Mapping the landscape of Artificial Intelligence applications against COVID-19

Joseph Bullock[1,2‡], Alexandra Luccioni[3‡], Katherine Hoffmann Pham[1,4‡],
Cynthia Sin Nga Lam[5], Miguel Luengo-Oroz[1]

1 United Nations Global Pulse, New York, NY, USA
2 Institute for Data Science, Durham University, Durham, United Kingdom
3 Mila Québec Artificial Intelligence Institute, Université de Montréal, Montréal, Québec, Canada
4 NYU Stern School of Business, New York, NY, USA
5 Global Coordination Mechanism on NCDs, World Health Organization, Geneva, Switzerland

‡These authors contributed equally to this work.
Contact: joseph@unglobalpulse.org, sasha.luccioni@mila.quebec, katherine@unglobalpulse.org, lams@who.int, miguel@unglobalpulse.org

Abstract

COVID-19, the disease caused by the SARS-CoV-2 virus, has been declared a pandemic by the World Health Organization, with over 2.5 million confirmed cases as of April 23, 2020 [1]. In this review, we present an overview of recent studies using Machine Learning and, more broadly, Artificial Intelligence, to tackle many aspects of the COVID-19 crisis at different scales including molecular, clinical, and societal applications. We also review datasets, tools, and resources needed to facilitate AI research. Finally, we discuss strategic considerations related to the operational implementation of projects, multidisciplinary partnerships, and open science. We highlight the need for international cooperation to maximize the potential of AI in this and future pandemics.

Executive Summary

- There is a broad range of potential applications of AI covering medical and societal challenges created by the COVID-19 pandemic; however, few of them are currently mature enough to show operational impact.
- From a molecular perspective, AI can be used to: estimate the structure of SARS-CoV-2-related proteins, identify existing drugs that may be repurposed to treat the virus, propose new compounds that may be promising for drug development, identify potential vaccine targets, improve diagnosis, and better understand virus infectivity and severity.
- From a clinical perspective, AI can support COVID-19 diagnosis from medical imaging, provide alternative ways to track disease evolution using non-invasive devices, and generate predictions on patient outcomes based on multiple data inputs including electronic health records.
- From a societal perspective, AI has been applied in several areas of epidemiological research modeling empirical data, including forecasting the number cases given different public policy choices. Other works use AI to identify similarities and differences in the evolution of the pandemic between regions. AI can also help investigate the scale and spread of the "infodemic" to address the propagation of misinformation and disinformation including the emergence of hate speech.
- Sharing and hosting data and models, whether they be clinical, molecular, or scientific, is critical to accelerate the development and operationalization of AI to support the response to the COVID-19 pandemic.
- Applications targeting critical applications – such as clinical ones – should take into account existing regulatory and quality frameworks to ensure the validity of use and safety as well as minimize potential risks and harms.
- International AI cooperation based on multidisciplinary research and open science is needed to accelerate the translation of research into global solutions which can be tailored and adapted to local contexts.

Introduction

With the continued growth of the COVID-19 pandemic, researchers worldwide are working to better understand, mitigate, and suppress its spread. Key areas of research include studying COVID-19 transmission, facilitating its detection, developing possible vaccines and treatments, and understanding the socio-economic impacts of the pandemic. In this article, we discuss how Artificial Intelligence can contribute to these goals by enhancing ongoing research efforts, improving the efficiency and speed of existing approaches, and proposing original lines of research. We have conducted an extensive review of the rapidly emerging literature and identified specific applications of AI at three different scales: the molecular scale, including drug discovery-related research; the clinical scale, including individual patient diagnosis and treatment; and the societal scale, including epidemiological and infodemic research. We also review open-source datasets and resources that are available to facilitate the development of AI solutions.

The mobilization of the scientific community to address the pandemic is unprecedented in its scale. An automated search for papers posted on the ArXiv and in the COVID-19 Open Research Dataset (CORD-19) identified over 4,500 coronavirus-related papers posted between January 1 and April 5, 2020. Of these papers, over 100 included the phrases "machine learning", "artificial intelligence", "deep learning", or "neural network" in the title or abstract. As shown in Figure 1, the number of papers has grown dramatically since mid-March 2020.

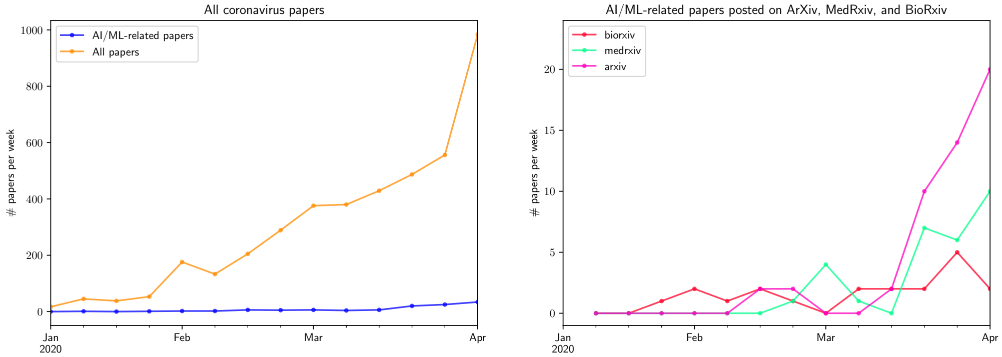

Fig 1. The publication of scientific articles and preprints related to COVID-19 between January 2 and April 5, 2020. Article counts were derived from the CORD-19 research dataset and the ArXiv API. We omitted 2020 articles with no specific publication date and articles published on January 1, 2020 since this appeared to be a default date for many articles; we also dropped CORD-19 articles missing both a title and summary. Note that the y axis scales differ between plots.

The purpose of this review is not to evaluate the impact of the described techniques, nor to recommend their use, but to show the reader the extent of existing applications and to provide an initial picture and road map of how Artificial Intelligence could help the global response to the COVID-19 pandemic. The scope of this review is restricted to applications of Machine Learning (ML) and Artificial Intelligence (AI), and we have therefore made judgment calls regarding whether certain methodologies fall into this category. For example, we have included applications where authors have explicitly described the use of models such as neural networks and decision trees, while excluding applications based on simple linear regression models. Furthermore, we note that many of the articles cited are still preprints at the time of writing this review; given the fast-moving nature of the crisis we strove to be comprehensive in our coverage, but their full scientific rigor should still be assessed by the scientific community through peer-reviewed evaluation and other quality control mechanisms. For specificity, we signify all preprints with † . Finally, since this article assumes a certain background knowledge of both Machine Learning and the nature of the SARS-Cov-2 virus, we invite our readers to consult [2]† for further explanation regarding the potential of ML for scientific research, and [3]† for additional information about the virology, clinical features, and epidemiology of COVID-19. Accessible overviews of SARS-CoV-2 proteins, the infection process, and molecular modeling can be found in [4–6].

Molecular scale: From proteins to drug development

At the most granular scale of the scientific response to COVID-19, biochemistry applications of AI have been used to better understand the proteins involved in SARS-Cov-2 infection and to inform the search for potential treatments. With respect to the virus itself, four types of structural proteins are of interest: nucleocapsid proteins (N), envelope proteins (E), membrane proteins (M), and spike proteins (S) [7,8]† . Also of interest are a number of non-structural proteins (NSPs), which are crucial for viral pathogenesis, including the 3-chymotrypsin-like (3C-like) protease (also known as 3CLPro, the main protease/MPro, or nsp5) and the papain-like protease (PLpro, part of nsp3). On the human side, research has focused on the angiotensin-converting enzyme 2 (ACE2) protein, a receptor that facilitates the virus' entry into host cells [9]. Potential applications of AI on this scale include predicting the structure of these associated proteins, identifying existing drugs which may be effective in targeting these proteins, and proposing new chemical compounds for further testing as potential treatments [10].

Protein structure prediction

Proteins have a 3D structure, which is determined by their genetically encoded amino acid sequence, and this structure influences the role and function of the protein [11]. Protein structure is traditionally determined through experimental approaches such as X-ray crystallography, but these can be costly and time-consuming. More recently, computational models have been used to predict protein structure. There are two primary approaches to the prediction task: *template modeling*, which predicts structure using similar proteins as a template sequence, and *template-free modeling*, which predicts structure for proteins that have no known related structure.

Senior *et al.* [12] have developed a system called AlphaFold which focuses on the latter challenge. The AlphaFold model is based on a dilated ResNet architecture [13, 14] and uses amino acid sequences, as well as features extracted from similar amino acid sequences using multiple sequence alignment (MSA), to predict the distance and the distribution of angles between amino acid residues. These predictions are used to construct a "potential of mean force" which is used to characterize the protein's shape [15]. This system has been applied to predict the structures of six proteins related to SARS-CoV-2 (the membrane protein, protein 3a, nsp2, nsp4, nsp6, and papain-like protease) [16]† .

Heo and Feig [17]† also use a dilated ResNet architecture, implemented as part of the transform-restrained Rosetta (trRosetta) [18]† pipeline, to predict the structure of the above proteins as well as proteins ORF6, ORF7b, ORF8, and ORF10. The trRosetta network has multiple output heads: one which predicts the distances between residues in a protein's amino acid sequence, and others which predict the orientations between these residues as characterized by five different angles. This approach may allow for better performance by jointly learning features that are relevant to predicting both distance and orientation [18]. Heo and Feig refine trRosetta and AlphaFold's predicted structures using molecular dynamics simulations and compare the results with structure predictions from a third approach – C-I-TASSER [19] – which incorporates nine different methods for contact prediction. While the authors find that the predicted structures generally have much variability between them, there is some consensus for predicted structures of the papain-like protease, part of nsp4, and the M protein.

Drug repurposing

In addition to better understanding the structure of key proteins involved in SARS-CoV-2 infection, a number of research efforts have focused on identifying known compounds which might be effective in mitigating infection – including, potentially, already-approved drugs. We have identified four distinct approaches to this problem, which are facilitated by AI: the construction of biomedical knowledge graphs, the prediction of protein-ligand binding affinities, molecular docking simulations, and the analysis of gene expression signatures.

Biomedical knowledge graphs

Biomedical knowledge graphs are networks capturing the relationships between different entities – such as proteins and drugs – in order to facilitate higher-level exploration of how they connect. Richardson *et al.* [20] use this technique to identify Baricitinib, a drug which is commonly used to treat arthritis via inhibition of JAK1/2 kinases, as a promising therapy for COVID-19 because it inhibits the AP2-associated protein kinase 1 (AAK1) enzyme and may, therefore, make it harder for the virus to infect host cells. Related work has described two approaches which potentially inform the graph construction. First, Segler *et al.* [21] describe an approach to mining a structured database of chemical reactions (Reaxys) using a three-part neural network pipeline combined with a Monte Carlo Tree Search approach (3N-MCTS), in order to understand how various compounds are formed hierarchically from reactions between simpler component compounds. Second, Fauquer *et al.* [22] describe a strategy for mining an unstructured scientific article database (PubMed) to identify stylized relationships between gene-disease pairs expressed in individual sentences (e.g. "GENE promotes DISEASE").

Ge *et al.* [23]† describe a similar approach to constructing a knowledge graph connecting human proteins, viral proteins, and drugs using databases that capture the relationships between these entities. The graph is used to predict potentially effective candidate drugs. This list is further refined using a Natural Language Processing (NLP) model, i.e. a Biomedical Entity Relation Extraction (BERE) approach [24]† applied to the PubMed database, filtered for mentions of the candidate drug compounds, coronaviruses, or their associated proteins. The authors identify a Poly (ADP-Ribose) Polymerase 1 (PARP1) inhibitor, CVL218, as a promising candidate, and it is currently undergoing clinical testing.

Prediction of protein-ligand binding affinities

Other studies attempt to predict protein-ligand binding affinities in order to tackle the drug repurposing problem. Ligands are small molecules which bind with a protein to trigger a signal, which can be activation or inhibition. Hu *et al.* [25]† use a multitask neural network to predict these affinities, identifying a list of 8 SARS-CoV-2 related proteins which they attempt to target using a database of 4,895 drugs. They suggest 10 promising drugs, along with their target proteins and binding affinity scores (which indicate the likelihood that the drug will act as an inhibitor). In an attempt to increase model interpretability, they also estimate the precise regions of each target protein where binding is likely to occur.

In a similar vein, Zhang *et al.* [26]† use their dense fully connected neural network architecture, trained to predict binding affinities on the PDBbind database, in order to identify potential inhibitors of the 3C-like protease. They develop a homology (template) model of the target protein using its SARS variant, and explore databases of existing compounds (e.g. ChemDiv and TargetMol) as well as tripeptides to find treatments which may be effective at targeting this protein. Nguyen *et al.* [27]† also build a SARS-based homology model of the 3C-like protease, and apply their Mathematical Deep Learning (MathDL) approach to identify potential inhibitors for this protease. In particular, their model uses mathematical representations of molecules as inputs, and relies on two main datasets: information on 84 SARS coronavirus protease inhibitors from the ChEMBL database, and a more general set of 15,843 protein-ligand binding affinities from the PDBbind database. They fit two different convolutional neural networks (CNNs) [28] on this dataset - a pooled (3DALL) CNN which is trained on both datasets together, and a multitask (3DMT) CNN which is trained on each dataset separately [29]. Using a consensus between these CNN models, the authors identify a list of 15 promising drug candidates from the DrugBank dataset.

Finally, Beck *et al.* [30]† use their own Molecule Transformer-Drug Target Interaction (MT-DTI) model of binding affinities to identify US Food and Drug Administration (FDA) approved antivirals which may be effective in targeting six coronavirus-related proteins (the 3C-like protease, RNA-dependent RNA polymerase, helicase, 3'-to-5' exonuclease, endoRNAse, and 2'-O-ribose methyltransferase). The MT-DTI model ingests string data in the form of simplified molecular-input line-entry system (SMILES) data and amino acid sequences, and applies a text-modeling approach that leverages ideas from the BERT algorithm [31]. The model identifies drugs that are expected to be effective in targeting each protein studied. Hofmarcher *et al.* [32]† likewise apply a text-based approach to SMILES data (ChemAI), which in turn relies on a Long Short-Term Memory (LSTM) [33] model called SmilesLSTM, to screen almost 900 million compounds from the ZINC database for effectiveness in inhibiting the SARS coronavirus 3C-like protease and the papain-like protease. They rank compounds according to predicted inhibitory effects, toxicity, and proximity to known compounds, and produce a list of 30,000 candidate compounds for screening.

Docking simulations

Another approach to drug repurposing and discovery involves molecular docking simulations. In a docking simulation, a wide range of candidate ligands interact with a protein in different orientations and conformations, generating a variety of poses (also known as the binding modes - i.e. the resulting interactions between the ligand and the protein as they bind). The poses are subsequently scored and used to predict the ligand's binding affinity. Since docking simulations are computationally expensive, some research has studied how to make the search more efficient by narrowing the pool of candidates which must be docked. For example, Ton *et al.* [34] develop a Deep Docking (DD) platform which trains a neural network to predict the outcomes of docking simulations, which they use to identify a set of 3 million candidate 3C-like protease inhibitors from a set of over 1 billion compounds extracted from the ZINC database. The authors then run a full docking simulation on the resulting compounds, presenting the top 1,000 results. On the other hand, Batra *et al.* [35]† train a random forest algorithm on SMILES data to predict binding affinities which would result from docking simulations and use this approach to select 187 promising molecules to target the coronavirus S-protein and the ACE2 receptor for the final docking simulation. They also identify 19,000 additional candidate compounds in the BindingDB dataset.

Gene expression signatures

A fourth approach to drug repurposing involves discovering therapies which have similar effects to other known effective treatments. To this end, Donner *et al.* [36] use the LINCS dataset of gene expressions from cells targeted by various perturbagens (chemical or genetic reagents to treat cells and alter intracellular processes). They learn an embedding with a deep neural network classifier that predicts the perturbagen associated with each signature (i.e., the gene expression that is specifically correlated with a biological state of interest, such as therapeutic response). In order to correctly classify signatures associated with the same perturbagen, the learned embedding should abstract away from the noise in the input data and identify core features that are associated with a perturbagen's effect. Following this logic, their approach can utilize similarity in the learned embedding space to predict pharmacological similarities in structurally different compounds, and hence expand the horizon of drug repurposing. Avchaciov *et al.* [37]† adapt this approach to find drugs that produce gene expression signatures that are similar to the COBP2 gene knockout, which might limit the replication of SARS-CoV-2 based on the gene's role in the replication of the related SARS coronavirus. They list twenty of the most promising drugs, many of which have already been identified as antivirals; since these drugs have been authorized for clinical trials or already approved, the authors argue that their approach could facilitate the rapid discovery of potentially effective therapies.

Drug discovery

Some research attempts to discover entirely new compounds for use in targeting SARS-Cov-2. Zhavoronkov *et al.* [38]† use a proprietary pipeline to find inhibitors for the 3C-like protease. Their models use three types of input: the crystal structure of the protein, the co-crystalized ligands, and the homology model of the protein. For each input type, the authors fit 28 different models, including Generative Autoencoders [39] and Generative Adversarial Networks [40]. The authors explore potential candidates using a reinforcement learning approach with a reward function that incorporates factors such as measures of drug-likeness, novelty, and diversity. Moreover, they confirm that the identified candidate molecules are dissimilar to existing compounds, suggesting that they have indeed found novel candidate drugs.

Tang *et al.* [41]† also apply a reinforcement learning approach to the discovery of compounds that inhibit the 3C-like protease. Specifically, the authors create a list of 284 molecules known to act as inhibitors in the context of SARS. They break these proteins into a series of 316 fragments, which can then be combined using an advanced deep Q-learning network with fragment-based drug design (ADQN-FBDD) that rewards three aspects of discovered molecules: a drug-likeness score, the inclusion of pre-determined "favorable" fragments, and the presence of known pharmacophores (which are abstract design patterns believed to be correlated with a compound's effectiveness [42]). The

4,922 results are heuristically filtered, and the 47 top compounds are assessed with molecular docking simulations, from which the researchers then select the top most promising compound and manually tailor it to produce suggested variants for testing.

In a third approach, Bung *et al.* [43][†] build a generative model for SMILES input strings. Treating the strings as a time series of characters, the model is a classifier that predicts the next character in the string. The model is first trained on 1.6 million molecules from the ChEMBL database, and then adapted to a smaller dataset of protease inhibitors using transfer learning. Reinforcement learning was then used to train the model to generate compounds with desirable properties. After filtering the resulting molecules and docking them, the authors propose 31 candidate inhibitors.

Finally, Gao *et al.* [29][†] apply a generative network complex (GNC) for drug discovery. Their pipeline involves gated recurrent unit (GRU) based encoders and decoders which ingest SMILES strings and propose new variants with the help of a DNN between the encoder and decoder that optimizes candidate variants. They also use a pretrained 2D fingerprint-based DNN (2DFP-DNN) as well as MathDL CNNs (described briefly above) to further predict the properties of the resulting drugs. The authors identify 15 novel candidate drugs and also test two proposed HIV drugs to estimate their efficacy against SARS-CoV-2. A similar approach to drug discovery is described in Chenthamarakshan *et al.* [44][†] ; the authors' Controlled Generation of Molecules (CogMol) framework uses a variational autoencoder (VAE) trained on SMILES strings to learn molecule embeddings. On these embeddings, the authors train a model to predict drug properties and protein binding affinities and use this to constrain the search for novel strings using Conditional Latent (attribute) Space Sampling (CLaSS). The authors also use a multitask DNN to predict toxicity in order to avoid proposing candidate drugs with a low probability of success later on in the testing pipeline. The authors focus their search on drugs which target nsp9, the 3C-like protease, and the receptor-binding domain (RBD) associated with the S protein, proposing 3,000 of the top candidates for further study.

In the process of mounting an immune response, B-cells in the body produce antibodies, which attack the part of the pathogen (the virus) known as an antigen. Magar *et al.* [45][†] thus take a different approach to discovering new therapies in which they search for antigen-neutralizing antibodies. They first construct a training dataset (VirusNet) consisting of 1,933 known antigen-antibody sequences from related illnesses such as HIV, SARS, Ebola, and influenza. The authors then train classification models such as XGBoost [46] on graph embeddings of these antigens and antibodies to predict whether an antibody will have a neutralizing effect on an antigen. Finally, the authors mutate the SARS coronavirus antibody sequence to generate 2,589 candidate antibody sequences. Given the subset of these antibodies which are predicted to be effective by the algorithm, they filter these mutations for valid and stable variants (which they identify through the use of molecular dynamics simulations), proposing 8 antibodies as potentially effective treatments.

Vaccine discovery

Another area of interest is vaccine discovery. In addition to producing virus-neutralizing antibodies via B-cells as described above (humoral immunity), the body also uses T-cells to attack the virus directly (cellular immunity). There is a subset of T-cells called memory cells which recognize the antigen of a formerly eliminated pathogen, and can quickly activate more effector T-cells upon re-exposure. These processes inform targets for vaccine design. As part of the response, helper proteins called major histocompatibility complex proteins (MHC I and MHC II proteins) present binding regions of antigens, called epitopes, for antibodies, B-cells, or T-cells to bind and attack. These MHC I and MHC II proteins are encoded by Human Leukocyte Antigen (HLA) gene complexes, which vary from person to person. Therefore, vaccine design involves two key objectives: (1) identifying suitable epitopes for targeting, and (2) ensuring that these epitopes can be presented by MHC proteins which are produced by different HLA alleles (variants of a gene) that occur in the population.

For example, Fast *et al.* [47] search for B- and T-cell epitopes. They identify 405 potential T-cell epitopes that can be presented by MHC I and MHC II proteins, as well as two B-cell epitopes on the S-protein. The search for the T-cell epitopes relies on two previously-developed neural networks to predict MHC presentation (NetMHCPan4 and MARIA). Upon identifying potential epitopes, the authors examine 68 different genetic variants of SARS-CoV-2 to study how the virus mutates, and identify parts of the virus that are more or less prone to evolution. They conclude that the S-protein epitopes may be a good target for vaccines because they contained no nearby mutations in their sample. In an alternative approach, Ong *et al.* [48][†] use their Vaxign-ML framework, which leverages supervised classification models such as XGBoost [46], in an effort to predict which viral proteins may serve as effective vaccine targets. While the authors find that the S-protein is the best candidate, they also identify five possible NSPs – most promisingly, nsp3 and nsp8 – as good candidates for the vaccine target.

Improving viral nucleic acid testing

Researchers are also applying Machine Learning in an attempt to improve the current virus nucleic acid detection test. Metsky *et al.* [49][†] combine ML with CRISPR (a tool which uses an enzyme to edit genomes by cleaving specific strands of genetic code) to develop assay designs for detecting 67 respiratory viruses, including SARS-CoV-2. The authors note that this technology can speed up the processing of test samples in order to assist overburdened diagnostic facilities, and help address the challenge of false positives, which occur as a result of sequence similarity between SARS-Cov-2 and other coronaviruses. ML models have been built to rapidly design assays which are predicted to be sensitive and specific, and cover a diverse range of genomes. The authors state that they are aiming to build a Cas13-based point-of-care assay for SARS-CoV-2 in the future.

Lopez-Rincon *et al.* [50] take another approach, in which they employ a Convolutional Neural Network (CNN) model to nucleic acid sequences to classify whether they are associated with SARS-CoV-2 and therefore potentially improve the specificity of diagnosis. They contrast SARS-CoV-2 with other human coronaviruses from the 2019nCoVR repository, as well as other genome sequences with the ORF1ab protein from GenBank. The authors use a 21-base pair convolution over the whole genome (as opposed to previous approaches which only examine sequences of fixed length) and visualize the network's convolution and max-pooling layers to understand which particular sequences help to identify SARS-CoV-2. Using the 21-base-pair sequences retained after the max-pooling layer, they subsequently fit a classification model (e.g., logistic regression) for the same task, using feature selection to identify the most predictive sequences. They also apply this classification approach to distinguish between hospitalized and asymptomatic cases. Using just 12 21-base-pair sequences, the authors are able to identify SARS-CoV-2 with over 99% accuracy and classify asymptomatic cases with 85% accuracy. Although the data of the study was limited, this work highlights that opportunities lie in bioinformatic processes for researchers to improve existing diagnostic tools.

Better understanding infection

Additional efforts have used Machine Learning to better understand SARS-CoV-2 infection severity and infectivity (how likely it is that a pathogen can infect a host) using protein sequences. For example, Gussow *et al.* [51]† use Support Vector Machines (SVM) [52] on genomes from different coronaviruses to identify which parts of coronavirus' protein sequences distinguish high case fatality rate (high-CFR) from low-CFR variants. Bartoszewicz *et al.* [53]† use reverse-complement neural networks built from CNN and LSTM architectures to detect whether a virus has the potential to infect a human host using its viral genome sequence, and apply machine learning interpretability techniques to identify the parts of the sequence that are most associated with infectivity. Finally, Randhawa *et al.* [54] use a Machine Learning with Digital Signal Processing (ML-DSP) approach which uses supervised learning approaches such as SVM [52] and K-nearest neighbors (KNN), as well as decision trees, to predict the taxonomic classifications of viruses based on their genomic sequences. Their findings support the classification of SARS-CoV-2 as a sarbecovirus of the betacoronavirus class and the hypothesis that it came from bats.

Clinical scale: From diagnosis to outcome predictions

To date, most clinical applications of AI to the COVID-19 response have focused on diagnosis based on medical imaging, with an increasing number of studies exploring non-invasive monitoring techniques. In recent literature, we have found several works that use AI to support diagnosis from computational tomography (CT) and X-Ray scans, in addition to others that use patient medical data to predict the evolution of the disease and original non-invasive measurements for monitoring purposes.

Medical imaging for diagnosis

Reverse Transcription Polymerase Chain Reaction (RT-PCR) tests are the key approach used for diagnosing COVID-19, however they have limitations in terms of resources, specimen collection, time required for the analysis, and performance [55]. As such, there is growing interest in other diagnostic methodologies which use medical imaging for the screening and diagnosis of COVID-19 cases [56]. This is notably due to the fact that COVID-19 exhibits particular radiological signatures and image patterns which can be observed in medical imagery [55, 57], but the identification of these patterns remains time-consuming even for expert radiologists. This makes image analysis from lung CT and X-Ray scans of COVID-19 patients a prime candidate for ML-based approaches, which could help accelerate the analysis of these scans, although the extent to which imaging can be used for diagnosis is still under discussion [58, 59].

Nonetheless, there are several approaches that aim to leverage Machine Learning for diagnosing COVID-19 from CT scans, via binary (i.e. healthy vs. COVID-19 positive) [60–62] † or multi-class (healthy patients vs. COVID-19 vs. other types of pneumonia) classification tasks using neural networks trained from scratch [63, 64]† [65]. These approaches use different architectures such as Inception [66], UNet++ [67] and ResNet [13], which can be trained directly either on raw CT scans, or scans labeled with regions of interest identified by radiologists. Some studies also adopt a hybrid approach, combining off-the-shelf software with bespoke ML approaches in order to achieve higher accuracy. For example, in Gozes *et al.* [68]† , a commercial medical imaging program is used for initial image processing and then combined with an ML pipeline. The two-step ML approach consists of a U-Net architecture [69] trained on medical images of lung abnormalities in order to pinpoint lung regions of interest and a Resnet-50 [13] trained on ImagetNet [70] and fine-tuned on COVID-19 cases in order to classify the images as COVID-positive or healthy. The resulting architectures are able to extract relevant features from the images and identify COVID-19 pneumonia even in cases where there are several competing potential diagnoses, and can be deployed both at hospitals to help radiologists accelerate the analysis of new cases and shared on the Internet to enable rapid review of new images.

X-Ray images, and specifically chest radiographs, also can be used for COVID-19 detection. Given the accessibility and potential portability of the imaging equipment needed, they can be an alternative in settings where access to advanced medical equipment such as CT scanners is limited. As shown in [71–73]† , there is potential in the use of Deep Learning approaches on X-Ray imagery, using architectures similar to the ones used for CT scans (e.g., ResNet [13] and Convolutional Neural Networks (CNNs) [28]). Further work aims to make predictions interpretable [74, 75]† and ensure that the models can be deployed in mobile and low-resource settings [76]† .

Studies which report operational deployment, such as [77]† , have opted for human-in-the-loop approaches to reduce the analysis time required while utilizing ML architectures. The authors use small manually-labeled batches of data for training an initial model based on the V-Net architecture [78]. This model then proposes segmentation of new CT scans, which can then be corrected by radiologists and fed back into the model, in an iterative process. This approach has enabled the development of a Deep Learning-based system for both automatic segmentation and the counting of infection regions, as well as assessing the severity of COVID-19, i.e. the percentage of infection in the whole lung. The authors show not only that the model improved its own performance incrementally, but also that the human time required for analysis of new images dropped from over 30 minutes initially to under 5 minutes after 200 annotated examples were used to train the model, reducing the effort required by radiologists to review a new scan. This is a promising direction which harnesses the power of ML alongside human annotation and expertise, which can be complementary and mutually beneficial.

While encouraging results have been achieved by many medical imagery-based AI diagnostics methods, in order for these methods to be used as clinical decision support systems, they should undergo clinical investigations and comply with regulatory and quality control requirements. In particular, their performance should be validated on a relevant and diverse set of training, validation, and test datasets, and they should demonstrate effectiveness in the clinical workflow [79]. We note that most of the papers we reviewed lacked provisions for these measures, relying on small and poorly-balanced datasets with flawed evaluation procedures and no plan for inclusion in clinical workflows.

Non-invasive measurements for disease tracking

There are also a number of original approaches that do not require specialized medical imaging equipment for diagnosing and tracking COVID-19. For example, one study used a GRU neural network [80] trained on footage from Kinect depth cameras to identify patient respiratory patterns [81]† , based on recent findings suggesting that COVID-19 has respiratory patterns which are distinct from those of the flu and common cold, notably that they exhibit tachypnea (rapid respiration) [82]. While these abnormal respiratory patterns are not necessarily correlated with a real-world diagnosis of COVID-

19, prediction of tachypnea could be a relevant first-order diagnostic feature that may contribute to large-scale screening of potential patients. Furthermore, new studies are being carried out which aim to understand how wearable device data can help COVID-19 surveillance, based on clinical research that demonstrates the value of aggregated signals from resting heart rates acquired from smart watches for influenza surveillance [83].

Finally, a growing number of efforts aim to utilize mobile phones for COVID-19 detection, using e.g. embedded sensors to identify COVID-19 symptoms [84]† , phone-based surveys to filter high-risk patients based on responses to key questions regarding travel and symptoms [85], or the analysis of cough sounds for preliminary COVID-19 diagnosis [86]† . While these are important efforts given the ubiquity and accessibility of mobile phone technology, these studies are not sufficiently advanced to evaluate their performance, so more extensive testing and clinical investigations are needed for deployment.

Patient outcome prediction

Forecasting potential patient outcomes is critical for preparation, planning, and optimization in overstretched health systems during the COVID-19 pandemic. It is important to know which factors can put patients at risk for hospitalization, developing acute respiratory distress syndrome (ARDS), and death from respiratory failure. In this vein, there have been several recent papers that propose triage approaches based on features contained in patients' medical data and blood tests, in order to help clinicians identify high-risk patients and those at risk of later development of ARDS [87,88]† , [89]. Using approaches such as the XGBoost algorithm [46] and Support Vector Machines [52], these approaches aim to identify key measurable features to predict mortality risk, which can later be tested for in hospitals upon patient admission and during the hospital stay. Clinical indicators that were identified using these ML-driven approaches include lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (CRP) [88]† ; alanine aminotransferase (ALT), myalgias, and hemoglobin [89]; and Interleukin-6, Systolic blood pressure and Monocyte ratio [87]† , although more research is needed to define specific thresholds and ranges of these indicators.

Furthermore, several complementary studies aim to leverage medical imagery for patient outcome prediction. These include carrying out severity assessment [90], predicting the need for long-term hospitalization based on CT imaging data [91]† , and patient risk stratification based on X-Ray images [92]† . Such approaches could help identify patients that might require intensive and long-term care, enabling hospitals to plan and manage their resources more effectively, as well as to monitor the state of patients and recognize when their condition worsens. A hybrid approach has also been proposed for this purpose, utilizing both CT findings as well as clinical features to predict the severity of COVID-19 [93]. The clinical features that were identified in this study, i.e. LDH and CRP, are similar to those identified in the purely clinical studies mentioned above; this overlap is promising for eventual clinical monitoring of these indicators. While these studies are limited both in scope and in data, they constitute important avenues of research that can be complemented and extended with clinical data from incoming cases around the world, thereby hopefully improving the prognosis of all patients and reducing the mortality of those that are critically ill.

Societal scale: Epidemiology and infodemiology

Epidemiology

The spread of the SARS-CoV-2 virus across the globe has received much policy attention, with advice at the national and local level changing daily in many locations as new information and model forecasts become available. Understanding how the virus is transmitted, and its likely effect on different demographics and geographic locations, is therefore crucial for public policy health care interventions.

The field of epidemiological research is incredibly vast, and the relevance and scale of the pandemic, in addition to the new data becoming available, has resulted in multiple modeling efforts. While most of these endeavors build on well-established classical models, such as susceptible-infected-recovered (SIR) models and fine-tune them to the COVID-19 situation, we focus here on cases specifically employing Machine Learning techniques for epidemiological modeling tasks.

Modeling and forecasting statistics

Most AI applications developed for epidemiological modeling have focused on forecasting national and local statistics such as: the total number of confirmed cases, mortality, and recovery rates.

Many authors have attempted to identify optimal approaches or model architectures for understanding and forecasting data, e.g. on a daily basis [94, 95]† [96]. These works employ modeling techniques such as an LSTM-GRU architecture [33, 80] for time series analysis and prediction [94]† , and CNN [28] based approaches in which numerical data has been combined and reshaped into "images" [95]† . In addition, new forecasting models for predicting the total number of confirmed cases have been developed [96]. In this study, the authors combine an adaptive neuro-fuzzy inference system (ANFIS) [97] with an enhanced flower pollination algorithm (FPA) [98] and salp swarm algorithm (SSA) [99] for optimizing the parameters of the model. In addition, they assess the robustness of their approach by training and testing on weekly confirmed influenza cases as collected by the US Centers for Disease Control and the WHO over two different four-year periods.

While these studies show how a range of different architectural choices can be made when building forecasting models, they demonstrate the complexities involved in choosing between such models and the non-trivial interplay between architectures, hyperparameters, and datasets. Moreover, since much of the data collected for COVID-19 modeling tasks is extremely limited, the choice of models and datasets can have significant effects on overall performance. In an attempt to address this, a simple framework, entitled "Group of Optimized and Multi-source Selection" (GROOMS), has been developed for exploring models and datasets during testing by ensuring that models of different categories, as defined by the authors, are tested in parallel for comparison [100]. Using the framework, the authors propose and test a polynomial neural network with corrective feedback (PNN+cf) [101] against other model architecture. This model is found to achieve optimal performance in predicting daily statistics on small datasets taken from Chinese health authorities.

Social media, other online data sources also provide a rich source of information for understanding public opinion, perception and behaviour. Such information can be incorporated into modeling efforts to augment existing data with the aim of providing more contextual understanding. For example, Liu *et al.* [102]† combine related internet search and news media activity with data from the Chinese Center for Disease Control and daily forecasts from GLEAM [103], an agent-based mechanistic model, in order to produce 2-day forecasts for a range of daily statistics. The authors first cluster provinces based on geo-spatial similarities in COVID-19 activity and then train a separate model on each cluster. The authors adapt an existing autoregressive model [104] [105]† for forecasting.

Similarly, data pertaining to Google search queries and news media volumes have been used as inputs to forecasting models for predicting daily trends [106]† . The authors assess the frequency of searches for different symptoms derived from a UK National Health Service survey of COVID-19 patients in which symptoms were recorded. Once the frequencies of each symptom search were calculated, they were weighted according to a probability distribution derived from these questionnaires. Using this data, along with prior daily statistics, the authors train an ElasticNet [107] model for forecasting future trends. Moreover, the authors investigate the transferability of their models between countries. This type of approach could be useful for probing the viability of training a model on data-rich countries and applying it to a data-poor ones, although the results of such a transferred model will have to be tailored for local contexts given that there may be different demographic characteristics and cultural norms.

Clustering

During the course of the outbreak, different countries experience different outbreak timing and growth depending on a range of factors including: international travel, demographics, socioeconomic factors, health care system characteristics and policy interventions. By assessing commonalities in virus propagation trends, as well as other country and regional data, it may be possible to cluster countries and regions and thereby use data from other similar areas to predict the outbreak in others. While useful at a high level, a significant limitation of many articles in this category is the heterogeneous data collection and reporting in different countries due to multiple factors including testing, case tracking, and reporting quality and standard.

A simple approach to clustering countries using an unsupervised k-means algorithm is given by Carollo-Larco *et al.* [108]. The authors cluster 155 countries using data relating to disease prevalence, average health status, air quality, gross domestic product (GDP), and universal health coverage, and find that their model was able to stratify countries according to the number of confirmed cases, although could not stratify them in terms of the number of deaths or the case fatality rate.

More sophisticated approaches have used the latent features of autoencoders, trained to predict infection rates, to identify similar groups of regions or countries. For example, Hu *et al.* [109]† have compiled a dataset of accumulated and new confirmed cases in 31 provinces and cities of China. After training a modified autoencoder (MAE) for real-time forecasting of new cases, the authors extract information from the autoencoder's latent variable layers to determine the model's most important features for each analyzed region. These features are then fed into a k-means clustering algorithm which groups similar regions for further analysis. This final step is designed to enable more efficient investigation of the regions showing infected/recovered characteristics of interest. Similarly, recent research has proposed training a Topological Autoencoder (TA), a simplified version of a Soft-supervised Topological Autoencoder [110], on the number of COVID-19 patients across 240 countries using data collected by the Center for Systems

Science and Engineering (CSSE) at Johns Hopkins University. The authors then study the latent variables of the TA to create a 2-dimensional clustering of countries [111].

Efficacy of public policy

In attempting to manage the pandemic, many national and local governments have introduced public policy interventions such as social distancing and the quarantining of individuals showing symptoms of COVID-19. The impacts of such measures can be modeled using agent-based models or by introducing regularizers in differential equations governing interaction models such as SIR. For instance, Hu *et al.* [112]† use data from WHO reports to train a modified autoencoder (MAE) to predict the number of cases and deaths on a daily basis. The authors encoded different intervention mechanisms according to their perceived strength and used this variable as an input to the model.

A different approach used data from Wuhan, China to build on the classical SIR model by adding a time-dependent regularizer to model the number of infected people who are in quarantine [113]† . Instead of specifying the form of this function and fitting parameters, the authors use a Neural Network to learn the "quarantine strength", $Q(t)$, based on the data, which in turn helps to determine the number of people who are able to infect others due to quarantine. While such work is heavily dependent on the available data and does not differentiate between symptomatic and asymptomatic individuals, the use of Neural Networks to augment well-understood techniques could serve as a powerful modeling tool.

Risk assessment

The models discussed in the previous sections mainly focused on predicting daily aggregate statistics for different regions or countries. Other work has specifically attempted to forecast risks of outbreaks in such regions, often by reducing aggregate statistical trends into a single risk score which facilitates interpretation, distills information for rapid analysis, and acts as a precursor to further investigation. However, it is important to note that such a distillation may not be robust to important changes in the underlying data or its coverage, and so should be interpreted with caution by policy makers.

Pal *et al.* [114]† train an LSTM [33] on variables including daily statistics and weather data to predict the long-term duration of the disease. In assessing which variables should be included in the model, the authors use an ordinary least-squares regression model to assess the p-value of all candidate features. The output of the LSTM is then used alongside explicit fuzzy rules (based on rates of death, confirmed cases, and recovery) to determine a risk category for the country or region.

A similar study looked at the Inherent Risk of Contagion (IRC), which is defined and calculated by the authors for similar geographic regions based on the acceleration of disease spread [115]† . The authors use k-means clustering to identify similar regions based on a non-linear combination of demographic and social characteristics and trained a Fully Connected Network (FCN) on data from Lombardy, Italy to forecast the IRC of the remaining provinces and municipalities of the country.

A more detailed approach was taken by Ye *et al.* [116]† , who develop a hierarchical community-level risk assessment. Given a location, the proposed $\alpha-$Satellite framework provides risk indices associated with different geographic levels (e.g. state, county, city). To achieve and test this framework the authors use data from the WHO and United

States Centers for Disease Control as well as county governments and other media for: new cases, death rates, and confirmed cases; demographic data; mobility data; and social media data which is to be used as an indication of public perception. For regions in which social media data is sparse, the authors use a cGAN [117]† trained on similar areas to generate social media content. The authors then attempt to incorporate how information at each of the different regional levels impacts the others, as well as how different attributes at each level influence the overall spread of the disease. After building a graph defining relationships between different geographic levels, the authors extract the latent variables from an autoencoder, which is designed to aggregate information propagated between different nodes on the graph. The autoencoder here plays the role of a dimensionality reduction algorithm to better understand the interplay between different geographic areas and their attributes.

Bayesian analysis

Although they are sometimes considered to be statistical rather than Machine Learning approaches, Bayesian analysis techniques can provide useful insights with respect to uncertainty and the handling of small datasets. In one study, Roy *et al.* [118]† develop a time-varying Bayesian autoregressive model for counts (TVBARC) with a linear link function for the estimation of time-dependent coefficients which could allow for better temporal modeling of the virus spread.

A more case-specific application of such methods is employed by [119], who seek to understand the rate of asymptomatic cases using data on 634 confirmed cases collected during the COVID-19 outbreak on-board the *Diamond Princess* cruise ship. The authors use a Bayesian time-series model with Hamiltonian Monte Carlo (HMC) algorithm and a No-U-Turn-Sampler [120] for model parameter estimation, to estimate the probability that a given patient is asymptomatic conditional on infection, along with the duration for which an individual is infected. The authors conclude that 17.9% of patients are asymptomatic. Although it is unclear if this result applies to the broader population, contained environments such as this one can be useful for tracking infection because they allow for comprehensive case data collection.

Infodemiology

The WHO defines an infodemic as "an over-abundance of information – some accurate and some not – that makes it hard for people to find trustworthy sources and reliable guidance when they need it" and deems it a second "disease" which needs fighting [121]. In this section, we highlight efforts to quantify the spread of information surrounding the pandemic and to understand its dynamics. Dealing with this vast amount of information requires the development and adoption of new tools, particularly surrounding the dissemination of misinformation and disinformation. While this is already an area in which much AI, and more specifically ML, research has been carried out, there is still a need for greater understanding of the underlying social dynamics during the pandemic.

Social media and online platforms have become key distribution channels for information surrounding the virus. Although national and international organizations have used these platforms to constructively communicate with the public, we are also seeing that populations can become overwhelmed with information, and that the propagation of misinformation and disinformation is increasingly prevalent. Furthermore, as highlighted in Figure 1, we have seen a significant increase in the number of published scientific articles related to the SARS-CoV-2 virus. Given that the virus is still relatively new and our understanding is quickly developing, many of these articles are disseminated via preprint archives, making it difficult to assess their quality. This does not mean that information contained within these articles cannot be valuable, but rather that effort should be made to distill this vast body of literature.

Spread and interaction

Understanding more about the dissemination of information is crucial to intervening proactively or reactively. While we acknowledge a wealth of literature on information propagation, network analysis, and social media interaction, in this section we discuss specifically those works applying such methods to the current infodemic.

At a high-level, work such as that by Singha *et al.* [122]† looks at global trends on Twitter by country. This work analyzes tweet volume according to specific themes discovered in coronavirus-related queries. The authors also analyze posts pertaining to specific myths surrounding the virus, examining the number of tweets containing certain terms they deem related to the myths, as well as the websites linked from them (categorized as either high-quality or low-quality sources).

In an effort to find early warning signals of a country or region experiencing an infodemic, Galotti *et al.* [123]† analyze social media posts on Twitter across 64 languages. The authors develop an Infodemic Risk Index (IRI), to quantify the rate at which a given generic user from a country or region is exposed to such unreliable posts from different classes of users, i.e. verified humans, unverified humans, verified bots, and unverified bots. The IRI considers the number of followers of the users which fall into each class, the number of messages those users post and their reliability (as measured by fact-checkers applied to samples of the user posts). This study highlights potentially actionable insights, observing that "the escalation of the epidemics leads people to progressively pay attention to more reliable sources thus potentially limiting the impact of the infodemics, but the actual speed of adjustment may make a major difference in determining the social outcome, and in particular between a [sic] controlled epidemics and a [sic] global pandemics".

In a broad-ranging study [124]† , interaction and engagement with COVID-19-related social media content is analyzed. From a collection of eight million comments and posts, selected using COVID-19 related keywords, from Twitter, Instagram, YouTube, Reddit, and Gab, the authors estimated engagement and interest in COVID-19 and comparatively assess the evolution of discourse on each platform. Interaction and engagement were measured using the cumulative number of posts and the number of reactions (e.g. comments, likes etc.) to these posts across the 45-day period. The authors then employed phenomenological [125] and classical SIR models to characterize the reproduction numbers. Specifically, they examined the average number of secondary cases (users that start posting about COVID-19) created by an "infectious" individual (already posting) on each of the social media platforms. As in epidemiological models, the authors simulated the likelihood of an infodemic in which discussion of COVID-19 will grow exponentially, at least in its initial stages. Moreover, the authors examined the spread of misinformation (which they identify using external fact-checking organizations). They find that information from both reliable and unreliable sources propagate in similar patterns, but that user engagement with posts from less-reliable sources is lower than engagement with content from reliable sources on major social media streams.

Similarly, Mejova *et al.* [126]† have examined the use of Facebook advertisements with content related to the virus. The authors used the Facebook Ad Library to search for all advertisements using the keywords “coronavirus” and “covid-19” and collected results across 34 countries, with most in the US (39%) and the EU (Italy made up 25% of the advertising market). While the majority of advertisements were paid for by non-profits to disseminate information and solicit donations, the authors found that around 5% of advertisements contained possible errors or misinformation.

Hate speech

Along with the propagation of misinformation and disinformation, the rise in hate speech in recent months has been of significant concern. As reported by the United Nations, there is an alarming rise of verbal abuses which might turn into physical violence against vulnerable and discriminated groups [127].

Velasquez *et al.* [128]† take a high-level approach to understanding the spread of hateful and malicious COVID-19 information and content within a variety of different social media channels, and attempt to characterize the methods by which such content moves between different channels. Concerningly, the authors find that hateful content is rapidly evolving and becoming increasingly coherent as time continues. As in [124]† , this study makes a comparison to the epidemiological R0 reproduction number in an attempt to determine the “tipping point” at which information will spread rapidly between information channels.

Others have looked at the emergence of Sinophobic behavior on social media, specifically Twitter and 4chan [129]† . This study uses data from October 2019 to March 2020 and uses word embeddings to assess context and word similarity over the entire five month period, as well as on a weekly basis. The authors also compared their findings to models trained on content taken from historical pre-COVID-19 content. The authors observed a distinct increase in Sinophobic content across the social media channels analyzed, and concluded that the Web is being “exploited for disseminating disturbing and harmful information, including conspiracy theories and hate speech targeting Chinese people”.

Understanding and fighting the spread of hate speech is of vital importance for the protection of human rights, in particular those of the most vulnerable and marginalized. By better comprehending the dynamics and the landscape of hateful speech, effective intervention mechanisms can be designed to disrupt and change the narrative.

Positive action

In the process of studying the features and dynamics of the infodemic, many of the works mentioned above suggest possible intervention options. In this section we explore several examples of such positive actions that are being considered and/or deployed to counter the infodemic.

The World Health Organization has taken steps to proactively confront the infodemic and bring together actors to assess aspects which still need to be addressed. Indeed, the WHO has been combating this infodemic through the use of its Information Network for Epidemics (EPI-WIN) platform for sharing information with key stakeholders [130], and is also working with social media and internet search companies to track the spread of specific rumors and to ensure that WHO content is displayed at the top of searches for terms related to the virus. Indeed, in April 2020, the WHO conducted a wide-ranging consultation on understanding and managing the infodemic, resources from which will be publicly available [121].

Efforts are also underway to curate specific news content related to the virus and perform both manual and automated fact-checking and relevance analysis. For instance, Pandey *et al.* [131]† have developed a pipeline for assessing the similarity between daily news headlines and WHO recommendations. The pipeline uses word embedding and similarity metrics, such as cosine similarity, to assess the relevance level of WHO recommendations to news article titles and content. If the similarity is above a certain threshold then the new article displays on the user's timeline with the associated relevant WHO recommendation. The setting of the similarity threshold is determined by human reviewers prior to release and then can be updated through user feedback. In the face of conflicting information, such methods could help identify accurate and trustworthy news articles which highlight important guidelines and promote official recommendations.

Another possible intervention strategy under consideration is the use of chatbots, which can be used to disseminate information while relieving pressure on other communication channels such as question-and-answer hotlines. For example, the WHO who has developed an interactive chatbot in multiple languages that allows users to explore pre-coded topics [132]. Finally, digital personal assistants could also be used to interactively disseminate official information, although governments and international actors would need ways to update recommendations as understanding changes.

Datasets and resources

The success of the global effort to use AI techniques to address the COVID-19 pandemic hinges upon sufficient access to data. Machine Learning, and Deep Learning in particular, requires notoriously large amounts of data and computing power in order to develop and train new algorithms and neural network architectures. In this section, we describe some of the datasets and data collection efforts that exist at the present time.

Case data

The current number and location of cases is essential for tracking the progress of the COVID-19 pandemic, calculating the growth rate of new infections, and observing the impact of preventive measures. Several datasets from organizations such as the WHO and national Centers for Disease Control (CDCs) exist for this purpose. They have been aggregated into public repositories hosted by institutions such as the Johns Hopkins CSSE or platforms like Github [133], in order provide day-level information on COVID-19 affected cases gathered from a variety of reliable sources. There are also other complementary data sources – including regional data on school closures, bank interest rates, and even community perceptions of the virus – which are continuously being added to a data portal hosted by the Humanitarian Data Exchange. A multitude of AI algorithms can be applied on this kind of data, including time series forecasting approaches such as LSTM networks [33] to predict the evolution of cases on a global and regional scale.

There is also an increasing quantity of tools and resources developed specifically for medical professionals and institutions, using data to help them prepare for managing the pandemic. For instance, CHIME is an open-source COVID-19 Hospital Impact Model for Epidemics based on SIR modeling, which uses the number of susceptible, infectious, and recovered individuals to compute the theoretical number of people infected over time and to predict outcomes in specific circumstances, and plan for the quantity of hospital beds that may be needed. While the CHIME project does not currently use ML techniques, it could benefit from these techniques in order to incorporate more features and data points such as hospital capacity information, to be used in applications such as dynamic ventilator allocation and surge capacity planning. Finally, there are also efforts underway to use de-identified, large-scale data to assess mobility changes and their impact on the local evolution of the epidemic, for instance in Italy and in North America.

Text data

Scientific Literature

As mentioned in previous sections, Machine Learning approaches can be used to analyze and parse the vast quantity of written information on COVID-19 and other coronaviruses, in order to make it easier for researchers and clinicians to use this information. Key questions of interest include:

1. What is known about the virus' transmission, incubation, and environmental stability?
2. What do we know about COVID-19 risk factors?
3. What do we know about non-pharmaceutical interventions?
4. What do we know about vaccines and therapeutics?
5. What has been published about ethical and social science considerations?
6. What has been published about medical care?

These questions can be studied using different sources, including the WHO Global Research Database on COVID-19, a curated literature hub for COVID-19 scientific information, and the COVID-19 Open Research Dataset (CORD-19), which is currently the largest open dataset available with over 52,000 relevant research articles. Several studies aiming to analyze this information have already been published, including [134] which uses a graph-based model to search through abstracts and to find relevant information, and [135]† , which extracts key terms and compares their usage with pre-COVID-19 articles. There are also several ongoing Kaggle challenges involving this data, with dozens of questions submitted daily and many teams involved. Other scientific research datasets that can be exploited include LitCOVID and the Dimensions AI COVID-19 dataset, which can contain important supplementary information such as clinical trial data when available. Using any of the sources mentioned above, NLP techniques can be applied to develop text mining tools and resources that can help the medical community find answers to key scientific questions regarding the nature and progress of COVID-19.

News and Social Media Data

Depending on the research questions addressed, data from scientific articles can also be completed with data from other sources, such as news articles and social media. Datasets such as COVID-19 TweetIDs, Covid19 Tweets, the Covid-19 Twitter dataset [136] †) which are maintained with general coronavirus-related tweets, can be useful for tracking the propagation of misinformation and unverified rumors on Twitter [137], as well as for monitoring the reactions of different populations to the virus. A potentially complementary source of information for this task is the COVID-19 Real World Worry Dataset [138] † , which includes labeled texts of individuals' emotional responses to COVID-19, and therefore can contain important data regarding public sentiment and the impacts of the pandemic on mental well-being in different regions of the world.

There is also important information available from official sources and news outlets, since the global media coverage of the pandemic is substantial and ongoing. For instance, the Institutional, News and Media Tweet Dataset [139] † brings together tweets based on a list of manually verified sources, and can be used to track official and institutional messaging around the pandemic in different countries. Repositories such as the COVID-19 Coronavirus News Article database and the COVID-19 Television Coverage Dataset can also be used to explore the question of how both print and television media outlets are covering the outbreak. These are rich sources of data for researchers interested in analyzing how media coverage evolves as the virus spreads globally, or tracking misleading reports and disinformation in the media.

Biomedical data

Clinical data

At this time, there are not many open-source datasets and models that can be used for diagnostic purposes. Some of the CT scan detection approaches described in the Diagnosis Section are available online and accessible to the public, for instance those of Wang et al. and Song et al. However, the data used to train the various models described is not systematically shared, although such sharing would be of great value to the ML research community. Several initiatives exist to crowdsource and open-source relevant data, for instance the Covid Chest X-Ray Dataset [140] for medical imagery and the COVID-19 Risk Calculator for symptoms, but these are challenging to assemble and maintain manually. Also, while data collection and ML model training can be carried out by computer scientists, data labeling, vetting, and annotation often need to be done by medical professionals such as radiologists or clinicians.

To address this lack of accessible data, there is an increasing number of initiatives and repositories that aim to share data and models; for instance, the Data4COVID Living Data Repository and the COVID-19 Dataset Clearinghouse have links to dozens of open-source data repositories from different geographical areas and levels of granularity. Also, initiatives such as United Against COVID-19 are particularly important, since they have become online platforms where data scientists and ML researchers can apply their skills to address requests for help from the research community, for instance by performing cleaning of clinical data, extracting actionable information regarding COVID-specific research questions, and collaborating to develop tools for deployment on the ground in hospitals. Such initiatives are promising given their potential to bridge the gap between those with medical and biological knowledge or experience, and those with computational and data skills.

Molecular data

In terms of genomic sequencing and drug discovery, there are several datasets available from pre-existing initiatives or which have been created from scratch for COVID-19 specifically. On the one hand, tracking the genome sequence of SARS-CoV-2 is crucial for designing and evaluating diagnostic tests, tracing the pandemic, and identifying the most promising intervention options. Notably, the GISAID Initiative, founded over a decade ago for the specific purpose of promoting the international sharing of influenza virus sequences and related clinical and epidemiological data, is tracking the genomic epidemiology of SARS-CoV-2. Other projects such as Nextstrain are looking at the genetic diversity of coronaviruses, in order to characterize the geographic spread of COVID-19 by inferring the lineage tree of hundreds of publicly shared genomes of SARS-CoV-2.

On the other hand, in terms of drug discovery, there are well-established initiatives such as the RCSB Protein Data Bank and the Global Health Drug Discovery Institute that have created centralized portals with data and resources for better understanding COVID-19 and for carrying out structure-guided drug discovery. In addition, CAS, a division of the American Chemical Society, has recently released the open-source Covid-19 antiviral candidate compounds dataset, containing information regarding antiviral compounds and molecules that have similar chemical structures to existing antivirals, to help the discovery of both new and repurposed treatments against the disease. Finally, another potentially interesting crowdsourced resource is the citizen science game Fold.it, which leverages collective intelligence against COVID-19 by proposing to design an antiviral protein.

Discussion

This research mapping exercise suggests that ML and AI can support the response against COVID-19 in a broad set of domains. In particular, we have highlighted emerging applications in drug discovery and development, diagnosis and clinical outcome prediction, epidemiology, and infodemiology. However, we note that very few of the reviewed systems have operational maturity at this stage. In order to operationalize this research, it is crucial to define a research road map and a funnel for AI applications to understand how this technology can immediately assist with the response, how it might help later on in the evolution of the current pandemic, and how it can be used to combat future pandemics. In the face of overstretched health care networks, we must strengthen our health systems to sustain services beyond the control and management of COVID-19 in order to truly protect the vulnerable, such as people living with noncommunicable diseases (NCDs). As members of a global community of researchers and data scientists, we identify three key calls for action.

First, we believe that scalable approaches to data and model sharing using open repositories will drastically accelerate the development of new models and unlock data for the public interest. Global repositories with anonymized clinical data, including medical imaging and patient histories, can be of particular interest in order to generate and transfer knowledge between clinical institutions. To facilitate the sharing of such data, clinical protocols and data sharing architectures will need to be designed and data governance frameworks will need to be put in place. It is important to reinforce that research with medical data is subject to strong regulatory requirements and privacy-protecting mechanisms. In particular, AI for clinical applications should demonstrate not only their performance on tests datasets , but also their effectiveness and safety when integrated into real clinical workflows. Overall, any AI application developed should undergo an assessment to ensure that it complies with ethical principles and, above all, respects human rights.

Second, the multidisciplinary nature of the research required to deploy AI systems in this context calls for the creation of extremely diverse, complementary teams and long-term partnerships. Beyond the examples shown in this review, other promising domains in which AI could be used to fight against COVID-19 include robotics (e.g., cleaning or disinfecting robots) and logistics (e.g., the allocation and distribution of personal protective equipment). Funding opportunities which encourage such collaborations and define key research directions may help accelerate the success of such partnerships.

Third, we believe that open science and international cooperation can play an important role in this pandemic that knows no borders. Proven solutions can be shared globally and adapted to other contexts and situations, prioritizing those that target local unmet needs. In particular, given that many international organizations, private sector companies and AI partnerships operate across international borders, they may be in the position to facilitate the knowledge dissemination and capacity building of national health systems. Regions with less capacity can benefit from global cooperation and concentrate their efforts on the most important local challenges. AI systems, methods, and models can act as a compact form of knowledge sharing which can be used and adapted to other contexts if they are designed to be widely deployable, requiring low energy and compute resources.

We acknowledge the difficulty of adding value through AI in the current situation. Nevertheless, we hope that this review is a first step towards helping the AI community understand where it can be of value, which are the promising domains for collaboration, and how research agendas can be best directed towards action against this or the next pandemic.

Acknowledgements

United Nations Global Pulse is supported by the Governments of Netherlands, Sweden and Germany and the William and Flora Hewlett Foundation. JB also is supported by the UK Science and Technology Facilities Council (STFC) grant number ST/P006744/1. AL is supported by grants from IVADO and Mila institutes. Thank you to our colleagues from the M.Tyers Laboratory (IRIC) for their advice.

References

1. WHO. Coronavirus disease (COVID-19) outbreak situation; 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

2. Raghu M, Schmidt E. A Survey of Deep Learning for Scientific Discovery. arXiv preprint arXiv:200311755. 2020;.

3. CDC. Coronavirus Disease 2019 (COVID-19) - Frequently Asked Questions; 2020. https://www.cdc.gov/coronavirus/2019-ncov/faq.html.

4. Corum J, Zimmer C. How Coronavirus Hijacks Your Cells. The New York Times. 2020;.

5. Corum J, Zimmer C. Bad News Wrapped in Protein: Inside the Coronavirus Genome. The New York Times. 2020;.

6. Dror R, Huang P. CS279 Computational Biology: Structure and Organization of Biomolecules and Cells; 2019. https://web.stanford.edu/class/cs279/.

7. Liu C, Zhou Q, Li Y, Garner LV, Watkins SP, Carter LJ, et al. Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases. ACS Central Science. 2020;.

8. Zhavoronkov A, Aladinskiy V, Zhebrak A, Zagribelnyy B, Terentiev V, Bezrukov DS, et al. Potential COVID-2019 3C-like Protease Inhibitors Designed Using Generative Deep Learning Approaches. chemRxiv preprint chemrxiv:11829102v2. 2020;.

9. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell. 2020;.

10. Zhavoronkov A. Artificial Intelligence for Drug Discovery, Biomarker Development, and Generation of Novel Chemistry. ACS Publications; 2018.

11. Senior A, Jumper J, Hassabis D, Kohli P. AlphaFold: Using AI for Scientific Discovery; 2020.

12. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved Protein Structure Prediction Using Potentials from Deep Learning. Nature. 2020;577(7792):706–710. doi:10.1038/s41586-019-1923-7.

13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

14. Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv preprint arXiv:151107122. 2016;.

15. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Protein Structure Prediction Using Multiple Deep Neural Networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). Proteins: Structure, Function, and Bioinformatics. 2019;87(12):1141–1148.

16. Jumper J, Tunyasuvunakool K, Kohli P, Hassabis D, AlphaFold Team. Computational predictions of protein structures associated with COVID-19; 2020.

17. Heo L, Feig M. Modeling of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Proteins by Machine Learning and Physics-Based Refinement. bioRxiv preprint bioRxiv:20200325008904v1. 2020;.

18. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved Protein Structure Prediction Using Predicted Interresidue Orientations. Proceedings of the National Academy of Sciences. 2020;.

19. Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y. Deep-Learning Contact-Map Guided Protein Structure Prediction in CASP13. Proteins: Structure, Function, and Bioinformatics. 2019;87(12):1149–1164.

20. Richardson P, Griffin I, Tucker C, Smith D, Oechsle O, Phelan A, et al. Baricitinib as Potential Treatment for 2019-nCoV Acute Respiratory Disease. The Lancet. 2020;395(10223):e30–e31.

21. Segler MH, Preuss M, Waller MP. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. Nature. 2018;555(7698):604–610.

22. Fauqueur J, Thillaisundaram A, Togia T. Constructing Large Scale Biomedical Knowledge Bases from Scratch with Rapid Annotation of Interpretable Patterns. In: Proceedings of the 18th BioNLP Workshop and Shared Task; 2019. p. 142–151.

23. Ge Y, Tian T, Huang S, Wan F, Li J, Li S, et al. A Data-Driven Drug Repositioning Framework Discovered a Potential Therapeutic Agent Targeting COVID-19. bioRxiv preprint bioRxiv:20200311986836. 2020;.

24. Hong L, Lin J, Tao J, Zeng J. BERE: An Accurate Distantly Supervised Biomedical Entity Relation Extraction Network. arXiv preprint arXiv:190606916. 2019;.

25. Hu F, Jiang J, Yin P. Prediction of Potential Commercially Inhibitors against SARS-CoV-2 by Multi-Task Deep Model. arXiv preprint arXiv:200300728. 2020;.

26. Zhang H, Saravanan KM, Yang Y, Hossain MT, Li J, Ren X, et al. Deep Learning Based Drug Screening for Novel Coronavirus 2019-nCov. ResearchGate preprint 2020020061v1. 2020;.

27. Nguyen D, Gao K, Chen J, Wang R, Wei G. Potentially highly potent drugs for 2019-nCoV. bioRxiv preprint bioRxiv:20200205936013v1. 2020;.

28. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436–444.

29. Gao K, Nguyen D, Chen J, Wang R, Wei G. Potentially Highly Potent Drugs for 2019-nCoV. bioRxiv preprint bioRxiv:20200205936013v1. 2020;.

30. Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting Commercially Available Antiviral Drugs That May Act on the Novel Coronavirus (2019-nCoV), Wuhan, China through a Drug-Target Interaction Deep Learning Model. bioRxiv preprint bioRxiv:20200131929547. 2020;.

31. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:181004805. 2018;.

32. Hofmarcher M, Mayr A, Rumetshofer E, Ruch P, Renz P, Schimunek J, et al. Large-Scale Ligand-Based Virtual Screening for SARS-CoV-2 Inhibitors Using Deep Neural Networks. arXiv:200400979 [cs, q-bio, stat]. 2020;.

33. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9(8):1735–1780.

34. Ton AT, Gentile F, Hsing M, Ban F, Cherkasov A. Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. Molecular Informatics. 2020;.

35. Batra R, Chan H, Kamath G, Ramprasad R, Cherukara MJ, Sankaranarayanan S. Screening of Therapeutic Agents for COVID-19 Using Machine Learning and Ensemble Docking Simulations. arXiv:200403766 [physics, q-bio]. 2020;.

36. Donner Y, Kazmierczak S, Fortney K. Drug Repurposing Using Deep Embeddings of Gene Expression Profiles. Molecular Pharmaceutics. 2018;15(10):4314–4325. doi:10.1021/acs.molpharmaceut.8b00284.

37. Avchaciov K, Burmistrova O, Fedichev P. AI for the Repurposing of Approved or Investigational Drugs against COVID-19; 2020.

38. Zhavoronkov A, Aladinskiy V, Zhebrak A, Zagribelnyy B, Terentiev V, Bezrukov DS, et al. Potential COVID-2019 3C-like Protease Inhibitors Designed Using Generative Deep Learning Approaches. Insilico Medicine Hong Kong Ltd A. 2020;307:E1.

39. Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial Autoencoders. arXiv preprint arXiv:151105644. 2015;.

40. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems; 2014. p. 2672–2680.

41. Tang B, He F, Liu D, Fang M, Wu Z, Xu D. AI-Aided Design of Novel Targeted Covalent Inhibitors against SARS-CoV-2. bioRxiv preprint bioRxiv:20200303972133. 2020;.

42. Qing X, Lee XY, De Raeymaecker J, Tame JR, Zhang KY, De Maeyer M, et al. Pharmacophore Modeling: Advances, Limitations, and Current Utility in Drug Discovery. Journal of Receptor, Ligand and Channel Research. 2014;7:81–92.

43. Bung N, Krishnan SR, Bulusu G, Roy A. De Novo Design of New Chemical Entities (NCEs) for SARS-CoV-2 Using Artificial Intelligence. 2020;doi:10.26434/chemrxiv.11998347.v2.

44. Chenthamarakshan V, Das P, Padhi I, Strobelt H, Lim KW, Hoover B, et al. Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models. arXiv preprint arXiv:200401215. 2020;.

45. Magar R, Yadav P, Farimani AB. Potential Neutralizing Antibodies Discovered for Novel Corona Virus Using Machine Learning. arXiv:200308447 [cs, q-bio]. 2020;.

46. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 785–794.

47. Fast E, Chen B. Potential T-Cell and B-Cell Epitopes of 2019-nCoV. bioRxiv preprint bioRxiv:20200219955484v2. 2020; p. 2020.02.19.955484. doi:10.1101/2020.02.19.955484.

48. Ong E, Wong MU, Huffman A, He Y. COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning. BioRxiv preprint BioRxiv:20200320000141v2. 2020;.

49. Metsky HC, Freije CA, Kosoko-Thoroddsen TSF, Sabeti PC, Myhrvold C. CRISPR-based surveillance for COVID-19 using genomically-comprehensive machine learning design. bioRxiv preprint bioRxiv:20200226967026. 2020;doi:10.1101/2020.02.26.967026.

50. Lopez-Rincon A, Tonda A, Mendoza-Maldonado L, Claassen E, Garssen J, Kraneveld AD. Accurate Identification of Sars-Cov-2 from Viral Genome Sequences Using Deep Learning. bioRxiv preprint bioRxiv:20200313990242v1. 2020;.

51. Gussow AB, Auslander N, Wolf YI, Koonin EV. Genomic Determinants of Pathogenicity in SARS-CoV-2 and Other Human Coronaviruses. bioRxiv preprint bioRxiv:20200405026450v2. 2020;.

52. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–297.

53. Bartoszewicz JM, Seidel A, Renard BY. Interpretable Detection of Novel Human Viruses from Genome Sequencing Data. bioRxiv preprint bioRxiv:20200129925354v2. 2020;.

54. Randhawa GS, Soltysiak MP, El Roz H, de Souza CP, Hill KA, Kari L. Machine Learning Using Intrinsic Genomic Signatures for Rapid Classification of Novel Pathogens: COVID-19 Case Study. bioRxiv preprint bioRxiv:20200203932350v3. 2020;.

55. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology. 2020; p. 200642.

56. Kanne JP, Little BP, Chung JH, Elicker BM, Ketai LH. Essentials for radiologists on COVID-19: an update—Radiology Scientific Expert Panel. Radiology. 2020; p. 200527.

57. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology. 2020; p. 200432.

58. Ng MY, Lee EY, Yang J, Yang F, Li X, Wang H, et al. Imaging profile of the COVID-19 infection: radiologic findings and literature review. Radiology: Cardiothoracic Imaging. 2020;2(1):e200034.

59. Weinstock MB RJea Echenique A. Chest x-ray findings in 636 ambulatory patients with COVID-19 presenting to an urgent care center: a normal chest x-ray is no guarantee. The Journal of Urgent Care Medicine. 2020; p. 13–18.

60. Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). medRxiv preprint medRxiv:2020021420023028. 2020;.

61. Chen J, Wu L, Zhang J, Zhang L, Gong D, Zhao Y, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv preprint medRxiv:2020022520021568. 2020;.

62. Gozes O, Frid-Adar M, Sagie N, Zhang H, Ji W, Greenspan H. Coronavirus Detection and Analysis on Chest CT with Deep Learning. arXiv preprint arXiv:200402640. 2020;.

63. Xu X, Jiang X, Ma C, Du P, Li X, Lv S, et al. Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia. arXiv preprint arXiv:200209334. 2020;.

64. Song Y, Zheng S, Li L, Zhang X, Zhang X, Huang Z, et al. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. medRxiv preprint medRxiv:2020022320026930. 2020;.

65. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology. 2020;.

66. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1–9.

67. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer; 2018. p. 3–11.

68. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, et al. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. arXiv preprint arXiv:200305037. 2020;.

69. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015. p. 234–241.

70. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009. p. 248–255.

71. Abbas A, Abdelsamea MM, Gaber MM. Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. arXiv preprint arXiv:200313815. 2020;.

72. Bukhari SUK, Bukhari SSK, Syed A, SHAH SSH. The diagnostic evaluation of Convolutional Neural Network (CNN) for the assessment of chest X-ray of patients infected with COVID-19. medRxiv. 2020;.

73. Hammoudi K, Benhabiles H, Melkemi M, Dornaika F, Arganda-Carreras I, Collard D, et al. Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19. arXiv preprint arXiv:200403399. 2020;.

74. Karim M, Döhmen T, Rebholz-Schuhmann D, Decker S, Cochez M, Beyan O, et al. DeepCOVIDExplainer: Explainable COVID-19 Predictions Based on Chest X-ray Images. arXiv preprint arXiv:200404582. 2020;.

75. Ghoshal B, Tucker A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. arXiv preprint arXiv:200310769. 2020;.

76. Li X, Li C, Zhu D. COVID-MobileXpert: On-Device COVID-19 Screening using Snapshots of Chest X-Ray. 2020;.

77. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, et al. Lung Infection Quantification of COVID-19 in CT Images with Deep Learning. arXiv preprint arXiv:200304655. 2020;.

78. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE; 2016. p. 565–571.

79. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. bmj. 2020;368.

80. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078. 2014;.

81. Wang Y, Hu M, Li Q, Zhang XP, Zhai G, Yao N. Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner. arXiv preprint arXiv:200205534. 2020;.

82. Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R. Features, Evaluation and Treatment Coronavirus (COVID-19). In: StatPearls [Internet]. StatPearls Publishing; 2020.

83. Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. The Lancet Digital Health. 2020;.

84. Maghdid HS, Ghafoor KZ, Sadiq AS, Curran K, Rabie K. A Novel AI-enabled Framework to Diagnose Coronavirus COVID-19 using Smartphone Embedded Sensors: Design Study; 2020.

85. Rao ASS, Vazquez JA. Identification of COVID-19 Can be Quicker through Artificial Intelligence framework using a Mobile Phone-Based Survey in the Populations when Cities/Towns Are Under Quarantine. Infection Control & Hospital Epidemiology. 2020; p. 1–18.

86. Imran A, Posokhova I, Qureshi HN, Masood U, Riaz S, Ali K, et al. AI4COVID-19: AI Enabled Preliminary Diagnosis for COVID-19 from Cough Samples via an App. arXiv preprint arXiv:200401275. 2020;.

87. Feng C, Huang Z, Wang L, Chen X, Zhai Y, Zhu F, et al. A Novel Triage Tool of Artificial Intelligence Assisted Diagnosis Aid System for Suspected COVID-19 pneumonia In Fever Clinics. 2020;.

88. Yan L, Zhang HT, Xiao Y, Wang M, Sun C, Liang J, et al. Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. medRxiv preprint medRxiv:L2020022720028027. 2020;.

89. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, et al. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. CMC-Computers, Materials & Continua. 2020;63:537–551.

90. Tang Z, Zhao W, Xie X, Zhong Z, Shi F, Liu J, et al. Severity Assessment of Coronavirus Disease 2019 (COVID-19) Using Quantitative Features from Chest CT Images. arXiv preprint arXiv:200311988. 2020;.

91. Qi X, Jiang Z, Yu Q, Shao C, Zhang H, Yue H, et al. Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study. medRxiv. 2020;.

92. Wang L, Wong A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. arXiv preprint arXiv:200309871. 2020;.

93. Shi W, Peng X, Liu T, Cheng Z, Lu H, Yang S, et al. Deep Learning-Based Quantitative Computed Tomography Model in Predicting the Severity of COVID-19: A Retrospective Study in 196 Patients. 2020;.

94. Bandyopadhyay SK, Dutta S. Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release. medRxiv preprint medRxiv:2020032520043505. 2020;.

95. Huang CJ, Chen YH, Ma Y, Kuo PH. Multiple-Input Deep Convolutional Neural Network Model for COVID-19 Forecasting in China. medRxiv preprint medRxiv:2020032320041608. 2020;.

96. Al-qaness MAA, Ewees AA, Fan H, Abd El Aziz MAE. Optimization Method for Forecasting Confirmed Cases of COVID-19 in China. Journal of Clinical Medicine. 2020;9(3):674.

97. Jang JSR. ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics. 1993;23(3):665–685.

98. Yang XS. Flower pollination algorithm for global optimization. International Conference on Unconventional Computing and Natural Computation. 2012; p. 240–249.

99. Mirjalili SM, Gandomi AH, Mirjalili SZ, Saremi S, Faris H. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Adv Eng Softw. 2017;114:163–191.

100. Fong SJ, Li G, Dey N, Crespo RG, 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. 2020;6(1):132–140. doi:10.9781/ijimai.2020.02.002.

101. Ivakhnenko AG. Heuritic Self-Organization in Problems of Engineering Cybernetics. Automatica. 1970;6:207–219.

102. Liu D, Clemente L, Poirier C, Ding X, Chinazzi M, David JT, et al. A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. arXiv preprint arXiv:200404019. 2020;.

103. Balcan D, Gonçalves B, Hu H, Ramasco J, Colizza V, Vespignani A. Modeling the spatial spread of infectious diseases: the GLobal Epidemic and Mobility computational model. Journal of Computational Science. 2010;1(3):132–145.

104. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proceedings of the National Academy of Sciences. 2015;112(47):14473–14478. doi:10.1073/pnas.1515373112.

105. Lu FS, Hattab MW, Clemente L, Santillana M. Improved state-level influenza activity nowcasting in the United States leveraging Internet-based data sources and network approaches via ARGONet. bioRxiv preprint bioRxiv:20180614344580. 2018;.

106. Lampos V, Moura S, Yom-Tov E, Edelstein M, Majumder M, Hamada Y, et al. Tracking COVID-19 using online search. arXiv preprint arXiv:200308086. 2020;.

107. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(2):301–320. doi:10.1111/j.1467-9868.2005.00503.x.

108. Carrillo-Larco R, Castillo-Cara M. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach [version 1; peer review: awaiting peer review]. Wellcome Open Research. 2020;5(56). doi:10.12688/wellcomeopenres.15819.1.

109. Hu Z, Ge Q, Li S, Jin L, Xiong M. Artificial Intelligence Forecasting of COVID-19 in China. arXiv preprint arXiv:200207112. 2020;.

110. Hartono P. Mixing autoencoder with classifier: conceptual data visualization. arXiv preprint arXiv:191201137. 2019;.

111. Hartono P. Generating Similarity Map for COVID-19 Transmission Dynamics with Topological Autoencoder. arXiv preprint arXiv:200401481. 2020;.

112. Hu Z, Ge Q, Li S, Boerwincle E, Xiong M. Forecasting and evaluating intervention of Covid-19 in the World. arXiv preprint arXiv:200309800. 2020;.

113. Dandekar R, Barbastathis G. Neural Network aided quarantine control model estimation of COVID spread in Wuhan, China. arXiv preprint arXiv:200309403. 2020;.

114. Pal R, Sekh AA, Kar S, Prasad DK. Neural network based country wise risk prediction of COVID-19. arXiv preprint arXiv:200400959. 2020;.

115. Ronsivalle GB, Foresti L, Poledda G. A prototype model of georeferencing the Inherent Risk of Contagion from COVID-19. 2020;doi:10.13140/RG.2.2.30077.72163.

116. Ye Y, Hou S, Fan Y, Qian Y, Zhang Y, Sun S, et al. α-Satellite: An AI-driven System and Benchmark Datasets for Hierarchical Community-level Risk Assessment to Help Combat COVID-19. arXiv preprint arXiv:200312232. 2020;.

117. Mirza M, Osindero S. Conditional Generative Adversarial Nets. arXiv preprint arXiv:14111784. 2014;.

118. Roy A, Karmakar S. Bayesian semiparametric time varying model for count data to study the spread of the COVID-19 cases. arXiv preprint arXiv:200402281. 2020;.

119. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance. 2020;25(10).

120. Homan MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. J Mach Learn Res. 2014;15(1):1593–1623.

121. WHO. Infodemic management - Infodemiology; 2020. https://www.who.int/teams/risk-communication/infodemic-management.

122. Singha L, Bansala S, Bodea L, Budakb C, Chic G, Kawintiranona K, et al. A first look at COVID-19 information and misinformation sharing on Twitter. arXiv preprint arXiv:200313907. 2020;.

123. Gallotti R, Valle F, Castaldo N, Sacco P, Domenico MD. Assessing the risks of "infodemics" in response to COVID-19 epidemics. arXiv preprint arXiv:200403997. 2020;.

124. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 Social Media Infodemic. arXiv preprint arXiv:200305004. 2020;.

125. Fisman DN, Hauck TS, Tuite AR, Greer AL. An IDEA for short term outbreak projection: nearcasting using the basic reproduction number. PloS one. 2013;8(12).

126. Mejova Y, Kalimeri K. Advertisers Jump on Coronavirus Bandwagon: Politics, News, and Business. arXiv preprint arXiv:200300923. 2020;.

127. News U. COVID-19 stoking xenophobia, hate and exclusion, minority rights expert warns; 2020. https://news.un.org/en/story/2020/03/1060602.

128. Velásquez N, Leahy R, Restrepo NJ, Lupu Y, Sear R, Gabriel N, et al. Hate multiverse spreads malicious COVID-19 content online beyond individual platform control. arXiv preprint arXiv:200400673. 2020;.

129. Schild L, Ling C, Blackburn J, Stringhini G, Zhang Y, Zannettou S. "Go eat a bat, Chang!": An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. arXiv preprint arXiv:200404046. 2020;.

130. Zarocostas J. How to fight an infodemic. The Lancet. 2020;395(10225):676.

131. Pandey R, Gautam V, Bhagat K, Sethi T. A Machine Learning Application for Raising WASH Awareness in the Times of COVID-19 Pandemic. arXiv preprint arXiv:200307074. 2020;.

132. WHO. WHO and Rakuten Viber fight COVID-19 misinformation with interactive chatbot; 2020. https://www.who.int/news-room/feature-stories/detail/who-and-rakuten-viber-fight-covid-19-misinformation-with-interactive-chatbot.

133. Xu B, Gutierrez B, Mekaru S, Sewalk K, Goodwin L, Loskill A, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. Scientific Data. 2020;7(1):1–6.

134. Ahamed S, Samad MD. Information Mining For COVID-19 Research From A Large Volume Of Scientific Literature. arXiv preprint arXiv:200402085. 2020;.

135. Jr IF, Fister K, Fister I. Discovering Associations In COVID-19 Related Research Papers. arXiv preprint arXiv:200400673. 2020;.

136. Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, et al. A large-scale COVID-19 Twitter chatter dataset for open scientific research–an international collaboration. arXiv preprint arXiv:200403688. 2020;.

137. Chen E, Lerman K, Ferrara E. COVID-19: The First Public Coronavirus Twitter Dataset. arXiv preprint arXiv:200307372. 2020;.

138. Kleinberg B, van der Vegt I, Mozes M. Measuring Emotions in the COVID-19 Real World Worry Dataset. arXiv preprint arXiv:200404225. 2020;.

139. Yu J. Open access institutional and news media tweet dataset for COVID-19 social science research. arXiv preprint arXiv:200401791. 2020;.

140. Cohen JP, Morrison P, Dao L. COVID-19 image data collection. arXiv preprint arXiv:200311597. 2020;.