

[1]

The research paper here uses not so complicated machine learning models to predict if a person has covid- 19 or not compared to this paper [3] which uses a deep learning algorithm called encoder which uses original patient features to get reconstructed patient features which the algorithm itself generates and is different from the original patient features which predict if a person having those particular features is covid- 19 susceptible or not. Although these provide accurate results, the accuracy of SVM which is 98.81% for this data which is more than enough, also that the cost of implementation of encoders are also very high, it is more plausible to with SVM's taking cost factors as well into consideration. In [4] a different dataset is used and it turns out that the naive- bayes classification turns out to be the best algorithm for the dataset, it actually depends which algorithm suits which dataset, one algorithm may produce the highest accuracy in one dataset while another may prove to be the best dataset for another algorithm.

Conclusion

Various models have come up to predict number of cases, severity of an outbreak, etc. quite successfully and accurately but there is always a limitation to any study, there isn't any study which can predict the possibility of occurrence of COVID- 19 at the individual level using the data of the individual tested upon. This study fulfils that missing piece and we found out that for the given dataset the SVM algorithm works the best producing an accuracy of 98.81%.

Citations and References

[1] Santosh, K.C. COVID-19 Prediction Models and Unexploited Data. *J Med Syst* **44**, 170 (2020). https://doi.org/10.1007/s10916-020-01645-z

[2] Wake, RM; Morgan, M; Choi, J; Winn, S (2020) ***Reducing nosocomial transmission of COVID-19: implementation of a COVID-19 triage system.*** Clin Med (Lond), 20 (5). e141-e145. ISSN 1473-4893

[3] Li Y, Horowitz MA, Liu J, Chew A, Lan H, Liu Q, Sha D and Yang C (2020) Individual-Level Fatality Prediction of COVID-19 Patients Using AI Methods. Front. Public Health 8:587937. doi: 10.3389/fpubh.2020.587937

[4] Murtas R, Morici N, Cogliati C, Puoti M, Omazzi B, Bergamaschi W, Voza A, Rovere Querini P, Stefanini G, Manfredi M, Zocchi M, Mangiagalli A, Brambilla C, Bosio M, Corradin M, Cortellaro F, Trivelli M, Savonitto S, Russo A

Algorithm for Individual Prediction of COVID-19–Related Hospitalization Based on Symptoms: Development and Implementation Study

JMIR Public Health Surveill 2021;7(11):e29504

URL: https://publichealth.jmir.org/2021/11/e29504

DOI: 10.2196/29504

[5]