Journal of Medical Systems (2020) 44:93
https://doi.org/10.1007/s10916-020-01562-1

EDUCATION & TRAINING

AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data

K. C. Santosh [1]

Received: 11 March 2020 / Accepted: 17 March 2020

Abstract
The novel coronavirus (COVID-19) outbreak, which was identified in late 2019, requires special attention because of its future epidemics and possible global threats. Beside clinical procedures and treatments, since Artificial Intelligence (AI) promises a new paradigm for healthcare, several different AI tools that are built upon Machine Learning (ML) 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. However, unlike other healthcare issues, for COVID-19, to detect COVID-19, AI-driven tools are expected to have active learning-based cross-population train/test models that employs multitudinal and multimodal data, which is the primary purpose of the paper.

Keywords COVID-19 · Artificial intelligence · Machine learning · Active learning · Cross-population train/test models · Multitudinal and multimodal data

Introduction

The novel coronavirus (COVID-19) is a global threat since it was identified in late 2019 [1]. About COVID-19, the Centers for Disease Control and Prevention report [2] has clearly mentioned the following (U.S. Feb. 24, 2020):

> Person-to-person spread of COVID-19 appears to occur mainly by respiratory transmission. How easily the virus is transmitted between persons is currently unclear. Signs and symptoms of COVID-19 include fever, cough, and shortness of breath [3]. Based on the incubation period of illness for Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) coronaviruses, as well as observational data from reports of travel-related COVID-19, CDC estimates that symptoms of COVID-19 occur within 2–14 days after exposure.

According to the World Health Organization (WHO) report [4], as of today (March 09, 2020), China confirmed 80,904 cases and 3123 of them were died. Outside China, 28,673 were confirmed and 686 of them were died from 104 countries, where Italy is found to be the most influenced country after china: 7375 confirmed cases (366 deaths). Similarly, there are 7382 confirmed cases (248 deaths) in Republic of Korea, and U.S. is no exception (see Fig. 1). Based on the confirmed cases, fatality rate, as of today, is still less than other respiratory diseases: study of 72,000 COVID-19 patients finds 2.3% death rate [5]. Figure 2 provides confirmed cases of COVID-19 across the world. Considering its global coverage, the WHO has already been declared public health emergency and its possible global threats, including consequences [6]. The devastating case in Wuhan China and future epidemics require special attention [3, 7, 8].

At this point, it is important to note that coronavirus was not a surprise case, since several years ago, in 2003, a novel coronavirus was identified in patients with SARS, and it is thought to be caused by an unknown infectious agent [9–12]. Since then, apart from clinical procedures and treatments, Artificial Intelligence (AI) promises a new paradigm

This article is part of the Topical Collection on *Education & Training*

✉ K. C. Santosh
santosh.kc@ieee.org

[1] Department of Computer Science, University of South Dakota, 414 E Clark St, Vermillion, SD 57069, USA

Published online: 18 March 2020

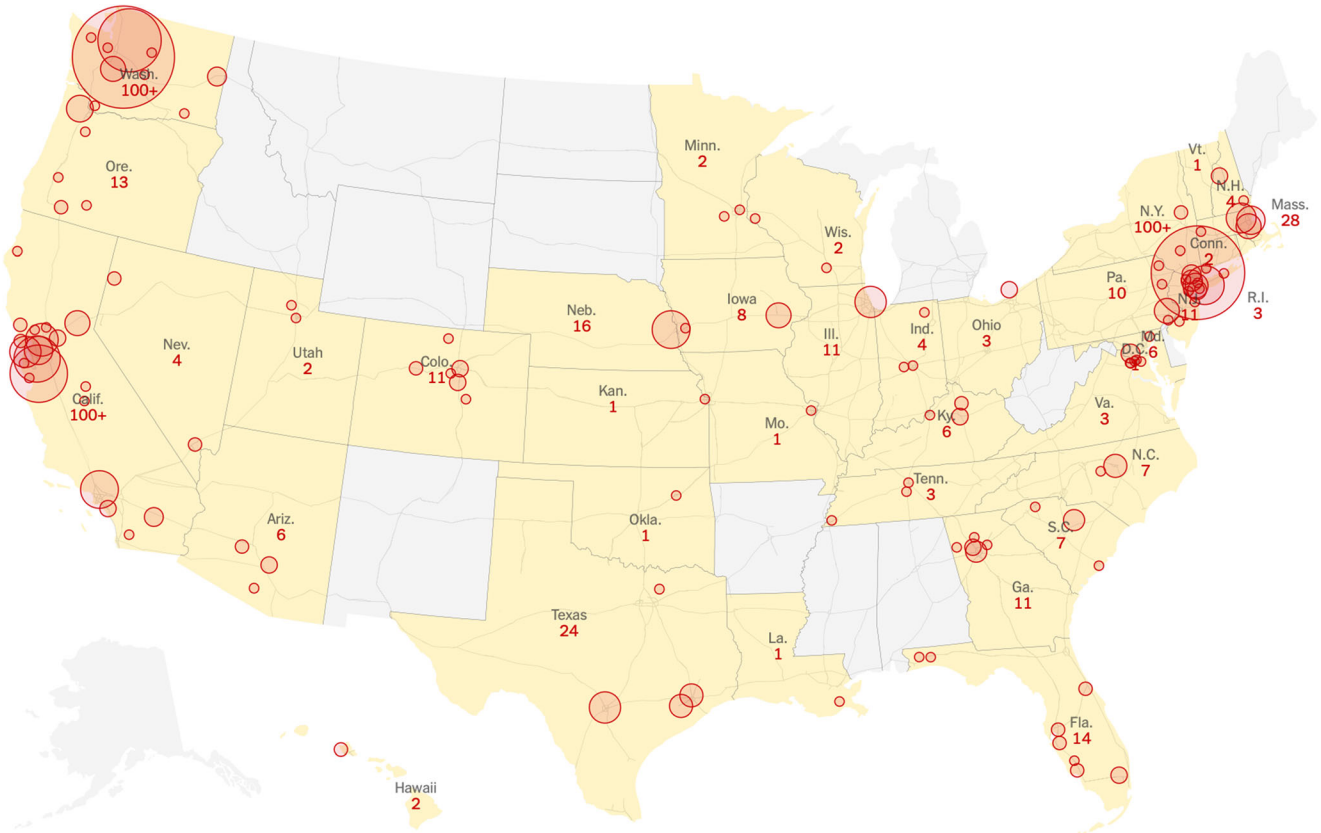

Fig. 1 Known locations of coronavirus cases by county in the US. Circles are sized by the number of people there who have tested positive, which may differ from where they contracted the illness. More than 100 cases have been identified in New York. (source: https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html, March 09, 2020)

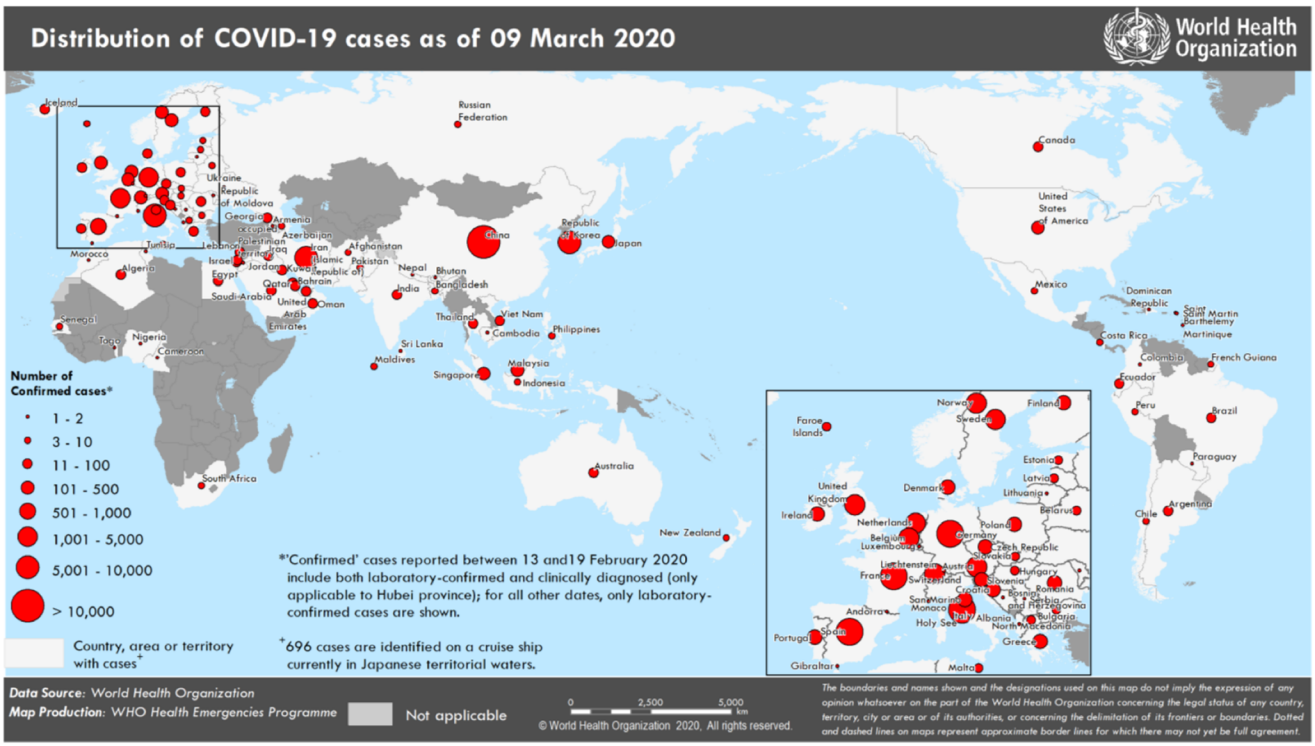

Fig. 2 Countries, territories or areas with reported confirmed cases of COVID- 19 (source: https://www.who.int/docs/default-source/coronaviruse/situation-reports/ 20,200,309-sitrep-49-covid-19.pdf?sfvrsn = 70dabe61_4, March 09, 2020)

for healthcare. Several different AI tools that are built upon Machine Learning (ML) algorithms are employed to analyzing data and decision-making processes [13].

AI-driven tools can be used to identify novel coronavirus outbreaks as well as forecast their nature of spread across the globe. However, the fundamental theory behind AI-driven tools is that they require sufficient training data (of all possible cases). Often, traditional machine learning requires a clean set of annotated data so that classifiers can possibly be well trained, which falls under scope of supervised learning. Over the past 5 decades or more, tremendous progress has been made in resolving many issues of several different projects. However, we failed to reach the point: "to model an accurate classifier, how big the size of training samples should be?" Do we still wait for collecting fairly large amount of data? Deep Learning (DL), as an example, requires a large amount of data to be trained [14, 15]. The primary idea behind the use of DL is not only to avoid feature engineering but also to extract tiny features in radiology data (pixel-level nodule, for example) [16].

Collecting large amount of data is not trivial, and one has to wait for a long. Most of the reported AI-driven tools are limited to proof-of-concept models for coronavirus case. AI experts state the fact that limited data may skew results away from the severity of coronavirus outbreak. The Wall Street Journal [17] reported that coronavirus reveals limits of AI health tools: some diagnostic-app makers are holding off updating their tools, highlighting the shortage of data on the new coronavirus and the limitations of health services billed as AI when faced with novel, fast-spreading illnesses (Parmy Olson, February 29, 2020). In a nutshell, social medias, newspapers, and health reports, we note that conventional AI-driven tools for real-world cases (with less data) may not provide optimal performance.

To detect COVID-19, AI-driven tools are expected to have *active learning*-based *cross-population train/test models* that employs *multitudinal and multimodal data*.

In the following, within the framework of COVID-19, the concept of active learning will be discussed (Section 2). Cross-population train/test models and its usefulness for COVID-19 are discussed in Section 3. The necessity and the use of multitudinal and multimodal data are explained in Section 4. The paper concludes in Section 5.

Active learning (AL)

As compared to passive learning (traditional machine learning classifiers), active learning is used to a learning problem, where the learner has some role in determining on what data it will be trained [13]. When it is an emergency (COVID-19) [6], it requires a special attention so that data analysis and decision-making can be made consistently without waiting several days, months, and years for data collection. Again, exploiting real-time data (on-the-fly) is the must since one cannot wait for years to train machine and learn from them nor manual annotation/analysis is possible. This means that instead of having a conventional set of train, validation, and test set, we need AI-driven tools that can learn over time without having complete knowledge about the data, which we call Active Learning (AL). In other words, AL mechanism helps self-learn i.e., Incremental Learning (IL) over time in the presence of experts (if required) [18]. The ILs aim is to iteratively help learn model to adapt to new data without forgetting its existing limited knowledge. Figure 3 provides a schematic diagram of an AL mechanism, where different data types are used. While learning, the changes in data over time can be assessed with the help of Anomaly Detection (AD) techniques. In AL-based tool, AD helps find/identify rare items, events or observations that bring suspicions by differing significantly from the majority of the data or with respect to a set of normal data for that particular event.

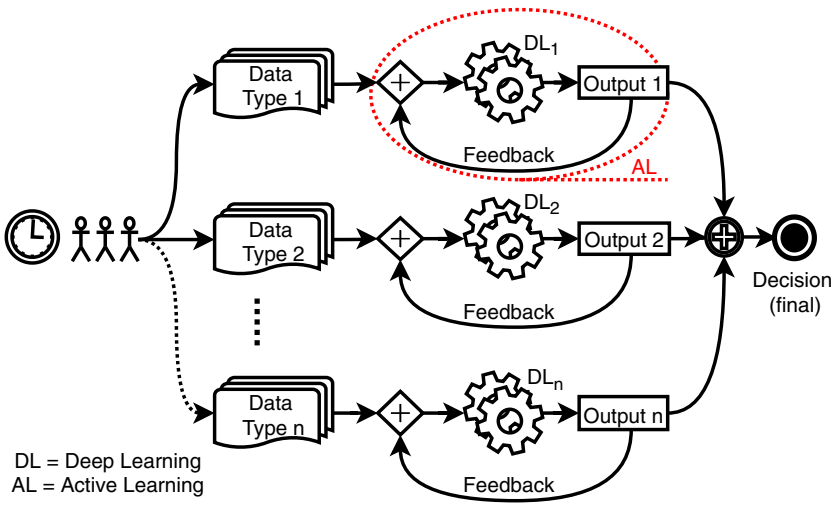

Fig. 3 For time-series data, a schema of Active Learning (AL) model is provided. For better understanding, AL (in dotted red circle) is used with Deep Learning (DL) for all possible data types. In AL, expert's feedback is used in parallel with the decisions from each data type. Since DL are data dependent, separate DLs are used for different data type. The final decision is made based on multitudinal and multimodal data

Cross-population train/test AI-driven models

Beside the use of AL in machine learning, cross-population train/test models are the must in such scenarios, since we do not have enough data from the particular regions. This means that there is a need to automatically detect nVirus in Italy from the model trained in Wuhan, China. In other words, for such a respiratory disease, it is essential to have cross-population train/test-based AI-driven models so that automated detection can be possible. In parallel, the collected data can be used for training models over time, which are based on the decisions. Conventionally, in the literature, such a concept does not exist.

Multitudinal and multimodal data

More often, AI-driven tools are limited to one data type. Decisions that are solely based on one data type (regardless of the data size) may be skewed away from the severity of coronavirus influence. In such a case, use of mutitudinal and multimodal data can help support decision-making process with higher confidence. Since coronaviruses are enveloped viruses with a positive-sense single-stranded RNA genome and a nucleocapsid of helical symmetry, the most popularly used data for AI-driven tools mostly employ RNA sequences. Besides, Electronic Health Record (EHRs), Computerized Tomography (CT) scans [19–21], Chest X-rays (CRRs, see Fig. 4), and other data are considered and tested. Alibaba launched a new AI-based system to detect coronavirus infection via CT scans with an accuracy of up to 96% [19]. As mentioned before, AD in image consists in finding portion of the images (a set of pixels) with anomalous and unusual patterns. With small changes in image pixels, spatio-temporal signatures can be deviated from standard threshold(s). Since AD is not just limited to image data; it can be applied for all types of data that are ranging from a vector (1D array/signal/pattern), 2D matrix (image, for instance) to multi-dimensional data. In the recently reported work [20], Chest CT (see Fig. 5, as an example) – which is used to diagnose COVID-19 – can be complemented to the reverse-transcription polymerase chain reaction (RT-PCR) tests. Therefore, instead of using different machine learning models for one data type (and even small size data [22]) and looking for ensemble techniques to combine results, for AI researchers, it is wise to use mutitudinal and multimodal data to check whether different data types can help yield consistent decisions about the coronavirus outbreak over time.

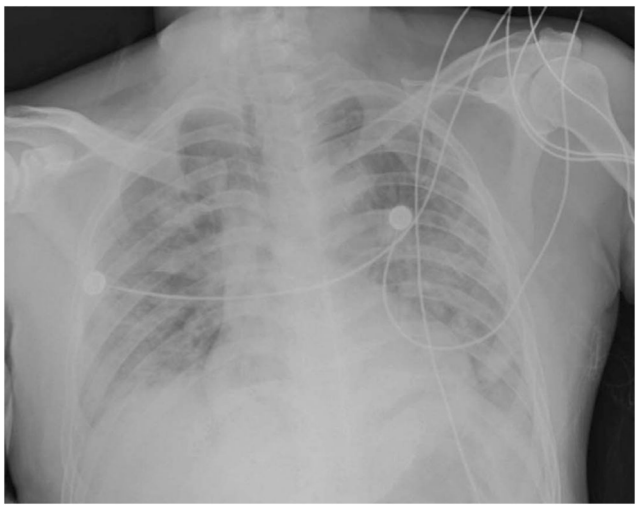

Fig. 4 Chest X-ray: Bilateral focal consolidation, lobar consolidation, and patchy consolidation are clearly observed (check lower lung [1])

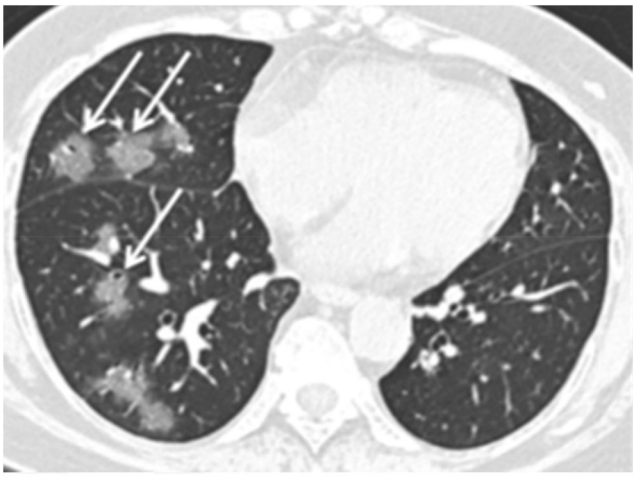

Fig. 5 Chest CT: An axial CT image shows ground-glass opacities with a rounded morphology (arrows) in the right middle and lower lobes [21]

Conclusions

Considering the possible future epidemics of the COVID-19, in this paper, the importance of the AI-driven tools and their appropriate train and test models have been introduced and discussed. The primary purpose of the paper is that AI scientists do not always wait for the complete datasets to train, validate, and test the models. Instead, AI-driven tools are required to be implemented from the beginning of data collection, in parallel with the experts in the field, where active learning needs to be employed. To achieve higher confidence during decision-making process, rather than relying on one data type, several data types are expected to be employed. For this, under the scope of active learning, the use of multitudinal and multimodal data have been discussed. Besides, considering the spread rate of COVID-19 (across the globe), AI-driven tools are expected to work as cross-population train/test models.

Compliance with ethical standards

Conflict of interest Author declared no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., Yuan, M. L., Zhang, Y. L., Dai, F. H., Liu, Y., Wang, Q. M., Zheng, J. J., Xu, L., Holmes, E. C., and Zhang, Y. Z., A new coronavirus associated with human respiratory disease in china. *Nature* 44(59):265–269, 2020. https://doi.org/10.1038/s41586-020-2008-3.
2. Centers for Disease Control and Prevention, Morbidity and Mortality Weekly Report (MMWR), Public Health Response to the Coronavirus Disease 2019 Outbreak – United States, February 24, 2020.
3. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., and Zhang, L., Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: A descriptive study. *The Lancet* 395(10223):507–513, 2020. http://www.sciencedirect.com/science/article/.
4. WHO Report, Coronavirus disease 2019 (COVID-19) Situation Report – 49 (2020 (accessed March 09, 2020)), https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200229-sitrep-40-covid-19.pdf?sfvrsn=849d0665_2.
5. Wu, Z., and McGoogan, J.M., Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72,314 cases from the Chinese center for disease control and prevention. *JAMA*, 2020. https://doi.org/10.1001/jama.2020.2648.
6. Medscape Medical News, The WHO declares public health emergency for novel coronavirus (January 30, 2020). https://www.medscape.com/viewarticle/924596.
7. Long, J.B., and Ehrenfeld, J.M., The role of augmented intelligence (ai) in detecting and preventing the spread of novel coronavirus. *Journal of Medical Systems* 44(59), 2020. https://doi.org/10.1007/s10916-020-1536-6.
8. Paules, C. I., Marston, H. D., and Fauci, A. S., Coronavirus infections – more than just the common cold. *JAMA* 323(8):707–708, 2020. https://doi.org/10.1001/jama.2020.0757.
9. Drosten, C., Gunther, S. et al., Identification of a novel corona virus in patients with severe acute respiratory syndrome. *N Engl J Med* 348:1967–1976, 2003. https://doi.org/10.1056/NEJMoa030747.
10. Peiris, J., Lai, S., Poon, L., Guan, Y., Yam, L., Lim, W., Nicholls, J., Yee, W., Yan, W., Cheung, M., Cheng, V., Chan, K., Tsang, D., Yung, R., Ng, T., and Yuen, K., Coronavirus as a possible cause of severe acute respiratory syndrome. *The Lancet* 361(9366), 2003. https://doi.org/10.1016/S0140-6736(03)13077-2.
11. Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., Zhu, H., Zhao, W., Han, Y., and Qin, C., From sars to mers, thrusting coronaviruses into the spotlight. *Viruses* 11(1), 2019. https://www.mdpi.com/1999-4915/11/1/59
12. de Wit, E., van Doremalen, N., Falzarano, D., and Munster, V. J., Sarsandmers: Recentinsights into emerging coronaviruses. *Nature Reviews Microbiology* 14(8):523–534, 2016. https://doi.org/10.1038/nrmicro.2016.81.
13. Sammut, C., and Webb, G.I. (eds.), Encyclopedia of machine learning and data mining. Springer, 2017. https://doi.org/10.1007/978-1-4899-7687-1.
14. Chen, J., Wu, L., Zhang, J., Zhang, L., Gong, D., Zhao, Y., Hu, S., Wang, Y., Hu, X., Zheng, B., Zhang, K., Wu, H., Dong, Z., Xu, Y., Zhu, Y., Chen, X., Yu, L., and Yu, H., Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: A prospective study. *medRxiv*, 2020. https://www.medrxiv.org/content/early/2020/03/01/2020.02.25.20021568.
15. Guo, Q., Li, M., Wang, C., Wang, P., Fang, Z., Tan, J., Wu, S., Xiao, Y., and Zhu, H., Host and infectivity prediction of wuhan 2019 novel coronavirus using deep learning algo- rithm. *bioRxiv* (2020), https://www.biorxiv.org/content/early/2020/02/02/2020.01.21.914044.
16. Dewey, M., and Schlattmann, P., Deep learning and medical diagnosis. *The Lancet* 394:1710–1711, 2019. https://doi.org/10.1016/S0140-6736(19)32498-5.
17. The Wall Street Journal, Coronavirus reveals limits of AI health tools (2020 (accessed February 29, 2020)), https://www.wsj.com/articles/coronavirus-reveals-limits-of-ai-health-tools-11582981201.
18. Bouguelia, M., Nowaczyk, S., Santosh, K. C., and Verikas, A., Agreeing to disagree: Active learning with noisy labels without crowdsourcing. *Int. J. Machine Learning & Cybernetics* 9(8):1307–1319, 2018. https://doi.org/10.1007/s13042-017-0645-0.
19. Technology Org, AI algorithm detects coronavirus infections in patients from CT scans with 96% accuracy (2020 (accessed March 02, 2020)), https://www.technology.org/2020/03/01/ai-algorithm-detects-coronavirus-infections-in-patients-from-ct-scans-with-96-accuracy/
20. Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., and Xia, L., Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: A report of 1014 cases. *Radiology* 0(0):200642, 2020. https://doi.org/10.1148/radiol.2020200642.
21. Bernheim, A., Mei, X., Huang, M., Yang, Y., Fayad, Z.A., Zhang, N., Diao, K., Lin, B., Zhu, X., Li, K., Li, S., Shan, H., Jacobi, A., and Chung, M., Chest ct findings in coronavirus disease-19 (covid-19): Relationship to duration of infection. *Radiology* 0(0):200463, 2020. https://doi.org/10.1148/radiol.2020200463.
22. Fong, S. J., Li, G., Dey, N., Crespo, R. G., and Herrera-Viedma, E., Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak. *International Journal of Interactive Multimedia and Artificial Intelligence* 6:51–61, 2020. https://doi.org/10.9781/ijimai.2020.02.005.