Furthering Machine Learning Applications in COVID-19 Research with Diverse Aggregated Datasets and Generalized Ensemble Model

Abhijit Gupta, Aikaterini Kargioti, Megha Joshi, Shayna Sragovicz

Abstract

The COVID-19 pandemic has posed issues on a worldwide scale. With current testing methods often lacking in accuracy and efficiency, novel artificial-intelligence-enhanced applications are proposed for COVID-19 diagnosis. Despite the encouraging results of recent literature, small and inconsistent datasets impede our ability to train and test machine learning models (ML models).

Efforts to develop strong ML models for specific COVID-19 applications can impact the overall understanding of disease transmission. However many of the conclusions derived from such models are found to have issues due to lack of consistent and high quality data. Therefore, it is important to homogenize this process so that data used in ML models is of a quality that will allow for meaningful results. Furthermore, the lack of broad scale data in terms of patient type, location, duration of disease, etc is often not accounted for, leading to variability in conclusions formed from current COVID-19 ML models.

This grant will propose two areas of investigation. First, the grant addresses the major discrepancies and inconsistencies in the current public datasets provided for machine learning research. Given the lack of standardization, comparisons between various models include a host of confounding variables stemming from the testing data itself. This grant proposes ways to aggregate, standardize, and pre-process data so that it is fit for training. The expected dataset will include a range of case severity, and the validation data will include a range of scans from healthy patients to patients with severe respiratory diseases other than COVID-19. In addition to collecting and organizing data, this grant introduces ways to consider the expected lack of diversity in the population representation of patients from minority and underrepresented groups. This grant advocates for transparency in test data reporting.

Additionally, this grant hypothesizes that an ensemble model leveraging prior models can attain higher performance metrics than the current baselines. Current COVID-19 ML models can be grouped across multiple input types (CT scans, X-Ray), multiple methods (ResNet, UNet, VNet, etc.), and multiple outputs (binary classification, severity, highlighted regions), and rely on unstandardized data sources and procedures. As a result, different models excel at different subsections of the general problem. By stacking heterogeneous models through ensemble learning, this grant aims to build a robust and generalizable model to classify and assess COVID-19 patients. Specific focus towards data standardization and preprocessing enables a scalable system that adapts as new models are developed.

The hope for this proposal is that through the development of standardized datasets and models, machine learning research will provide support in curbing the effects of this and future pandemics

Lay Abstract

Current COVID-19 testing methods often lack accuracy and efficiency. Artificial-intelligence-enhanced applications as a proposed method for COVID-19 diagnosis have shown encouraging results. However, small and inconsistent datasets limit our ability to create robust machine learning models (ML models). It is important to have a standardized process to identify issues with current data so that ML models can provide accurate results. Current models also do not include a wide range of quantitative and qualitative data which leads to a range of inaccurate conclusions.

The first goal of this grant is to ensure that all data being used in machine learning experiments is standardized and diversified. This grant aims to build a diverse dataset in terms of disease severity and demographic representations. Additionally, while multiple ML models exist to diagnose COVID-19 patients, they often struggle to generalize beyond their training data. This grant hypothesizes that combining the various models will improve robustness and highlight each model's strengths, improving prediction accuracy.

Introduction and Significance

The COVID -19 pandemic has strongly impacted the lives of people around the world. While the severity of the COVID-19 pandemic continues to concern health experts and public officials, there still remains to be found a conclusive and efficient method of diagnosing the virus. To this day lab testing yields false negatives especially at early steps of transmission [1], RT-PCR detection presents issues [2], and there exists concern about the affordability and accessibility of COVID-19 testing to the general public and especially minority groups. AI among other fields and technologies is investigated as a potential solution to those issues.

Current medical-imaging-based COVID-19 diagnosis methods generally involve a three-step process with CT or X-Ray scans. The process involves a pre-scan process when the patient is positioned with the assistance of a physician, an imaging process when the medical image is acquired in the form of raw data, and a diagnosis process when images are constructed based on the data and interpreted by a physician. Imaging devices have been used to enhance conventional image workflow. The addition of imaging devices such as cameras and other sensors in tomographic scans shows improvements in traditional image workflow. [3] AI as an additional enhancement to conventional imaging protocols allows via a combination of sensor data (from ToF, RGB, infrared and other sensors) and algorithms for the fast and efficient positioning of the patient and the accurate creation of a human mesh image during the imaging process. [4, 5, 6] It reduces the amount of patient-physician interaction while providing auditory and visual guidance for the patient. [7] Furthemore AI-ehnhansed medical imaging methods can reduce the amount of radiation exposure by accurately positioning the patient. [8] AI has also shown encouraging results in accurate scanning parameters identification including the start and end points of the medical imaging process for CT scans by determining the position of specific joints on the human body. [9]

There have been strong efforts to develop robust networks for COVID-19 specific applications. [10,11] Figure 1 summarizes the various types of ML networks utilized in COVID-19 medical applications as well as their datasets when available. Wang et al developed a novel approach for risk aware identification of suspected COVID-19 cases utilizing the Social Internet of Things. [12] Such applications can strongly impact our understanding of the viral transmission and impact the testing needed to be carried out. AI enhanced COVID-19 medical imaging applications provide a more accurate and efficient solution for the detection of COVID-19, while limiting the possibility of transmission via the physician-patient interaction. Zheng et al proposes a two-step weakly-supervised deep learning method that involves segmenting the image of the lung via a U-Net and assessing the infectiousness via a 3D deep neural network. [13] Cao et al utilized a U-Net convolutional network in a Deep Learning Quantitative CT Pipeline that allows a better qualification of COVID-19 via a more accurate CT segmentation of pulmonary opacities. [14] Narin et al ResNet50, InceptionV3 and Inception- ResNetV2 for the detection of COVID-19 pneumonia in radiographs of chest X-Ray images. [15] More recent work aims to provide more information to assist practitioners carrying out the diagnosis. Karim et al proposes the use of explainable deep neural networks for the end to end process of CXR images. [16]

Despite the encouraging results found in the literature there still remain issues with COVID-19 AI applications. The datasets available to researchers are limited, hindering the training and testing process for AI networks while also impending the accuracy of any proposed model. Furthermore, the breakdown of the studied populations included in the data sets raises questions regarding diversity and inclusion.

Purpose	Model	Dataset
Patient positioning	FAST Integrated Workflow [7]	
	Automatic Patient Centering for MDCT [8]	63 patients (36 men, 27 women, mean age: 51, age range: 22-83), chest CT: 18, abdominal: 45)
Body estimation	DARWIN: Deformable Patient Avatar Representation With Deep Image Network [4]	1063 patients from 3 different hospitals
	Patient MoCap [9]	180000 video frames

	Hierarchical Kinematic Human Mesh Recovery [5]	
Infectiousness assessment	Risk-Aware Identification of Highly Suspected COVID-19 Cases [12]	Hospital data of infected patients
Disease qualification	Deep learning-based detection for COVID-19 from chest CT [13]	Training: 499 CT volumes, testing: 131 CT volumes
	Deep Learning-based Quantitative CT Pipeline [14]	Training: CT images from 10 patients
	Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images [15]	50 patient images, 50 healthy individual images
	Automatic detection from X-Ray images utilizing Transfer Learning [16]	1427 X-Ray images (Covid-19: 224, common pneumonia: 700, normal conditions: 504)

Figure 1. Examples of ML models developed for different COVID-19 medical applications.

Rationale for Proposed Study

The outbreak of the COVID-19 pandemic has led to a major crisis in the healthcare industry. The virus not only poses a threat to individuals who have been around a carrier, but to healthcare workers who are constantly placing themselves in a precarious situation when treating these patients. It is important to streamline the workflow required to treat and diagnose COVID-19 to minimize threats that frontline healthcare workers are facing. One way to do this is to improve COVID-19 testing methods. While there are already a variety of networks created to predict trends and diagnose patients based on a variety of data, there is an overall absence of standardized and quality controlled data. Therefore it is important to democratize data related to COVID-19 that could be essential for the prediction of and trend pattern of the virus.

In order to facilitate the improvement of prediction and diagnosis networks, it is hypothesized that choosing the highest preform networks such as U-Net [50-55], UNet++, and VB-Net will enable standardization of the methodology to assess and diagnose the virus based on a variety of datasets. Furthermore these networks are widely used because of their ability to make conclusions from CT and X-ray scans which are more easily accessible around the world. A secondary aspect of this hypothesis is to assess the current data that exists related to COVID-19 and where potential issues are in the current identified data. It is then important to consider a diverse dataset that has quality controlled and accurate data as well as data that represents a variety of individuals to involve both the qualitative and quantitative aspects of diagnosis and trend prediction

The second hypothesis, then, is a robust and accurate dataset will facilitate the workflow between network-practitioner relationships. Qualitative data suggests that the use of networks in diagnosis and prediction is creating subtle yet meaningful changes in the relationships between healthcare practitioners and their patients. In the case of a pandemic, it is hypothesized that the use of artificial intelligence for diagnosis and prediction can limit the amount of contact a practitioner might have with a patient thereby limiting the threat it poses to others while still providing high quality diagnosis.

Specific Aim 1

The first aim of this grant proposal is to develop a robust and diverse dataset on which neural network models can be tested. A robust dataset is a large dataset in comparison to the existing public datasets with enough variability to inhibit generalization in a neural network's training. A diverse dataset is one with test data that represents a wide array of disease severity and demographics including patients from minority and underrepresented groups. A robust dataset will allow for more accurate model predictions, improving the reliability of the networks. A diverse dataset will ensure that all varieties of patients are able to be examined by the networks, regardless of their identity.

Specific Aim 2

A second aim of this grant proposal is to combine existing COVID-19 machine learning models to develop an ensemble model with improved accuracy and performance relative to current baselines. Many prior works have built strong models despite limited training data, using innovative and advanced machine learning theory. By combining these models in a systematic and data-driven way, the best aspects of each model can be highlighted to improve predictive power. Independent of a larger dataset, training ensemble models to leverage various models will improve COVID-19 classification and severity quantification.

Experimental Design

Specific Aim 1

Given the severity of the pandemic, efforts to aggregate and standardize data for neural network training is of the utmost importance when conducting artificial intelligence research. Standardized data will reduce the confounding variables that may impact the performance of a model. Instead of questioning the integrity of the dataset, targeted research can focus on more productive questions regarding a network's performance. Currently there exist a variety of public datasets, each with a variable quantity and quality of images. The COVIDGR dataset developed by Tabik et al. [17] contains 377 positive and negative images and the COVID-19 CT segmentation dataset [18] contains only about 100 images. Each dataset is hosted on its own virtual forum, and while it can be accessed publicly, differs widely from another public dataset. In order to improve efficiency and accuracy in training, it is vital that research includes a standardized, large dataset. To do this, we will develop an online platform to host image submissions and invite owners of already existing datasets to submit unprocessed data. However, with a variety of images being submitted, we assume that the data will need to be preprocessed to standardize each image. We plan to develop a method to pre-process all images submitted to the online database by adjusting the size and pixels of each picture. We will develop a clear metric to determine how to pre-process the images based on the picture's brightness, coloring, and clarity. Our online platform will include two collections of images: the first will consist of all raw images submitted by individual contributors for free and public access, and the second will consist of all pre-processed images. We urge users to test models using the pre-processed images in order to ensure the standardization of all data that is being trained.

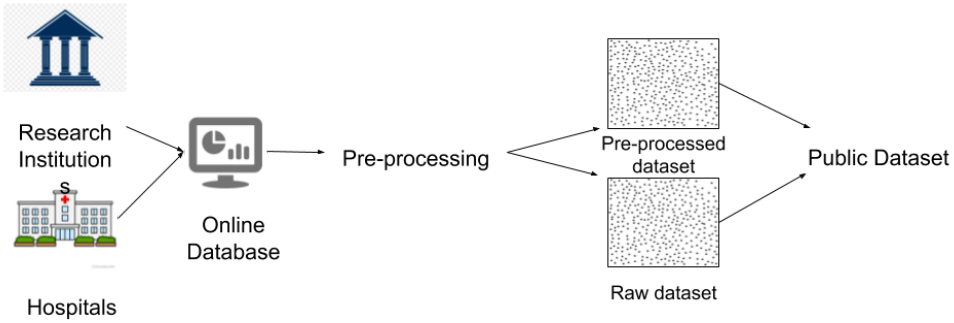

Figure 2. Expected aggregation method: a combination of research institutions and hospitals will submit data to an online database that is then preprocessed and made publicly available online in two forms: raw and preprocessed.

In addition to working directly with other research laboratories around the country, we also plan to communicate with public and private hospitals who are the source of the medical imaging. A significant obstacle in the success of neural network accuracy is diversity in training data. It is our moral, ethical, and scientific duty to ensure that our research benefits the entire population affected by the COVID-19 pandemic. To do this, we plan to be wholly transparent and communicative about the source and representation of our dataset. We will include statistics regarding the demographics of our dataset and will dedicate resources to ensuring the diversity of our dataset. In an effort to consider diversity in the public dataset, we recognize that hospitals that release images will need to be in compliance with Health Insurance

Portability and Accountability Act (HIPAA) and privacy laws regarding the identity of their patients. In order to respect privacy laws, we will be allocating funds for a legal advisor in order to develop a lawful way to gather demographic statistics, in whatever form possible. Regardless of the format of our diversity reporting, we intend to be clear and transparent about the population from which our datasets are collected.

Specific Aim 2

Independent of aggregating a larger, more diverse, dataset, this grant aims to consolidate current leading COVID-19 machine learning models and use an ensemble model to improve prediction metrics relative to current baselines. Current COVID-19 ML models can be grouped across multiple input types (CT scans, X-Ray), multiple methods (ResNet, UNet, VNet, etc.) and multiple outputs (binary classification, severity, highlighted regions), and rely on unstandardized data sources and procedures. Consequently, model performance often does not generalize beyond the specific subproblem addressed, rendering these black box functions potentially dangerous for real-world applications. On the flip side, each model has been engineered and trained with a specific optimization in mind, and thus different models excel at different subsections of the general problem.

By stacking heterogeneous models through ensemble learning, this grant aims to build a robust and generalizable model to classify and assess COVID-19 patients. Such a model would take in a variety of inputs such as CT-scans, X-Rays, demographic data, etc. and produce results including binary classification (does a patient have COVID-19), severity (how likely will they need a ventilator, ICU bed), and highlighted regions (where should a health practitioner look to assess the situation firsthand). Specific focus towards data standardization and preprocessing will enable a scalable system that adapts as new models are developed. As a starting point, the models listed in Figure 1 can be included, in addition to models listed in [11].

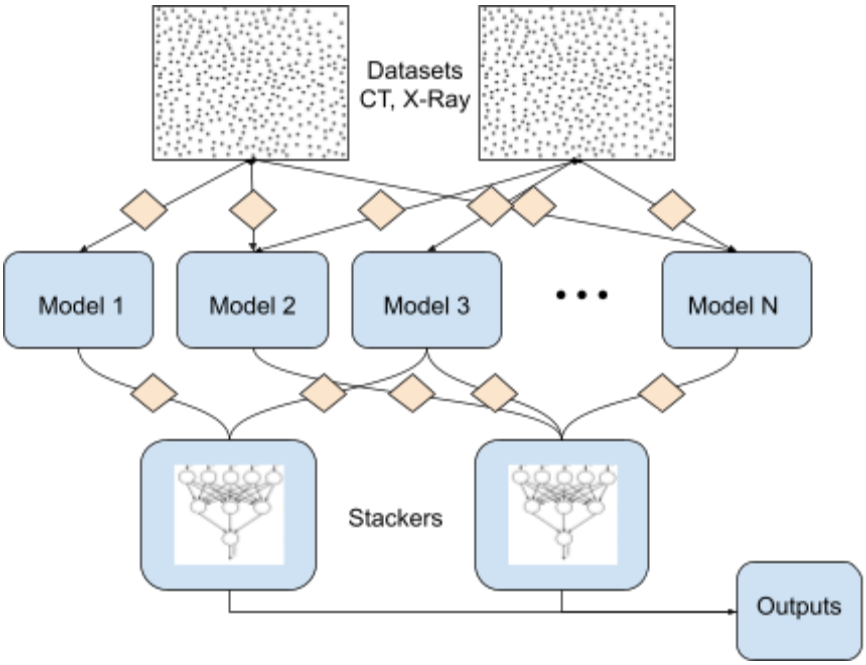

Figure 3. Preliminary ensemble model structure. White rectangles represent different datasets, blue rounded rectangles represent machine learning models, orange diamonds represent data transforms, arrows show data flow.

Figure 3 above demonstrates the broad implementation of the ensemble model in top-down view. To begin with, datasets of CT scans and X-ray images are aggregated from patients with COVID-19 of various severities, patients with separate respiratory diseases, and healthy patients. Depending on the success of Specific Aim 1, this grant's aggregated datasets could be used, else prior collected datasets such as from [4], [8], [13]. In the next layer, a set of N machine learning models are collected. Each model should take in imaging data and produce one of many outputs including binary classification, severity analysis, or lesion localization. These models will be taken from prior works such as [12], [14],

[15]. To ensure model quality, a set of metrics will be compiled that individual models must satisfy to be included in the ensemble ML model.

A particular model within the ensemble model may take in either CT scan, X-Ray images, or both. Depending on the specific input parameters, connections to the two datasets are made and data is loaded in. As the ensemble model brings in models from multiple sources, the specific input dimensions may vary model to model, so preprocessing data transforms will be created connecting any dataset-model pair. In this manner, the ensemble model can easily scale as new ML models are introduced, regardless of the specific input formats required. This is demonstrated in Figure 2, where Model 1 only uses CT scan input, Model 3 uses X-Ray, and Model 2 uses both sources.

The outputs from the N models can be primarily characterized as classification or numeric severity. Depending on the output type, the result is fed into one of two neural network 'stacks', again with data transforms as necessary to ensure homogeneity. Each neural network accepts as input the individual models' output, and mines for underlying patterns to arrive at a single prediction. For the binary classification stack, this will be a categorical result of COVID-19 or other. For the numerical stack, this will be a likelihood of ventilator need. Note some individual models may produce both output types, in which case the data is split and sent to both stacks (Model 3 in Figure 2). Finally, the two outputs are combined and displayed as the output of the ensemble model.

Potential Pitfalls

Aim One is entirely data-driven, and relies on the quality of data inputted in order to have an accurately trained system. A large issue with this is mislabelling of datasets. More specifically, the mislabelling of COVID-19 and pneumonia, which has currently been a problem in testing and labeling of data [19]. The solution that is provided to this foreseen pitfall is to have health experts choose random points to test to ensure that the quality of data used is up to a certain standard with little to no error.

Another aspect of the Aim One that could be addressed is that there are a variety of data types that the models will be working with. They are primarily divided into two types of scans: those that use CT scans and those that use X-rays. The problem to consider is whether the models are given an equal opportunity to report their accuracy based on the data they are given. In this case, consistency and reliability are important to control. While it is challenging to gather an equal amount of data for both types of scans, to combat the inconsistency in the amount of data there will be reporting ratios for each model to ensure transparency in analysis. The final pitfall with Aim One is the types of patient data that are getting reported which would most likely lead to a lack of diversity in the data. In order to develop conclusive results it is important to gather a diverse set of data which includes patients from minority communities and certifying that the range of severity is accounted for with each patient dataset received. The solution for this is once again transparency. The ambiguity or lack of important qualitative data will be noted and efforts will be made to receive more information about race and age from hospitals where data is gathered.

When utilizing machine learning models combined in an ensemble model as presented in Aim Two there are pitfalls to consider. The first being splitting data inappropriately. When creating training sets to teach the model, data is often split at random. However real life data is rarely so random and most likely contains a variety of changes based on variables and historical trends. The second problem, then, is the possibility of a hidden variable. Hidden variables can occur due to the lack of data presented and the way that the testing for COVID-19 patients was conducted, most likely not in a uniform way [20]. Finally if less than 12 models are present then a simple average is preferred. If there are greater than 12 models then a weighted average or neural network is required [21]. This would be hard to achieve given the limited amount of data and networks available for COVID-19 and therefore would require additional work and analysis.

Furthermore with ensemble models, many weaker models may be drowned out when there is a range of weaker models and robust models. In order to combat this it is important to consider the models that are being included and to define a criteria for why each model is being included. If it is not necessary to the ensemble model or has roles that are overlapped by stronger models, then it should not be included in the ensemble.

Acknowledgements

The authors of this grant would like to thank Yale University for funding this research, the 2020 Yale Summer Online Research Workshop for guidance on reading and writing scientific literature, Dr. Belperron for leading the workshop, and Dean Chang for supervising the course.

References

1. Lauren M. Kucirka, Stephen A. Lauer, Oliver Laeyendecker, Denali Boon, Justin Lessler, (2020), Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure, *Annals of Internal Medicine*
2. Tahamtan A, Ardebili A, (2020), Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert Rev Mol Diagn*, vol. 20(5), 453-454.
3. V. Singh, Y.-J. Chang, K. Ma, M. Wels, G. Soza, and T. Chen, (2014), "Estimating a patient surface model for optimizing the medical scanning workflow," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 472-479.
4. V. K. Singh, K. Ma, B. Tamersoy, Y. Chang, A. Wimmer, T. F. Odonnell*, et al.*, (2017), DARWIN: Deformable patient avatar representation with deep image network, *Medical Image Computing and Computer Assisted Intervention*, 497-504.
5. R. Li, C. Cai, G. Georgakis, S. Karanam, T. Chen, and Z. Wu, (2019), Towards robust RGB-D human mesh recovery
6. G. Georgakis, R. Li, S. Karanam, T. Chen, J. Kosecka, and Z. Wu, (2020), Hierarchical hinematic human mesh recovery
7. *Siemens CT scanner SOMATOM Force, SOMATOM Drive or SOMATOM Edge Plus*. Available: https://www.siemens-healthineers.com/computed-tomography/technologies-and-innovations/fast-integrated-workflow
8. J. Li, U. K. Udayasankar, T. L. Toth, J. Seamans, W. C. Small, and M. K. Kalra, (2007), Automatic patient centering for MDCT: effect on radiation dose, *American Journal of Roentgenology,* vol. 188, 547-552.
9. F. Achilles, A. E. Ichim, H. Coskun, F. Tombari, S. Noachtar, and N. Navab, (2016), Patient MoCap: Human pose estimation under blanket occlusion for hospital monitoring applications, *Medical Image Computing and Computer Assisted Intervention*, 491-499.
10. L. A. Bullock Joseph, Pham Katherine Hoffmann, Lam Cynthia, Luengo-Oroz Miguel A., (2020), Mapping the landscape of artificial intelligence applications against COVID-19
11. Shi,F.,Wang,J.,Shi,J.,Wu,Z.,Wang,Q.,Tang,Z.,He,K.,Shi,Y.,Shen,D, (2020), Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE Reviews in Biomedical Engineering
12. B. Wang, Y. Sun, T. Q. Duong, L. D. Nguyen and L. Hanzo, (2020), Risk-Aware Identification of Highly Suspected COVID-19 Cases in Social IoT: A Joint Graph Theory and Reinforcement Learning Approach, IEEE Access, vol. 8, 115655-115661.
13. C. Zheng, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma*, et al.*, (2020), Deep learning-based detection for COVID-19 from chest CT using weak label, *MedRxiv*
14. Y. Cao, Z. Xu, J. Feng, C. Jin, X. Han, H. Wu*, et al.*, (2020), Longitudinal assessment of COVID-19 using a deep learning–based quantitative CT pipeline: Illustration of two cases, *Radiology: Cardiothoracic Imaging,* vol. 2, e200082.
15. A. Narin, C. Kaya, and Z. Pamuk, Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks, (2020), *arXiv:2003.10849*.
16. M. Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, O. Beyan, et al., (2020), Deepcovidexplainer: Explainable covid-19 predictions based on chest x-ray images, arXiv preprint arXiv:2004.04582.
17. S Tabik, A Gómez-Ríos, JL Martín-Rodríguez, I Sevillano-García, M Rey-Area, D Charte, E Guirado, JL Suárez, J Luengo, MA Valero-González, et al, (2020), COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on Chest X-Ray images. arXiv preprint arXiv:2006.01409 (2020).
18. Sakinis, Tomas, and Håvard Bjørke Jenssen, (2020), Free Medical Segmentation Online. *MedSeg*, www.medseg.ai/.
19. Samsami, Majid, et al. "COVID-19 Pneumonia in Asymptomatic Trauma Patients; Report of 8 Cases." *Archives of Academic Emergency Medicine*, Shahid Beheshti University of Medical Sciences, 6 Apr. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7158871/.

20. Riley, Patrick. “Three Pitfalls to Avoid in Machine Learning.” *Nature News*, Nature Publishing Group, 30 July 2019, www.nature.com/articles/d41586-019-02307-y.
21. Heeswijk, Mark van, et al. “Adaptive Ensemble Models of Extreme Learning Machines for Time Series Prediction.” *SpringerLink*, Springer, Berlin, Heidelberg, 14 Sept. 2009, link.springer.com/chapter/10.1007/978-3-642-04277-5_31.