The application of Machine learning for COVID-19

Yuhan Yang[1,a], Zhulin Chen[2,b], Lei Sun[3,c], Yang Lu[4,d], Yongjie Li[5,e]

[1]Rutgers University, New Jersey, USA

[2]University of Edinburg, UK

[3]East China University of Science and Technology, Shanghai, China

[4]Dalhousie University, Halifax, Nova Scotia

[5]University of Electronic Science and Technology of China, Chengdu, China

[a]Melody950218@gmail.com, [b]S2001758@ed.ac.uk, [c] 1207639682@qq.com, [d]Yang.Lu@dal.ca, [e]2017200604026@std.edu.cn,

Keywords: Machine Learning, Covid-19, Gene Analysis, CT.

Abstract: Machine learning method develops very fast and has been applied in various industries. In novel coronavirus pneumonia diagnosis, recognition and treatment, machine learning has shown great advantages, which is faster, more efficient and less labor cost than traditional methods, and has made a great contribution to epidemic prevention. This paper focuses on novel coronavirus pneumonia gene sequencing and CT diagnosis.

1. Introduction

Machine learning has found a wide variety of applications on the prediction of the Coronavirus disease (COVID-19) since its outbreak. We provide a general overview on the application of machine learning in the prediction of COVID-19 affairs. At first, due to the lack of some fundamental data and accurate statistic results, many models proved to be inefficient. However, Ardabili et al in [1] showed after a comparative analysis of diverse models that only two models yielded reasonable results, which include Multi-Layered Perception (MLP) and Adaptive Network-based Fuzzy Interference Systems (ANFIS). By conducting experiments in two different scenarios: the infection progress on previous days and daily sampling on consecutive days, the authors found MLP most accurate in rendering outbreak prediction and ANFIS could also generate rational results. Despite so much unknown on the efficacy of machine learning models, this work provided a promising insight at the beginning of the outbreak.

Santosh [2] made use of AI-tools based on machine learning to predict the COVID-19 outbreaks and their spread speeds in different areas. To realize this, this paper introduced the active learning-based cross-population train or test models which employ the multitudinal as well as multi-model data. In its experiments, the active learning technique was based on multitudinal as well as multi-model data, and the forecasting of spread speeds took the advantage of cross-population train and test models of AI tools. Poirior et al in [3] managed to predict the outbreak of COVID-19 in China with the aid of novel-digital-data-based and mechanistic-model-based machine learning. They proposed a new interpretable methodology integrating disease estimates from mechanistic models with digital traces based on machine learning. The results were shown to be accurate enough to forecast the outbreak two days in advance in 27 out of 32 provinces in China. Pinter et al [4] rendered a hybrid machine learning method for COVID-19 pandemic prediction and also evaluate its efficiency in predicting the outbreak in Hungary. Similar to [5], this hybrid method was built upon the Adaptive Network-based Fuzzy Interference Systems (ANFIS) and Multi-Layered Perception-Imperialist Competitive Algorithm (MLP-ICA), which aimed to forecast the time series of infected individuals as well as mortality rate. And a validation for nine days is demonstrated by experimental results compared with the real outbreak trend. Theoretically, without significant outbreaks, this model is considered to be reliable. A simple average aggregated machine learning method was proposed to predict the number, size, and length of the COVID-19 pandemic across India.

Also, Pourhomayoun et al [6] utilized machine learning and AI models to help predict the mortality risks of COVID-19. The document data of 117000 patients from different countries were used for evaluation. This paper provided a AI model demonstrating a 93% overall accuracy, which was mainly intended for helping the overwhlemed hospitals or other health facilities to determine which patient might require attention in priority according to the predicted mortality rate. Moreover, several machine learning methods were also employed, including Support Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Tree, Random Forest, Logistic Regression and K-Nearest Neighbor (KNN). Through them, the most apparent symptoms and features were detected and a confusion matrix method was further introduced to analyze the sensitivity of the model. The classical Kermack-Mckendrick SIR model was tested in [7] by Ndiaye et al to forecast the COVID-19 pandemic. Using the public data, this paper estimated the main key pandemic parameters and predicted the inflection point and the probable time of pandemic ending in the real world, specifically for Senegal. Promising results were then rendered.

Similarly, the growth and trend of COVID-19 pandemic were also implemented with a machine learning model as well as cloud computing by Tuli et al in [11]. They showed that with the aid of iterative weighting while fitting Generalized Inverse Weibull distribution, a more promising framework can be generated for the prediction. In [12], Yan et al developed a machine-learning-based prognostic model for predicting the survival of severe patients and it proved to reach over 90% accuracy according to the experimental evaluation over the clinical dataset of Tongji hospital, Wuhan. This method only needed three clinical features so that it realized a low-cost and efficient criticality classification and survival prediction even before targeted intervening. In [13], interested reader can also find some further reviews of the summary, challenges and suggestion about the application of machine learning or other AI tools in COVID-19 related problems.

2. Machine learning for genetic analysis of Covid-19

2.1 Taxonomy study of 2019-nCov

Taxonomy study of 2019-nCov can assist to locate the origin of the virus and to its relationship with relative species, and thus help us quickly comprehend the occurrence and transmission of the disease, leading to a quick response to infectious disease outbreaks. After the affection of 2019-nCov was firstly reported from Wuhan in Hubei province in China, the results of full-genome sequence and phylogenetic analysis reveals that 2019-nCov belongs to the betacoronaviruses, which are one of the genera of Coronaviridae [14][15]. Coronaviridaes are enveloped, positive-sense, single-stranded RNA viruses and they can cause respiratory infections in human . For 2019-nCov, researches indicate that it is likely originating from bat coronavirus RaTG13 because they share 96.2% overall genome sequence identity. Traditional methods used in classification of 2019-nCov are based on alignment of sequences.

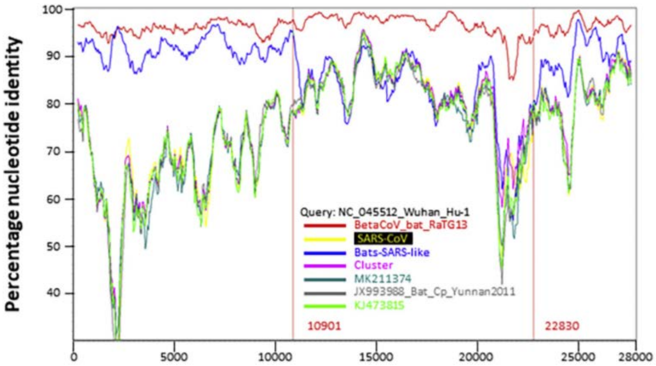

Figure 1 Genome nucleotide position

2.2 Drug targets identification and machine learning methods

Drug targets identification is the process of identifying specific parts of target protein that can be modulated by specific small molecules [17]. For 2019-nCov, finding its drug binding pockets can promote the efficiency of drug design and drug repurposing. There are many researches focusing on searching similar binding pockets of 2019-nCov and other RNA viruses, such as influenza, SARS-Cov for drug repurposing. Traditional methods of drug targets identification are based on structure analysis. For 2019-nCov, learning its genome, it possesses 4 structural proteins: surface glycoprotein (S), envelope protein (E), integral membrane glycoprotein (M) and nucleocapsid protein (N), 16 non-structural proteins (nsps) and eight accessory proteins (3a, 3b, p6, 7a, 7b, 8b, 9b, and orf14) (Figure 3). These proteins collaborate fulfilling the process of cell entry, translation and transmission for the virus and by studying these proteins' structure more therapeutics options will be available. For instance, recent studies indicate that 2019-nCov enter cells through S protein binding with angiotensin-converting enzyme 2 (ACE2), which is the host cell receptor that SARS-Cov used for entrance. Several drugs that was used for SARS-Cov treatment have been proposed for clinical test in the treatment of 2019-nCov focusing on blocking the interactions between S protein and ACE2 to hinder 2019-nCov access into host cells.

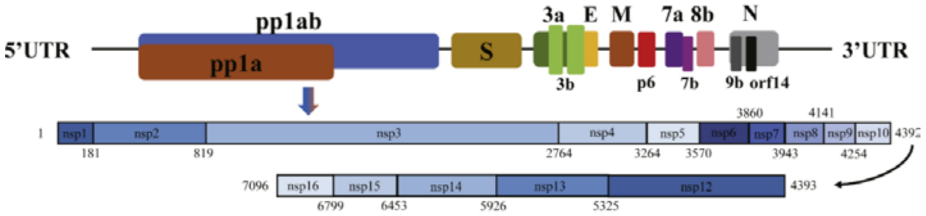

Figure 2 genome composition of 2019-nCov

In addition to that, some methods of drug targets identification focusing on the lifecycle of 2019-nCov in host cells. As a member of coronavirus, 2019-nCov's genome enter cells through entry receptors and translation of structural and nonstructural protein get started. Nonstructural proteins facilitate the replication of 2019-nCov's genome and the assembly of newly synthesized genome and structural proteins through forming a RNA replicase-transcriptase complex. Following by these steps, the virion was released from host cells through exocytosis (Figure 3).

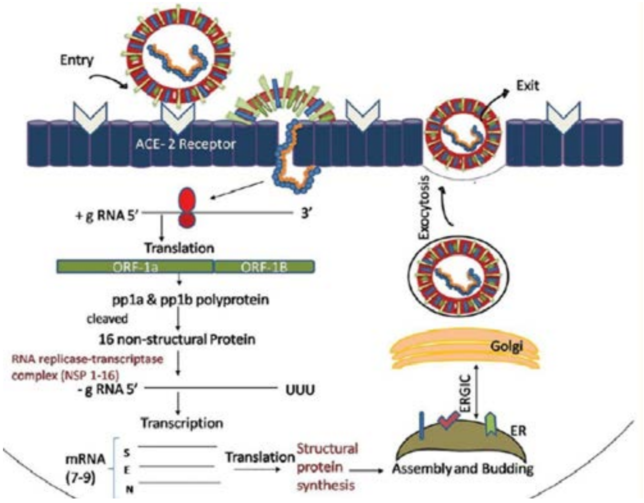

Figure 3 Details of coronavirus lifecycle in host cells

3. CT analysis with machine learning for COVID-19

The novel coronavirus (COVID-19) outbreak that was discovered at the end of 2019 requires special attention because it will be epidemic for a long time in the future and pose a threat to the health of the global population. In addition to clinical procedures and traditional treatments, artificial intelligence (AI) has brought new inspiration to healthcare. Artificial intelligence-driven tools can help identify the COVID-19 pandemic, and advances in machine learning can help in many aspects of the COVID-19 pandemic. From a clinical perspective, machine learning can be used to detect COVID-19 and predict patient outcomes. The traditional manual method of screening suspected cases is widely used, but this process takes a long time, and the number of false-negative patients in the test results is significant. Therefore, it is necessary to use machine learning as an alternative diagnostic method to support CT scan diagnosis, to fight diseases accurately and quickly.

Although the new coronavirus exhibits unique radiological features and image patterns in CT scans, identifying these features is still tricky and time-consuming, even for experienced radiologists. Therefore, due to the complexity of imaging data, researchers have developed a variety of deep learning-based models, which are based on CT chest scans and used for automatic screening and diagnosis of COVID-19 detection. The general principle of action is to input the preprocessed CT image, and then the model extracts the region of interest (ROI) from it and passes ROI to the neural network, which uses an ensemble method to classify each ROI and predict the infection rate. After data augmentation process, classification models are based on several networks. The first Model contain 4 convolutional layers followed by max-pooling layer, at the end 1 flatten, 1 drop-out and 2 fully connected layers were added. Input is images collected and generated through GANs and was fed to convolutional layer. Transfer Learning Using Inception-V3 and ResNet model without image augmentation are trained based on the dataset as well. The validation accuracy of first Model remains almost constant at 83% and 80%. The training curve indicates that more epochs might increase accuracy of models. Transfer learning model can increase performance/accuracy by almost 2-3%. The validation accuracy of third model stands form at 73% which could vaguely suggest that the augmentation of dataset through GAN is effective concerning the accuracy.

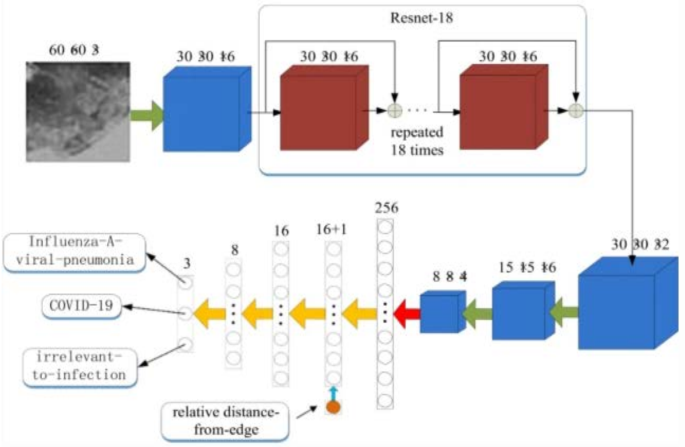

Figure 4 GAN method

New measurements are brought up to better extract the feature. An article conduct a study that utilized CNN models and introduce a new measurement called relative distance from-edge, which could increase the accuracy of traditional CNN Res Net.[4] The dataset in this study is a total of 618 transverse-section CT samples including 219 from 110 patients with COVID-19, and 224 CT samples were having Influenza-A viral pneumonia including H1N1, H3N2, and 175 are from healthy people. The author compared traditional CNN, ResNet-18 with the net adding relative distance from-edge, which serves as the extra weight to learn the relative location of the patch in the image. The structures of two nets are shown in figure 2. Using ResNet to extract features from CT images together with a location-attention mechanism model, compared to without the location-attention model, could more accurately distinguish COVID-19 cases from others, with an overall accuracy rate of 86.7%. Novel deep neural networks were proposed, like a network named as CovXNet based on depthwise dilated convolutions [5]. Firstly. This initial training phase is transferred with additional fine-tuning layers that are next trained with a smaller number of chest X-raysof COVID-19. Most importantly, different forms of CovXNets are designed and trained with X-ray images of various resolutions and for further optimization of their predictions, a stacking algorithm is employed. Finally, a gradient-based discriminative localization is integrated to distinguish the abnormal regions of X-ray images referring to different types of pneumonia. The experimentations provide very satisfactory detection performance with accuracy of 90.2% for multiclass COVID /normal /Viral/Bacterial pneumonias.

4. Conclusion

In this paper, we review the application of machine learning methods in covid-19, including generating antagonistic neural network, support vector machine, gene sequence alignment and analysis, etc. machine learning method has a very good effect, and has a role in promoting epidemic prevention and other work. In the future, machine learning methods can be more applied to vertical application industries, improve the efficiency of traditional industries, and promote the development of society.

References

[1] N. Zhu, et al.A novel coronavirus from patients with pneumonia in China, 2019N Engl J Med (2020).

[2] de Groot RJ, Baker SC, Baric R, Enjuanes L, Gorbalenya AE, Holmes KV, et al. Family Coronaviridae. In: King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ, editors. Virus taxonomy. Ninth report of the international committee on taxonomy of viruses, Elsevier Academic Press; 2012. pp.

[3] Kahn, Jeffrey S. MD, PhD*; McIntosh, Kenneth MD† History and Recent Advances in Coronavirus Discovery, The Pediatric Infectious Disease Journal: November 2005 - Volume 24 - Issue 11 - p S223-S227.

[4] P. Zhou, et al.Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat originbioRxiv (2020)p. 2020.01.22.914952.

[5] G. Stecher, K. Tamura, S. KumarMolecular evolutionary genetics analysis (MEGA) for macOSMol. Biol. Evol. (2020), 10.1093/molbev/msz312pii: msz312.

[6] K.S. Lole, et al.Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombinationJ. Virol., 73 (1) (1999), pp. 152-160.

[7] Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infection, Genetics and Evolution. 2020; 79: 104212. pmid:3200475.

[8] Schenone M, Dančík V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. Nat Chem Biol. 2013;9(4):232-240. doi:10.1038/nchembio.1199.

[9] Zhou, Y., Hou, Y., Shen, J. et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov 6, 14 (2020). https://doi.org/10.1038/s41421-020-0153-3.

[10] H. LuDrug treatment options for the 2019-new coronavirus (2019-nCoV) Biosci Trends (2020 Jan 28), 10.5582/bst.2020.01020.

[11] Wu A, Peng Y, Huang B, et al. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 2020;27:325-328.

[12] Prajapat M, Sarma P, Shekhar N, et al. Drug targets for corona virus: A systematic review. Indian J Pharmacol. 2020;52(1):56-65. doi:10.4103/ijp.IJP_115_20.

[13] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, J. S. McLellan, Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science eabb2507 (2020). doi:10.1126/science.abb2507pmid:32075877.

[14] Stohr, S. et al. Host cell mTORC1 is required for HCV RNA replication. Gut 65, 2017–2028 (2016).

[15] Plotkin SA. Vaccines: past, present and future. Nat Med 2005;11:Suppl:S5-S11.