ON THE USE OF MACHINE LEARNING FOR PREDICTING COVID-19 CASES: AN OVERVIEW OF EVOLVING DATASETS AND DISCRIMINATIVE FEATURES

Oyelakin A. M.[1], Salau-Ibrahim T. T.[1], Ogidan B.S.[2], Azeez R.D.[2], Ajiboye I. K.[3]

[1]Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria.
[2] ICT Centre, Al-Hikmah University, Ilorin, Nigeria
[3] Abdulraheem College of Advanced Studies, An Affiliate of Al-Hikmah University, Ilorin, Nigeria

Correspondence Email: amoyelakin@alhikmah.edu.ng

ABSTRACT: *The corona virus pandemic is one of the largest global health crises in recent times. Several studies are being proposed to demonstrate how Machine Learning techniques can be used for the prediction of the disease. While some of the studies focus on the prediction of the diseases in patients, some other ones focus on predicting its mode of spread across countries of the world. Corona virus was formally announced in China in December, 2019. Since then, the disease has been spreading in large number. More than 3.4 million people have been infected as at 3rd May, 2020. As parts of the efforts to help in predicting the presence of the disease in patients as well as its geographical spread, some Machine Learning-based techniques are being proposed. This study first explores the various promises of Machine Learning as a sub field of Artificial Intelligence in the prediction of corona virus disease. We provided information on how Machine Learning approaches can aid for predictive analytics. We equally emphasized the use of comprehensive and reliable datasets for reliable predictive modeling of the virus. However, since the disease is novel, it is of interest to study some of the datasets that are evolving. This is because having representative datasets for training and testing predictive COVID-19 models is very important. It is believed that future researchers in the area of corona virus classification and prediction can benefit immensely from the findings of this study.*

Keywords*: COVID-19, COVID-19 prediction, Machine Learning-based Predictive Models*

1. INTRODUCTION

The coronavirus pandemic is a worldwide health crisis and the number of people infected keeps increasing rapidly. The disease is tagged COVID-19 disease. It is a respiratory virus that was first reported officially in China in December 2019. The political, economic and social impacts of corona virus the world are magnificent. Since then, the disease has been spreading in large number. More than 3.4 million people have been infected as at 3rd May, 2020. Several studies are being published on the promises that Machine Learning techniques have for the prediction of corona virus. While some of the studies focus on the prediction of the diseases on patients, some other ones focus on predicting its spread across countries of the world. Of recent, datasets of different kinds are being collected, validated and released in to the public domain.

However, since the disease is novel, there is a need for huge and representative datasets for the Machine Learning prediction tasks. There have been several studies whereby Machine Learning is applied for disease diagnosis and prediction. In such studies, physicians record several medical parameters related to the disease and then evaluate some clinical status of the prediction.

World Health Organization (2020) explained that Corona virus spreads through airborne transmission, when tiny droplets remain in the air even after the person with the virus leaves the area. With the growing devastating effects of Corona virus on lives and socio-economic activities throughout the globe, the earlier diagnosis and prediction is of great importance. This is because the rate at which people are being infected worldwide created increased pressure on health workers and health facilities. Machine learning researchers can build models that can be used for earlier diagnosis and prediction of the disease in patients, which can facilitate their being recommended for clinical testing in the approved testing laboratories (Godfried, 2020). As pointed out by Naude (2020) data scientists are fast taking up the challenge of supporting the efforts of health workers in relation to the prediction and treatment of COVID-19, through the use of technology. A particular sub-field of Artificial Intelligence that has been identified to be of very importance is Machine Learning. In different areas including health sector, Machine Learning has been widely used for various classification problems in different domains (Hall, 1999). There is no doubt that the world is facing one of the greatest health challenges in recent times. Since the coming of this pandemic, World Health Organization has released several situational reports and technical guidance so that every country of the world can take caution against its spread and achieve recovery of the infected.

This study first explores the various promises of Machine Learning as a sub field of Artificial Intelligence in the prediction of corona virus disease. This is because having representative datasets for training and testing predictive models is very important. This study first explores the various promises of Machine Learning as a sub field of Artificial Intelligence in the prediction of corona virus disease. Then, mention was made of the characteristic nature of some of the preliminary datasets that are being released so far. We carry out preliminary investigation of these datasets and then come up with some challenges that researchers may have in using those datasets for reliable, verifiable and effective Machine Learning-based prediction models. The main reason for this approach is to provide earlier insights to research communities in this area.

2. MACHINE LEARNING FOR COVID-19 BASED CLASSIFICATION

Machine Learning is a very large interdisciplinary field that is deeply based on concepts from computer science, statistics, cognitive science, engineering, optimization theory and many other mathematical-related disciplines in mathematics (Ghahramani, 2004). Machine learning algorithms have been widely used for various classification problems in several fields, including health sector. Machine learning involves training an algorithm to perform tasks by learning from patterns in data rather than performing a task it is explicitly programmed to do (Hall, 1999; Chartrand, Cheng, Vorontsov, Drozdzal, Turcotte, et al., 2020). In a Machine Learning classification problem, each training point belongs to one of N various classes. The goal is to construct a function which, given a new data point, will correctly predict the class to which the new point belongs.

Alpaydin (2010) pointed out that data plays an indispensable role in machine learning as the learning algorithm is used to discover and learn knowledge or properties from the data. In order to build an effective machine learning model, it is essential to train the model on and test it against data that come from the same target distribution. Similarly, to have a very effective Machine Learning model, the quantity and quality of dataset used is necessary. Machine learning does involve a common task of constructing algorithms that can learn from and make predictions on data (Ron Kohavi & Foster Provost, 1998*)*. Such algorithms work by making data-driven predictions or decisions by building a mathematical model from input data (Bishop, 2006). Studies that use machine learning approaches for COVID-19 predictive models are growing as, there have been studies that focus on the epidemic trend in the world (Li et al, 2020), covid-19 prediction in patients based on clinical data used for training the model (Onoja, 2020; Jiang, Coffee, Bari, Wang, Jiang, et al., 2020) as well as those studies that are more centered on recovery rate in different countries of the world (Yan, Zhang, Xiao, Wang, M., … Yuan, 2020). Similarly, Several studies including Kuniya (2020), Li, Zhang, Jiang, Liu, Chen, Zhang and Wang (2020) have proposed Machine Learning based approaches for the prediction of corona virus in patients as well as its spread across the world. Santosh (2020) pointed out that Artificial Intelligence (AI) promises a new paradigm for healthcare, as several different AI tools that are built upon Machine Learning algorithms are employed for analyzing data and decision-making processes. This means that AI-driven tools help identify COVID-19 outbreaks as well as forecast their nature of spread across the globe.

3. PROMISING FEATURES FOR TRAINING MACHINE LEARNING BASED COVID 19 MODELS

In every machine learning problem, it is expected that the most important features are used to train and test for the best classifier performance. A feature is the input variable or attribute which is any representative information that is extracted from the raw data set. In respect of covid-19 prediction, Barstugan, Ozturk and Ozkaya (2020) argued that expert radiologists have identified that computerized tomography images are promising for detecting coronavirus as COVID-19 shows different behaviours from other viral pneumonia. Also, Farid, Selim, and Khater (2020) proposed a novel CT image analysis for building a predictive model for the new corona virus. Apart from this, some other studies have identified that the genetic and phenotypic structure of COVID-19 in pathogenesis is important (Mousavizadeh & Ghasemi, 2020). Based on the two popular approaches for detecting novel corona virus, the most important of these features that can be used for training Machine Learning-based models can be identified. A computerized tomography scan make use of computers and rotating X-ray machines to create cross-sectional images of the body.

These images have been found capable to show the soft tissues, blood vessels, and bones in various parts of the body. There are some new studies that are doing clinical characterization of COVID-19 by summarizing the effect of COVID-19 on various body organs (Naji, 2020). Some of these studies identified the need to use CT images, Protein genomes etc in patients that are infected in patients for the prediction of the disease. Preliminary investigations showed that some of the features used by some Machine Learning-based COVID-19 predictive models are promising enough.

In Machine Learning, feature selection is carried out on the chosen dataset with a view to achieving the following: increasing the predictive ability, increase interpretability and reduce computational time of the proposed model. Given a feature space, the focus is on how to have an optimal mapping which will serve as the one that does not result into increase in the minimum probability of error.

Overview of COVID-19 Datasets

As at May 05, 2020, 15:36 GMT, the number of coronavirus cases globally stood at 3,676,502 while 253,473 have died and about 1,211,848 have recovered from the infections (Worldometer, 2020;WHO, May 4, 2020). Several preliminary datasets on the corona virus pandemic are being collected and released to the public domain. The datasets come with a lot of promises. However, these datasets need to be investigated thoroughly before they can be used for building Machine Learning models that will be reliable, verifiable and effective. Some of the datasets identified so far are still being updated from time to time, as the disease is spreading across countries and communities. The focus of some of these datasets is the clinical and epidemiologic description or characteristics of the affected COVID-19 patients. Based on preliminary findings, it is believed that these set of data will be needed in every Machine learning based model. As the repository of coronavirus-related databases are growing, one cannot say for sure that these datasets can be used for a reliable and verifiable Machine Learning-based predictive models yet. It is important to validate every dataset released. IT has been identified that some data sources only report the infection rate, the number of those that have recovered as well as those that died across countries of the world.

Coronavirus-related databases/datasets are growing rapidly. Several public datasets that are related to coronavirus diseases have been released in recent times. For instance, we have them from Johns Hopkins Center for Systems Science and Engineering (JHU CSSE), Global Health Data from the World Bank, and OpenStreetMap data are current. We equally have Dataset on COVID from Next Web as well as from Harvard University. Human Data has released COVID-19 related Dataset that can be used for classification problems.

There are comprehensive dataset from Kaggle that focuses on the novel corona virus. Last but not the least, coronavirus-source-data have been released by Our World in Data. There is also Corona virus tweet dataset that was released by The NextWeb. All these datasets are currently available for download (The Next Web, 2020). Preliminary investigation in some of these datasets showed that some of these datasets are still being collected as the COVID-19 disease keeps spreading and affected countries of the world are battling with new infections and increased fatality.

Several researches in the past have identified that building effective Machine learning models is based on some factors. One of such factor is the choice of dataset. These datasets contain Early Epidemiological Cases of the diseases in some countries of the world. Some of the datasets identified in this study are the ones by the following: platforms Corona virus Dataset by Kaggle, Novel Corona virus (COVID-19) Cases Data released by humdata, Corona virus Data hub, Data.europa.eu, Corona virus Source Data ourworldindata.org, Google's comprehensive datasets for Corona virus researches and Corona Virus Tweets Dataset by the Next Web. Roser, Ritchie, Ortiz-Ospina and Hasell (2020) prepared and provided a complete COVID-19 dataset in two formats namely: .xslx and .csv. The dataset is updated daily as the virus is spreading and new cases being reported in the six continents.

4. METHODOLGY

The methodology employed in this study is to give a description of the promises that Machine learning approach has for predicting the COVID-19 disease. We equally investigate the datasets that are fast evolving as the corona virus is ravaging the world. Then, our literature search shifted to identify some works that promise the use of Machine learning prediction the virus among patients as well as the ones that investigate its geographical spread. The study specifically searched and identified good resources from the following online research databases: medRxiv, arxiv, mdpi, and Research gate as well as technical reports from World Health Organanisation and some other reputable health and technological company's repositories.

We provided an overview of the current set of publicly available datasets on the novel corona virus, popularly called COVID-19. Then, some of the Machine-learning based challenges that these datasets may have and that have to be overcome before being used for reliable research results are mentioned.

Challenges of Predicting corona virus with current datasets

Some of the challenges that may arise in some of these datasets can depend on how to choose the best data repository for the work which may depend on the nature of the research approach. Another one can be based on the need to do validation of some of the records contained in the dataset since the datasets are released continuously and the disease keeps spreading. Generally, in Machine learning, there have been challenges such as curse of small dataset or curse of dimensionality in large datasets. Small data can cause some contemporary machine learning algorithm to overfit. It is expected that the researcher identifies some of these challenges and other data pre-processing challenges and make the right decision early enough while proposing a Machine Learning based COVID-19 predictive model.

5. DISCUSSIONS

This study gave an overview of works that argued in favour using Machine Learning as a sub-field of Artificial Intelligence for the prediction of COVID-19. Specifically, the promises in respect of identifying the disease in patient as well as forecasting its spread in communities are discussed. The work equally mentioned some preliminary COVID-19 datasets being collected and publicly released. Some of these datasets are based on the clinical features observed in countries that are battling with the corona virus. We identified that some of the recent Machine Leaning studies surveyed are based on the statistical data collected based on the infection rate, the recovery rates as well as the fatality date (death rate). We equally found out that some of the datasets identified so far are still being updated from time to time, as the disease is spreading across countries and communities. Lastly, it was discovered that the focus of some of these datasets is on the clinical and epidemiologic description or characteristics of the affected COVID-19 patients.

CONCLUSION AND FUTURE WORK

This study provided an overview of the strengths that Machine learning algorithms for the prediction of COVID-19. Some of the research directions in the use of Machine Learning algorithms for predicting the disease were also discussed. The emerging datasets that are being publicly made available are identified and the need for reliable, verifiable and comprehensive datasets was pointed out. The research also found out that some of these datasets are being updated from time to time as the COVID-19 cases keep growing in most of the affected nations.In future, the focus will be on proposing Machine Learning-based framework that will give promising prediction and characterization of corona virus in patients as well as its spread in communities of the affected countries. Emphasis will be on proposing hybrid technique that can classify the virus in both real time and offline basis. Comprehensive datasets that will focus on the main features that can be used for predicting infections as well as geographical spread will be required.

REFERENCES

[1] **Alpaydin E.** (2010). Introduction to Machine Learning, MIT Press https://github.com/mindis/002_MachineLearning_eBook/blob/master/Alpaydin%20-%20Introduction%20to%20Machine%20Learning%20(MIT%2C%202004).pdf

[2] **Barstugan M., Ozturk Saban & Ozkaya Umut** (2020). Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods

[3] **Bishop Christopher M.** (2006). Pattern Recognition and Machine Learning. New York: Springer. p.7. *ISBN 0-387-31073-8.*

[4] **Chartrand G. et al.** (2020). Deep learning: a primer for radiologists. Radiographics. 2017; 37 (7):21132131 https://pubs.rsna.org/doi/pdf/10.1148/rg.2017170077

[5] **Farid A. A., Selim G. I., & Khater H. A. A**. (2020). *A Novel Approach of CT Images Feature Analysis and Prediction to Screen for Corona Virus Disease (COVID-19),* International Journal of Scientific and Engineering Research 11(3):1141, 1–9. https://doi.org/10.20944/preprints202003.0284.v1

[6] **Ghahramani Z.** (2004). Unsupervised learning, in Advanced lectures on machine learning, ed: Springer, 2004; pp. 72112. https://doi.org/10.1007/978-3-540-28650-9_5

[7] **Godfried Isaac** (2020). Machine Learning methods to aid in Corona virus Response, Battling Coronavirus with Data Science and Artificial Intelligence, https://towardsdatascience.com/machine-learning-methods-to-aid-in-coronavirus-response-70df8bfc7

[8] **Hall Mark A.** (1999). Correlation-based Feature Selection for Machine Learning, a PhD Thesis at University of Waikato, retrieved from https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

[9] **Jiang, X. et al.** (2020). Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity. *CMC-Computers, Materials & Continua, 63(1),* 537–551

[10] **John Hopkins** University & Medicine, Corona Virus Resource Centre retrieved from https://coronavirus.jhu.edu/data/new-cases on 17th April, 2020

[11] **Kuniya, T.** (2020). Prediction of the Epidemic Peak of Coronavirus Disease in Japan, Journal of Clinical Medicine, 2020. 1–7, https://doi.org/10.3390/jcm9030789, retrieved from https://www.mdpi.com/2077-0383/9/3/789

[12] **Li Mengyuan, Zhang Zhilan, Jiang Shanmei, Liu Qian, Chen Canping, Zhang Yue, Wang Xiaosheng** (2020). Predicting the epidemic trend of COVID-19 in China and across the world using the machine learning approach, https://doi.org/10.1101/2020.03.18.20038117

[13] **Mousavizadeh Leila and Ghasemi Sorayya** (2020). Genotype and phenotype of COVID-19: Their roles in pathogenesis,Journal of Microbiology, Immunology and Infection https://doi.org/10.1016/j.jmii.2020.03.022, retrieved on online 6th May, 2020

[14] **Naude Wim** (2020).Artificial Intelligence against COVID-19: An Early Review, *AI has not yet made an impact, but data scientists have taken up the challenge*, https://towardsdatascience.com/artificial-intelligence-against-covid-19-an-early-review-92a8360edaba

[15] **Naji, H.** (2020). *Clinical Characterization of COVID-19*. EJMED, European Journal of Medical and Health Sciences 2(2), March 2020 https://doi.org/10.24018/ejmed.2020.2.2.194

[16] **Onoja A.A.** (2020). A Propose Machine Learning approach for Monitoring Individual's Health Status on Corona virus (COVID19) cases, March 202 Project: Applying Support Vector Machine and AdaBoost Method to Ordinal Classifition

@019https://www.researchgate.net/publication/338821819_Host_and_infectivity_prediction_of_Wuhan_2019_novel_coronavirus_using_deep_learning_algorithm

[17] **Ron Kohavi and Foster Provost** (1998). *"Glossary of terms". Machine Learning*. 30: 271–274. *doi:10.1023/A:1007411609915*.

[18] **Roser Max, Ritchie Hannah, Ortiz-Ospina Esteban and Hasell Joe** (2020).Coronavirus (COVID-19) DeathsStatistics and Research, retrieved from

https://ourworldindata.org/covid-deaths

[19] **Santosh K.C.** (2020). AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data,

[20] **The Next Web** (2020). Google offers comprehensive coronavirus datasets to researchers for free, retrieved from https://thenextweb.com/neural/2020/03/31/google-offers-comprehensive-coronavirus-datasets-to-researchers-for-free

[21] **WHO** (2020). Laboratory testing strategy recommendations for COVID-19, Interim Guidance published on 21st March, 2020 retrieved from

https://apps.who.int/iris/bitstream/handle/10665/331509/WHO-COVID-19-lab_testing-2020.1-eng.pdf

[22] **WHO** (2020). WHO Situation Report for 4th May 2020 retrieved from

https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200504-covid-19-sitrep-105.pdf?sfvrsn=4cdda8af_2

[23] **Worldometer** (2020). COVID-19 Coronavirus Pandemic retrieved from

https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1?

[24] **Yan, L., Zhang, H., Xiao, Y., Wang, M., Yuan, Y.** (2020). *Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in Wuhan*