IEEE
Computational Intelligence
MAGAZINE

NOVEMBER 2020
VOLUME 15 NUMBER 4
WWW.IEEE-CIS.ORG

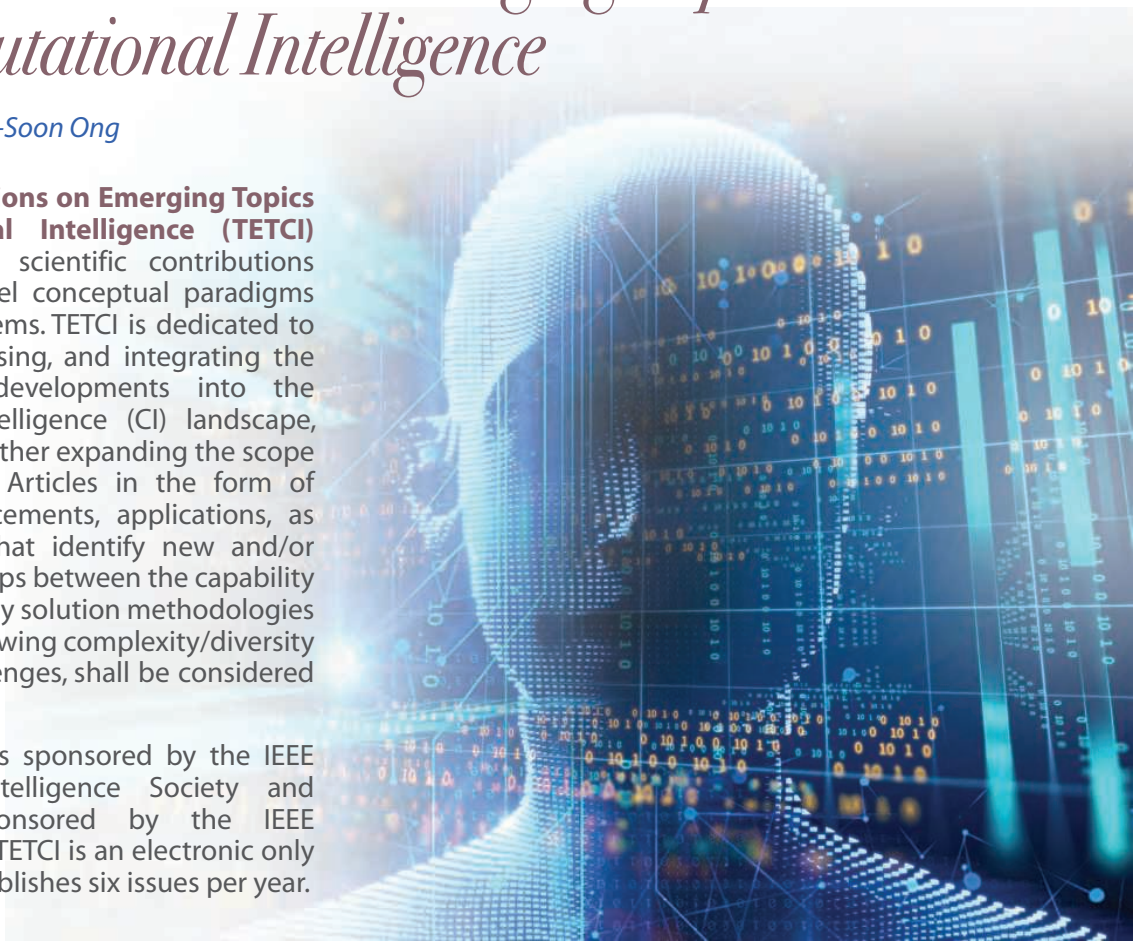
Computational Intelligence for Combating COVID-19

IEEE Computational Intelligence Society

IEEE

IEEE Transactions on Emerging Topics in Computational Intelligence

Editor-in-Chief: Yew-Soon Ong



The IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI) publishes original scientific contributions that establish novel conceptual paradigms for intelligent systems. TETCI is dedicated to embracing, harnessing, and integrating the latest scientific developments into the computational intelligence (CI) landscape, with the goal of further expanding the scope and depth of CI. Articles in the form of theoretical advancements, applications, as well as surveys, that identify new and/or address existing gaps between the capability limits of present-day solution methodologies and the rapidly growing complexity/diversity of real-world challenges, shall be considered for publication.

This Transactions is sponsored by the IEEE Computational Intelligence Society and technically co-sponsored by the IEEE Computer Society. TETCI is an electronic only publication and publishes six issues per year.

TETCI welcomes manuscripts in any emerging topic in computational intelligence, including but not limited to the following concepts and methodologies:

- Artificial Life
- Ambient Intelligence
- Brain Computer Interfaces
- Computational Neuroscience
- Cultural Learning
- Social Reasoning
- Non-Fuzzy Computing with Words
- Artificial Hormone/Endocrine/Glial Cell Networks
- Computational Intelligence for IoT
- Computational Intelligence for Smart-X Technologies

For submission and other information, please visit TETCI website at:
https://cis.ieee.org/publications/t-emerging-topics-in-ci

IEEE

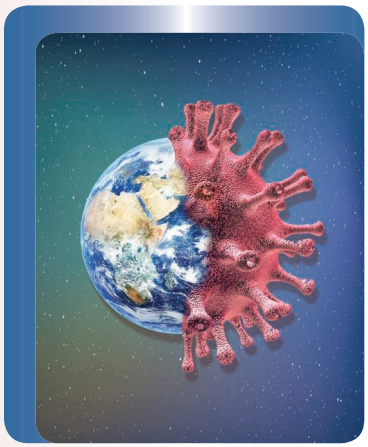
IEEE Computational Intelligence Society

IEEE COMPUTER SOCIETY

Digital Object Identifier 10.1109/MCI.2020.3019901

IEEE Computational Intelligence MAGAZINE

Volume 15 Number 4 ❑ November 2020
www.ieee-cis.org



on the cover

PLANET EARTH—©SHUTTERSTOCK.COM/MAD CHECKPOINT,
UNIVERSE BACKGROUND—©SHUTTERSTOCK.COM/KPP

Features

Departments

IEEE Computational Intelligence Magazine (ISSN 1556-603X) is published quarterly by The Institute of Electrical and Electronics Engineers, Inc. **Headquarters:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997, U.S.A. +1 212 419 7900. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society, or its members. The magazine is a membership benefit of the IEEE Computational Intelligence Society, and subscriptions are included in Society fee. Replacement copies for members are available for US$20 (one copy only). Nonmembers can purchase individual copies for US$213.00. Nonmember subscription prices are available on request. Periodicals postage paid at New York, NY and at additional mailing offices. Postmaster: Send address changes to IEEE Computational Intelligence Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-1331 U.S.A. PRINTED IN U.S.A. Canadian GST #125634188.

Digital Object Identifier 10.1109/MCI.2020.3019867

Digital Object Identifier 10.1109/MCI.2020.3019868

Chuan-Kang Ting
National Tsing Hua University, TAIWAN

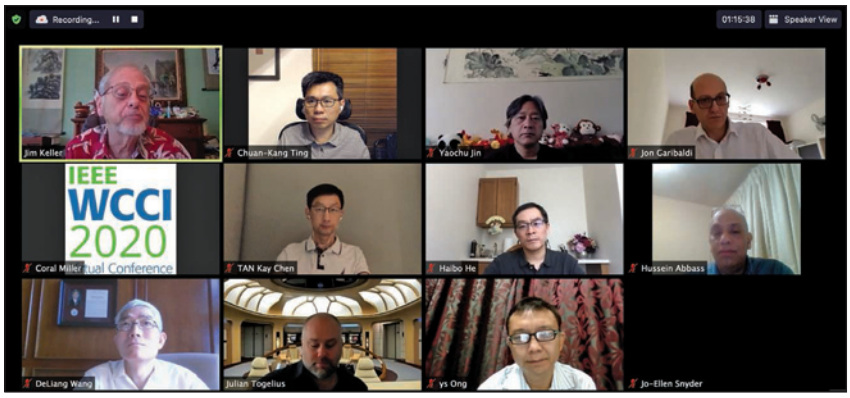
Editor's Remarks

Awaiting the Light at the End of the Tunnel

Undoubtedly, the COVID-19 pandemic has become the gravest worldwide issue in 2020. The world has faced common threats before; nevertheless, the fact that we are bound together this time by an invisible threat unsettles us all even more. We are all shocked by how fast the virus has spread globally, giving us almost no time to prepare and upending our lives in every way. Although things look bleak while this pandemic lasts, we can still find solace in the presence of our family and friends, and in humankind's ability to find solutions using our intelligence.

Like many others, I have also been caught in inevitable changes due to the pandemic. Conferences and meetings, such as the IEEE WCCI 2020, the CIS AdCom meeting, and the CIM AE meeting, have all moved online. Virtual conferences are ways to compromise in these dangerous times, and while I would have much preferred face-to-face meetings with their joys of conversing in person, I am surprised to find these online meetings pleasantly refreshing and appealing. These meetings give me a special sense of connectivity—seeing friends from so many different time zones all gathered together on the computer screen, drinking coffee, beer, or tea due to differences in countries or time. Hopefully we will all see each other in person soon.

Such is the gravity of our current situation that we should all shoulder our responsibility and take part in preventing further spread of the pandemic. With the support of IEEE CIS, especially James Keller, Pablo Estévez, and Bernadette Bouchon-Meunier, the CIM set up this fast-track special issue to provide solutions and to contribute its part in global prevention. This issue includes five articles that use computational intelligence technologies to help fight COVID-19: an in-depth review of the current studies on computational intelligence for combatting COVID-19, a Bayesian updating



Screenshot of Panel on "Landscape of Publications in Computational Intelligence" at IEEE WCCI 2020.

Digital Object Identifier 10.1109/MCI.2020.3019865
Date of current version: 14 October 2020

This issue includes five articles that use computational intelligence technologies to help fight COVID-19.

approach to estimate COVID-19 infection dynamics, a time series model for predicting the number of confirmed cases of COVID-19, a platform facilitating demonstration and analysis of COVID-19 data, and an intelligent optimization method for diversified community prevention of COVID-19.

We sincerely hope that the articles in this issue will be of assistance to the world in slowing down the spread of this pandemic. With the responsibility that each of us understands and is willing to take on, we may be nearing the light at the end of the tunnel yet!

Digital Object Identifier 10.1109/MCI.2020.3019869

Bernadette Bouchon-Meunier
CNRS – Sorbonne Université, FRANCE

President's Message

Living in a Virtual World

In this year 2020, we have learned to live in a virtual world. Besides the disastrous effects of the pandemic on our health and privacy, can we say that this virtual world is a good or a bad thing? The reality is certainly more mixed than it appears at first glance.

None of the conferences sponsored or co-sponsored by the CIS will ultimately be held physically in 2020.

In particular, the flagship conference IEEE WCCI 2020, planned to be held in Glasgow, Scotland, had to move to a virtual form. It was a challenge to handle more than 1800 papers to be presented online and to organize the virtual conference, taking into account all the technical constraints and time zone issues. I express my thanks to all those who contributed to its success. In the end, more than 2300 participants were able to attend, thanks to the outstanding quality of the program as well as the reduced registration fees and the various grants provided by the CIS to its members. Most of them, students and young professionals presenting a paper, or members not presenting a paper, could attend the conference free of charge. All the lectures and paper presentations were available on demand and easier to follow in the case of parallel sessions than in a physical conference.

In this sense, a virtual conference can be regarded positively, as it allows a larger number of participants to attend by avoiding travel expenses, paying little or no registration fees and being able to cope with a busy personal schedule. In particular, participants from developing countries or students benefit from the virtual form of conferences, and it seems that more women also attend than in the case of physical conferences.

The negative aspect is the lack of physical networking, friendly discussions, ease of interaction with plenary or keynote lectures during coffee breaks, even if it is possible to ask live questions to the authors after their presentations. Another negative aspect is clearly the difficulty of having a schedule that is compatible with all normal working hours in the world. It is of course possible to only watch the videos, but the concept of conference is then lost and replaced by a solitary visualization of a series of slide presentations.

In any case, there is nothing else possible in 2020 and I can only recommend that you attend all the CIS-sponsored virtual conferences that will be held until the end of the year. I am sure that you will appreciate their scientific programs and you will be happy to virtually meet the authors, given the effort made by the organizers to create conditions as close as possible to normal conferences. In particular, you will not miss the second flagship conference of this year, IEEE SSCI 2020, held virtually in Canberra, Australia on December 1–4, 2020, and co-located with an IEEE CIS Summer School on Artificial Intelligence.

Beyond conferences, the pandemic has changed the way we meet CIS members and we have had more virtual meetings this year in technical and administrative committees or chapters, than we had physical meetings in the past. It will certainly be interesting to preserve virtual meetings in the future, in complement to physical meetings we are used to and we will be happy to organize again.

Stay safe and be optimistic. We will meet again in person in the future!

Digital Object Identifier 10.1109/MCI.2020.3019870
Date of current version: 14 October 2020

Haibo He, Jon Garibaldi, Kay Chen Tan, Julian Togelius, Yaochu Jin, and Yew Soon Ong

CIS Publication Spotlight

IEEE Transactions on Neural Networks and Learning Systems

Change Detection in Graph Streams by Learning Graph Embeddings on Constant-Curvature Manifolds, by D. Grattarola, D. Zambon, L. Livi, and C. Alippi, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 31, No. 6, June 2020, pp. 1856–1869.

Digital Object Identifier: 10.1109/TNNLS.2019.2927301

"The space of graphs is often characterized by a nontrivial geometry, which complicates learning and inference in practical applications. A common approach is to use embedding techniques to represent graphs as points in a conventional Euclidean space, but non-Euclidean spaces have often been shown to be better suited for embedding graphs. Among these, constant-curvature Riemannian manifolds (CCMs) offer embedding spaces suitable for studying the statistical properties of a graph distribution, as they provide ways to easily compute metric geodesic distances. In this paper, we focus on the problem of detecting changes in stationarity in a stream of attributed graphs. To this end, we introduce a novel change detection framework based on neural networks and CCMs, which takes into account the non-Euclidean nature of graphs. Our contribution in this paper is twofold. First, via a novel approach based on adversarial learning, we compute graph embeddings by training an autoencoder to represent graphs on CCMs. Second, we introduce two novel change detection tests operating on CCMs. We perform experiments on synthetic data, as well as two real-world application scenarios: the detection of epileptic seizures using functional connectivity brain networks and the detection of hostility between two subjects, using human skeletal graphs. Results show that the proposed methods are able to detect even small changes in a graph-generating process, consistently outperforming approaches based on Euclidean embeddings."

©IMAGE CLUB

Completely Automated CNN Architecture Design Based on Blocks, by Y. Sun, B. Xue, M. Zhang, and G. G. Yen, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 31, No. 4, April 2020, pp. 1242–1254.

Digital Object Identifier: 10.1109/TNNLS.2019.2919608

"The performance of convolutional neural networks (CNNs) highly relies on their architectures. In order to design a CNN with promising performance, extensive expertise in both CNNs and the investigated problem domain is required, which is not necessarily available to every interested user. To address this problem, we propose to automatically evolve CNN architectures by using a genetic algorithm (GA) based on ResNet and DenseNet blocks. The proposed algorithm is completely automatic in designing CNN architectures. In particular, neither preprocessing before it starts nor postprocessing in terms of CNNs is needed. Furthermore, the proposed algorithm does not require users with domain knowledge on CNNs, the investigated problem, or even GAs. The proposed algorithm is evaluated on the CIFAR10 and CIFAR100 benchmark data sets against 18 state-of-the-art peer competitors. Experimental results show that the proposed algorithm outperforms the state-of-the-art CNNs hand-crafted and the CNNs designed by automatic peer competitors in terms of the classification performance and achieves a competitive classification accuracy against semiautomatic peer competitors. In addition, the proposed algorithm consumes much less computational resource than most peer competitors in finding the best CNN architectures."

Digital Object Identifier 10.1109/MCI.2020.3019871
Date of current version: 14 October 2020

IEEE Transactions on Fuzzy Systems

Fast and Scalable Approaches to Accelerate the Fuzzy k-Nearest Neighbors Classifier for Big Data, by J. Maillo, S. García, J. Luengo, F. Herrera, and I. Triguero, *IEEE Transactions on Fuzzy Systems*, Vol. 28, No. 5, May 2020, pp. 874–886.

Digital Object Identifier: 10.1109/TFUZZ.2019.2936356

"One of the best-known and most effective methods in supervised classification is the k-nearest neighbors algorithm (kNN). Several approaches have been proposed to improve its accuracy, where fuzzy approaches prove to be among the most successful, highlighting the classical fuzzy k-nearest neighbors (FkNN). However, these traditional algorithms fail to tackle the large amounts of data that are available today. There are multiple alternatives to enable kNN classification in big datasets, spotlighting the approximate version of kNN called hybrid spill tree. Nevertheless, the existing proposals of FkNN for big data problems are not fully scalable, because a high computational load is required to obtain the same behavior as the original FkNN algorithm. This article proposes global approximate hybrid spill tree FkNN and local hybrid spill tree FkNN, two approximate approaches that speed up runtime without losing quality in the classification process. The experimentation compares various FkNN approaches for big data with datasets of up to 11 million instances. The results show an improvement in runtime and accuracy over literature algorithms."

Multitasking Genetic Algorithm (MTGA) for Fuzzy System Optimization, by D. Wu and X. Tan, *IEEE Transactions on Fuzzy Systems*, Vol. 28, No. 6, June 2020, pp. 1050–1061.

Digital Object Identifier: 10.1109/TFUZZ.2020.2968863

"Multitask learning uses auxiliary data or knowledge from relevant tasks to facilitate the learning in a new task. Multitask optimization applies multitask learning an optimization to study how effectively and efficiently tackle the multiple optimization problems, simultaneously. Evolutionary multitasking, or multi-factorial optimization, is an emerging subfield of multitask optimization, which integrates evolutionary computation and multi-task learning. This article proposes a novel and easy-to-implement multitasking genetic algorithm (MTGA), which copes well with significantly different optimization tasks by estimating and using the bias among them. Comparative studies with eight state-of-the-art single-task and multitask approaches in the literature on nine benchmarks demonstrated that, on average, the MTGA outperformed all of them and had lower computational cost than six of them. Based on the MTGA, a *simultaneous optimization strategy* for fuzzy system design is also proposed. Experiments on simultaneous optimization of type-1 and interval type-2 fuzzy logic controllers for couple-tank water level control demonstrated that the MTGA can find better fuzzy logic controllers than other approaches."

IEEE Transactions on Evolutionary Computation

Feature Extraction and Selection for Parsimonious Classifiers With Multiobjective Genetic Programming, by K. Nag and N. R. Pal, *IEEE Transactions on Evolutionary Computation*, Vol. 24, No. 3, June 2020, pp. 454–466.

Digital Object Identifier: 10.1109/TEVC.2019.2927526

"The objectives of this paper are to investigate the capability of genetic programming to select and extract linearly separable features when the evolutionary process is guided to achieve the same and to propose an integrated system for that. It decomposes a c-class problem into c binary classification problems and evolve c sets of binary classifiers employing a steady-state multi-objective genetic programming with three minimizing objectives. Each binary classifier is composed of a binary tree and a linear support vector machine (SVM). The features extracted by the feature nodes and some of the function nodes of the tree are used to train the SVM. The decision made by the SVM is considered as the decision of the corresponding classifier. During crossover and mutation, the SVM-weights are used to determine the usefulness of the corresponding nodes. It also uses a fitness function based on Golub's index to select useful features. To discard less frequently used features, it employs unfitness functions for the feature nodes. The method is compared with 34 classification systems using 18 datasets. The performance of the proposed method is found to be better than 432 out of 570, i.e., 75.79% of comparing cases."

IEEE Transactions on Games

Winning Is Not Everything: Enhancing Game Development With Intelligent Agents, by Y. Zhao, I. Borovikov, F. de Mesentier Silva, A. Beirami, J. Rupert, C. Somers, J. Harder, J. Kolen, J. Pinto, R. Pourabolghasem, J. Pestrak, H. Chaput, M. Sardari, L. Lin, S. Narravula, N. Aghdaie, and K. Zaman, *IEEE Transactions on Games*, Vol. 12, No. 2, June 2020, pp. 199–212

Digital Object Identifier: 10.1109/TG.2020.2990865

"Recently, there have been several high-profile achievements of agents learning to play games against humans and beat them. In this article, we study the problem of training intelligent agents in service of game development. Unlike the agents built to "beat the game," our agents aim to produce human-like behavior to help with game evaluation and balancing. We discuss two fundamental metrics based on which we measure the human-likeness of agents, namely skill and style, which are multi-faceted concepts with practical implications outlined in this article. We report four case studies in which the style and skill requirements inform the choice of algorithms and metrics used to train

agents; ranging from A* search to state-of-the-art deep reinforcement learning (RL). Furthermore, we show that the learning potential of state-of-the-art deep RL models does not seamlessly transfer from the benchmark environments to target ones without heavily tuning their hyperparameters, leading to linear scaling of the engineering efforts, and computational cost with the number of target domains."

IEEE Transactions on Cognitive and Developmental Systems

Concrete Action Representation Model: From Neuroscience to Robotics, by J. Nassour, T. D. Hoa, P. Atoofi, and F. Hamker, *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 12, No. 2, June 2020, pp. 272–284.

Digital Object Identifier: 10.1109/TCDS.2019.2896300

"How can robotics benefit from neuroscience to build a unified framework that computes actions for both locomotion and manipulation tasks? Inspired by the hierarchical neural control of movement from cortex to spinal cord, the authors propose a model that generates a concrete action representation in robotics. The action program is composed of four basic modules: 1) pattern selection; 2) spatial coordination; 3) temporal coordination; and 4) sensory motor adaptation. The first and the fourth are considered for behavior initiation. The model is implemented on a humanoid robot to generate rhythmic and nonrhythmic movements. The robot is able to perform tasks like perturbation recovery, and drawing based on different motor programs generated by the same model. Unifying motor control in robotics through a hierarchical structure increases the capacity to gain an accurate and deep understanding of transfer of motor skills between different tasks."

IEEE Transactions on Emerging Topics in Computational Intelligence

Pedestrian Flow Optimization to Reduce the Risk of Crowd Disasters Through Human-Robot Interaction, by C. Jiang, Z. Ni, Y. Guo, and H. He, *IEEE Transactions on Emerging Topics in Computational Intelligence*, Vol. 4, No. 3, June 2020, pp. 298–311.

Digital Object Identifier: 10.1109/TETCI.2019.2930249

"Pedestrian flow in densely populated or congested areas usually presents irregular or turbulent motion state due to competitive behaviors of individual pedestrians, which reduces flow efficiency and raises the risk of crowd accidents. Effective pedestrian flow regulation strategies are highly valuable for flow optimization. Existing studies seek for optimal design of indoor architectural features and spatial placement of pedestrian facilities for the purpose of flow optimization. However, once placed, the stationary facilities are not adaptive to real-time flow changes. In this paper, we investigate the problem of regulating two merging pedestrian flows in a bottleneck area using a mobile robot moving among the pedestrian flows. The pedestrian flows are regulated through dynamic human-robot interaction (HRI) during their collective motion. We adopt an adaptive dynamic programming (ADP) method to learn the optimal motion parameters of the robot in real time, and the resulting outflow through the bottleneck is maximized with the crowd pressure reduced to avoid potential crowd disasters. The proposed algorithm is a data-driven approach that only uses camera observation of pedestrian flows without explicit models of pedestrian dynamics and HRI. Extensive simulation studies are performed in both MATLAB and a robotic simulator to verify the proposed approach and evaluate the performances."

Call for Papers for Journal Special Issues

Special Issue on "Effective Feature Fusion in Deep Neural Networks"

Journal: ***IEEE Transactions on Neural Networks and Learning Systems***
Guest Editors: Yanwei Pang (pyw@tju.edu.cn), Fahad Shahbaz Khan, Xin Lu, and Fabio Cuzzolin
Submission Deadline: November 30, 2020
https://cis.ieee.org/images/files/Documents/call-for-papers/tnnls/SI_EFDNN_TNNLS_CFP.pdf

Special Issue on "Deep Learning for Anomaly Detection"

Journal: ***IEEE Transactions on Neural Networks and Learning Systems***
Guest Editors: Guansong Pang (guansong.pang@adelaide.edu.au), Charu Aggarwal, Chunhua Shen, and Nicu
Submission Deadline: November 30, 2020
https://cis.ieee.org/images/files/Documents/call-for-papers/tnnls/TNNLS_SI_deep_learning_for_anomaly_detection_CFP.pdf

Digital Object Identifier 10.1109/MCI.2020.3019905

Guest Editorial

Vincent S. Tseng
National Chiao Tung University, TAIWAN

Stephen T.C. Wong
Houston Methodist Cancer Center, USA

Diane J. Cook
Washington State University, USA

Jiming Liu
Hong Kong Baptist University, CHINA

This Fast-Track Special Issue is in line with the COVID-19 Initiative of IEEE CIS, aiming to present the latest developments and insights in applying computational intelligence approaches into practical applications for combating COVID-19.

COVID-19 (COronaVIrus Disease 2019), announced by World Health Organization (WHO) in March 2020, has become a global pandemic in very short time and caused infections on tens of millions of people with hundreds of thousands of deaths, which unfortunately are continuing to rise. Not only are the healthcare systems worldwide under high threats, but also the global economy were damaged severely.

During the past few months, all countries are pushing to develop novel and effective solutions for coping with this crisis by leveraging the resources and efforts from the governments, industry and academia. Various emerging solutions and systems for combating COVID-19 have been under development and deployment, ranging from fast screening methods to accurate diagnosis using different kinds of clinical data (X-Ray, CT, vital signs, etc.), risk profiling, patient surveillance, propagation modeling, dashboard visualization and control, drug and vaccine design and social analytics. Computational intelligence plays an important and crucial role in building such kind of solutions since it takes different kinds of computational intelligence technologies, like Neural Networks, Fuzzy Systems and Evolutionary Computation, to deal with the underlying challenges. Moreover, integration of computational intelligence mechanisms with various types of medical systems/devices is essential for practical deployment in existing healthcare environments.

In light of the above observations, this "Fast-Track Special Issue," which is in line with the *COVID-19 Initiative* of *IEEE CIS*, aims at soliciting high-quality articles to share the latest developments and insights in applying computational intelligence approaches into practical applications for combating COVID-19. The overall goal of this special issue is to offer a venue for researchers and practitioners from academia and industry to present the latest technologies and developments in dealing with the challenges brought by COVID-19, with the hope to enlighten new and compelling solutions for combating COVID-19. We were successful in attracting 32 submissions, which were reviewed by at least three competent independent referees and one editor. Following a rigorous peer review process, 5 papers have been accepted for publication in this special issue.

The first paper, "Computational Intelligence Techniques for Combating COVID-19: A Survey" by V. Tseng et al., provides a comprehensive review and categorization of the representative research works on computational intelligence for fighting COVID-19. The collected literatures are categorized into five principles of computational intelligence according to the methods used. Meanwhile, a number of important issues that potentially can be addressed by computational intelligence techniques are listed and categorized. Based on the properties and advantages of each principle, the authors also point out what kinds of issues could be but have not yet been dealt with well. In particular, the authors provide insightful discussions for enlightening future research directions such as hybrid models for dealing with the problem of treatment design. Furthermore, this article reports several promising real-world software tools that applied the computational

Digital Object Identifier 10.1109/MCI.2020.3019872
Date of current version: 14 October 2020

intelligence techniques for prevention of COVID-19.

The second paper, titled "A Bayesian Updating Scheme for Pandemics: Estimating the Infection Dynamics of COVID-19" by S. Wang et al., tackles the important topic of assessing the impacts of different intervention strategies against COVID-19 by applying the technology of data assimilation to estimate epidemiological parameters using observable information. A Bayesian updating scheme for reliable and timely estimation of parameters in epidemic models is proposed. Unlike conventional compartmental models which did not work well for emerging pandemic, a concise renewal model with new parameters has been proposed for modeling the transmission dynamics. The proposed parameters were designed by disentangling the reduction of instantaneous reproduction into mitigation and suppression factors such that the proposed concise renewal model can quantify intervention impacts at a finer granularity. Several promising results are also reported on applying the proposed concise renewal model to estimate the effects of interventions in European countries, the United States, Wuhan as well as the resurgence risk in the USA.

In the third paper, "COVID-19 Time Series Forecast Using Transmission Rate and Meteorological Parameters as Features," M. Mousavi et al. tackle the challenge of forecasting future cases of the virus for a targeted region. The proposed model identifies past transmission rate, temperature, and humidity as key predictive features. They decompose each feature into stationary and non-stationary modes, train two recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) cells based on the corresponding modes, and sum the two RNN predictions to create a forecasted number of future COVID-19 cases. The proposed approach was validated through the evaluation on predicting transmission based on historic numbers of confirmed virus cases for two states in India. Several insights were provided through the evaluation results, in particular on that the data complexity and predictability varies between regions, which points to the need to further investigate the dependencies between transmission and meteorological factors in understanding spread of the virus.

Moreover, integration of computational intelligence mechanisms with various types of medical systems/devices is essential for practical deployment in existing healthcare environments.

The fourth paper, "Meaningful Big Data Integration for a Global COVID-19 Strategy" by J. Pita Costa et al., presents the Meaningful Integration of Data Analytics and Services (MIDAS) platform they developed to connect and integrate heterogeneous data sources including open and social media data, etc., which is ready for deep analytics, event monitoring and research. MIDAS includes the governance and ethnic review organization structure to ensure patient privacy and ethical issues are properly addressed. Such a technology platform can play important and helpful roles for broad spectrum of COVID-19 public health studies, including better understanding of the spread and geosocial impact of the disease, locating the origin of the disease and tracking its mutations, monitoring and following the pandemic across global regions and diverse populations, particularly for health disparities and vulnerable populations; and helping earlier preparation for disease prevention or containment. The large database accumulated in MIDAS also allows the experimentation and development of novel computational intelligent algorithms and tools in analyzing and predicting the COVID-19 pandemic as well as events of the next outbreak.

During the COVID-19 pandemic, community prevention and control can be quintessential in reducing the risks of viral infection and spread. As such, Traditional Chinese medicine (TCM), best known for its holistic approach to treatment and prevention of acute and chronic disorders, can play an important role in that it pays special attention to improving the inherent self-resistance and hence mitigating the likelihood of disease onset. In this regard, the fifth paper entitled "Intelligent Optimization of Diversified Community Prevention of COVID-19 using Traditional Chinese Medicine" by Y. Zheng et al. demonstrates the TCM principle of "treatment based on syndrome differentiation" when developing targeted TCM prevention programs for people with different needs, as opposed to a one-size-fits-all prevention program. In doing so, the authors have utilized an improved fuzzy clustering method to differentiate the population based on TCM health and medicine related characteristics. Thereafter, for each of the clusters identified, TCM experts would recommend a specific prevention program. The authors have employed a bio-inspired heuristic algorithm to further optimize the programs, so as to make the best use of the available resources. This article reports several promising results of applying the proposed method on TCM-based prevention of COVID-19 in 12 communities in Zhejiang province, China, during the peak of the pandemic.

We sincerely thank all of the authors who submitted their papers to this special issue, and to a large number of reviewers who dedicated their time and expertise for materializing a high-quality special issue on this very important and timely topic. In particular, we would also like to thank Prof. Chuan-Kang Ting, the Editor-in-Chief of IEEE Computational Intelligence Magazine (IEEE CIM) for his great efforts in initiating and developing this special issue together, and all members of the editorial team for their enthusiastic support during the editing process of this special issue.

Computational Intelligence Techniques for Combating COVID-19: A Survey

©SHUTTERSTOCK/IRINA SHI

Vincent S. Tseng
National Chiao Tung University, TAIWAN.

Josh Jia-Ching Ying
National Chung Hsing University, TAIWAN.

Stephen T. C. Wong
Houston Methodist Cancer Center, USA.

Diane J. Cook
Washington State University, USA.

Jiming Liu
Hong Kong Baptist University, CHINA.

***Abstract*—**Computational intelligence has been used in many applications in the fields of health sciences and epidemiology. In particular, owing to the sudden and massive spread of COVID-19, many researchers around the globe have devoted intensive efforts into the development of computational intelligence methods and systems for combating the pandemic. Although there have been more than 200,000 scholarly articles on COVID-19, SARS-CoV-2, and other related coronaviruses, these articles did not specifically address in-depth the key issues for applying computational intelligence to combat COVID-19. Hence, it would be exhausting to filter and summarize those studies conducted in the field of computational intelligence from such a large number of articles. Such inconvenience has hindered the development of effective computational intelligence technologies for fighting COVID-19. To fill this gap, this survey focuses on categorizing and reviewing the current progress of computational intelligence for fighting this serious

Digital Object Identifier 10.1109/MCI.2020.3019873
Date of current version: 14 October 2020

Corresponding Author: Vincent S. Tseng (vtseng@cs.nctu.edu.tw).

disease. In this survey, we aim to assemble and summarize the latest developments and insights in transforming computational intelligence approaches, such as machine learning, evolutionary computation, soft computing, and big data analytics, into practical applications for fighting COVID-19. We also explore some potential research issues on computational intelligence for defeating the pandemic.

I. Introduction

COVID-19 is an infectious disease caused by a novel coronavirus and has been declared by the World Health Organization (WHO) as a pandemic in March 2020. Since this disease was first identified in December 2019, it has become a global pandemic and has caused infections in millions of people. The coronavirus death toll surpassed 687,000 worldwide as of the end of July 2020, and the number of infections and deaths continues to rise. Such an extremely serious situation has led to high threat in healthcare systems worldwide and severe damage in the global economy being. To combat COVID-19, many countries are working to develop novel and effective mechanisms to overcome this disaster. Governments, industry leaders, and academics alike are devoting substantial resources and effort into mitigating the effects of the pandemic. Over the past few months, various emerging solutions and systems for combating COVID-19 have been developed and deployed. For example, fast screening methods utilizing different types of clinical data, including X-rays, computed tomography (CT) scans, and vital signs, have enabled timely diagnosis and disease monitoring. Computer systems are also being designed for risk profiling, patient surveillance, contact tracing, or propagation modeling by using social media data.

Owing to the advancement of computational intelligence, numerous integrations of computational intelligence mechanisms with various devices and systems have already achieved considerable success in dealing with the underlying challenges of epidemic diseases such as new influenzas [1], SARS [2], and MERS [3]. As a result, many systems and solutions for combating COVID-19 have adopted computational intelligence, and the design of proper computational intelligence mechanisms plays a crucial role in building such solutions. Since the integration of computational intelligence mechanisms with various devices and systems under different application conditions would require different types of computational intelligence techniques, including data analytics, computational modeling, high-performance computing, artificial intelligence, and in particular its subfield of machine learning, many researchers have devoted their efforts to developing systems of computational intelligence specifically for the fight against COVID-19.

By the end of July 2020, more than 200,000 scholarly articles were published regarding COVID-19, SARS-CoV-2, and other related coronaviruses [4]. However, these articles did not address in-depth the key issues in applying computational intelligence to combating the COVID-19 pandemic. Thus, it would be exhausting to filter and summarize studies related to computational intelligence from such a large number of articles. In light of the above observations, now is the time to systematically categorize and review the current progress of research on computational intelligence. Accordingly, this survey aims to assemble and summarize the highlights of the latest developments and insights in applying computational intelligence approaches, such as machine learning, evolutionary computation, soft computing, and big data analytics, to practical applications used to combat COVID-19.

The remainder of this paper is organized as follows. In Section II, we briefly survey the history of computational intelligence. In Sections III through VII, we categorize computational intelligence into its five principles and determine the urgent issues concerning COVID-19, which have been, or can be, resolved using computational intelligence approaches. We then review the current computational intelligence studies that have attempted to address these urgent issues based on these five principles. Then, in Section VIII, we review some current systems or applications for combating COVID-19 that have employed principles of computational intelligence. Finally, in Section IX, we conclude the article and discuss recommendations for future studies.

II. Overview

Computational intelligence techniques have already been successfully integrated into various systems for dealing with the underlying challenges of epidemic diseases. Before we introduce the specific issues which computational intelligence can be used to solve to fight COVID-19, we should first understand the history and various categories of this method. Based on the principles of computational intelligence, we can further clarify what types of issues can be dealt with when battling COVID-19 with computational intelligence.

A. Brief Introduction to Computational Intelligence

Computational intelligence was formally defined by Bezdek in 1994 [5] [6] such that a system is called "computationally intelligent" if the system deals with data on a basic level (such as pixels of an image), contains a module of pattern recognition, and does not utilize prior knowledge in the sense of artificial intelligence. According to Bezdek's definition, computational intelligence is one branch of artificial intelligence. Actually, the goals of both artificial intelligence and computational intelligence are the same, which is to realize general intelligence. Marks [7] clarified the difference between artificial intelligence and computational intelligence by claiming that the former is made from hard computing technologies, whereas the latter is made from soft computing technologies.

Therefore, we can presume that two types of machine intelligence exist: 1) artificial intelligence, which is developed by the concept of hard-computing and 2) computational intelligence, which is developed by the concept of soft-computing.

Based on our observations, there are several urgent issues related to COVID-19 that must be combatted. These issues can be categorized into five topics: tracking and predicting virus propagation (TPVP), characterization of symptoms of virus infections (CSVI), treatment design (TrD), precaution development (PD), and public health policy making (PHPM).

Compared to the hard-computing-based artificial intelligence, computational intelligence can adapt to many different conditions via the benefits of the concept of soft-computing. Hard computing techniques are designed using a Boolean logic based only on true or false values that information engineering relies on. One critical issue in Boolean logic is that Boolean values are unable to interpret natural language easily. However, based on fuzzy logic, soft computing techniques can deal with uncertain cases. This type of logic is one proprietary aspect of computational intelligence, and by aggregating data into partial facts, it is approximated to the manner in which the human brain acts [7].

B. Categorization of Computational Intelligence

As mentioned above, the notion of computational intelligence has been around for 30 years. During this period, new concepts have been constantly added to the field, thereby reinforcing the discipline. Today, we can broadly divide computational intelligence techniques into five categories: neural networks, fuzzy logic, evolutionary computation, computational learning theory, and probabilistic methods.

1) Neural Networks

Based on biological neural networks, an artificial neural network (called "neural network" for short) is designed as a network of artificial neurons or nodes. Artificial neural networks can be used for regression or classification modeling for prediction and automatic control. A large number of simulation data using limited data sets. This structure is the foundation of deep learning, which is good at representation learning. Accordingly, artificial neural networks process and learn information from data via the systems of distributed information processing [8]. By doing so, one of the crucial properties of artificial neural networks is fault tolerance, which is approximately modeled on the manner in which the human brain operates [6]. Based on these characteristics, neural networks have been widely applied to data analytics, clustering, classification, and automatic control engineering. In real-world applications, such methods aim to analyze and classify medical data, recognize human faces, detect computer fraud, and deal with the nonlinearity of a system for better process control [9]. Furthermore, neural network techniques can incorporate fuzzy logic concepts.

2) Fuzzy Logic

Fuzzy logic [10] can be seen as a formulation defined by multi-valued logic. Meanwhile, the true value of a variable's formulation can be any real number between, but not limited to, 0 and 1. It is often utilized to solve the problem of uncertainty, where the truth value may not be all true or all false. As a result, fuzzy logic has been successfully applied in the field of clinical realms, including a continuous blood glucose prediction system and a tuberculosis diagnosis platform based on chest X-ray, among other devices. We can also see this in use of a video camera to help stabilize an image endoscope. Other areas such as household appliances, business decision making, and financial analysis are also examples of applications of this principle [6]. A main application of fuzzy logic is approximate reasoning. However, the methods of fuzzy logic reasoning usually lack learning abilities, which are necessary for a multitude of tasks.

3) Evolutionary Computation

Evolutionary computation (EC) is global optimization method inspired by biological evolution [6]. It is a family of algorithms and is a branch of computational intelligence and natural computing. EC systems solve problems via populations, error and success, meta-heuristics, or stochastic optimization. An initial set of candidate solutions is generated and updated iteratively, such as the removal of less-desired solutions and the insertion of noise. A population of solutions is subject to natural selection or artificial selection and mutation, and therefore evolves and adapts—i.e., increases fitness (function quantizes how adapted/desired the solution is). EC is popular in computational intelligence because it results in near-optimal solutions in a wide spectrum of contexts [11] where there are many variants and extensions for specific data structures and problems.

4) Computational Learning Theory

Computational learning theory (referred to as learning theory for short) is a sub-field of artificial intelligence mainly for the research and development of learning strategies for machine learning. Computational learning theory is one of the principal methods in computational intelligence, which seeks for a way to achieve reasoning that recapitulates human rezoning humans. In psychology, learning is the process of enhancing or changing knowledge, skills, values, and world views through cognition and experience [6]. Inspired by psychology, computational learning theory is utilized to actualize the process of experience and decision making according to previous experiences.

5) Probabilistic Methods

A probabilistic method is a nonconstructive approach used to prove the existence of specified types of mathematical objects. It operates by showing that if one randomly selects an object from a specified category, the probability that the result will

become the specified type is strictly greater than zero. Although probabilistic methods are designed based on the probability theories, the results are determined with certainty and without any possible errors. Probabilistic methods were first introduced as the main foundation of fuzzy logic by Erdos and Spencer [6]. To evaluate the outcomes of a system based on computational intelligence, probabilistic methods are mostly defined by randomness [12]. Accordingly, probabilistic methods can provide proper solutions to problems based on prior knowledge.

C. Issues on Fighting COVID-19

Based on our observations, there are several urgent issues related to COVID-19 that must be combatted. These issues can be categorized into five topics: tracking and predicting virus propagation (TPVP), characterization of symptoms of virus infections (CSVI), treatment design (TrD), precaution development (PD), and public health policy making (PHPM). The issues of each topic are listed below:

1) *Tracking and Predicting Virus Propagation (TPVP)*
   - Surveillance and tracking of COVID-19-infected patients.
   - Modeling and predicting virus propagation and pathways.
   - Visual analytics techniques and applications for propagation modeling and monitoring.
2) *Characterization of Symptoms of Virus Infections (CSVI)*
   - Discovery of early markers/symptoms of viral infections.
   - Personalized and group-based risk profiling and prediction.
   - Real-time and early alerting systems for hazardous and forefront outbreaks.
   - Fast and accurate diagnosis of COVID-19 through analytics and modeling using various biomedical data, e.g., images, vital signs, genome, etc.
3) *Treatment Design (TrD)*
   - Treatment optimization and care planning for the best care of patients.
   - Prognosis and outcome prediction for patients for effective resource allocation.
   - Drug discovery and repurposing through big data analytics approaches.
4) *Precaution Development (PD)*
   - Vaccine design through machine learning approaches.
   - Intelligent analysis of social media and networks for contact tracing and safety control.
   - Integrations of intelligent computing mechanisms with information technology systems and the internet of things (IoT) for smart care in COVID-19.
5) *Public Health Policy Making (PHPM)*
   - Secure and privacy-preserving analysis of data in public health emergencies.
   - Public health policy making through big data analytics and model simulations.

To overcome these issues, researchers are actively conducting research to obtain various outcomes, and many have adopted computational intelligence. Since there are so many studies that address the issues associated with COVID-19, this survey focuses on studies designed using computational intelligence were selected for discussion in the following sections.

III. Neural Networks for Combating COVID-19

As mentioned earlier, an artificial neural network applies the principle of deep learning and achieves a high level of representation learning. Representation learning is a learning method that can automatically learn representations from data. Learning algorithms do not require humans to help them extract features. Accordingly, neural networks can be easily utilized for extracting important representations of virus propagation and the characteristic symptoms of a viral infection, among other factors. However, deep neural networks can extract useful knowledge mostly if the amount of data is sufficient, which means they can hardly deal effectively with incompleteness or, most importantly, data with missing values in the processing model. Fortunately, many countries have established a series of COVID-19 data collection mechanisms. As a result, building an effective deep neural network to fight COVID-19 is possible. For example, many researchers have built deep neural network models for characterizing viral infections by using CT chest images, as shown in Fig. 1, or X-ray images [23], [24]. Table I summarizes the issues that have been addressed by existing neural network methods. Owing to their excellent ability to extract important representations, most applications involving neural networks have addressed the issue of CSVI, as shown in Table I.

Roy *et al.* [14] proposed the application of artificial neural networks for analyzing lung ultrasonography (LUS) images.

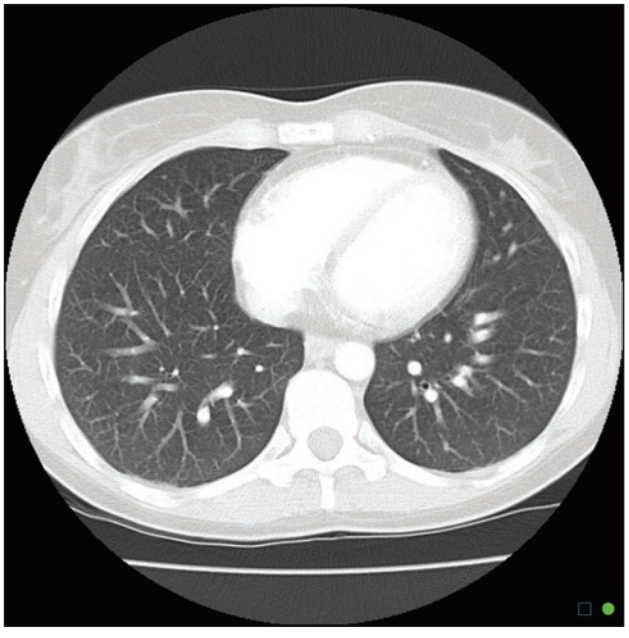


FIGURE 1 Example of computed tomography (CT) image [13].

They collected a new fully-annotated LUS image dataset from several hospitals in Italy, and the labels indicate the severity of the disease at the frame, video, and pixel levels (segmentation masks). Using these data, several artificial neural network models have been developed to solve the tasks related to the automatic analysis of LUS images. To predict the severity of the disease associated with an input frame, an extension of spatial transformer networks was proposed, which can provide localization of the diseased area in a weakly supervised manner. To conduct scoring at the video level, an effective frame score aggregation function was proposed, and three artificial neural networks, vanilla U-Net [25], U-Net++ [26], and Deeplabv3+ [27], have been adopted for the segmentation of COVID-19 imaging biomarkers at the pixel level.

Wang *et al.* [15] proposed a CNN-based model, COVID-Net, for detecting COVID-19 infected patients from a dataset of chest radiography images. The dataset is an open dataset consisting of 13,800 images collected from 13,725 patients. COVID-Net utilizes a novel lightweight residual block, the projection-expansion-projection-extension (PEPX), to improve representational capacity while maintaining reduced computational complexity. Furthermore, COVID-Net is designed to make predictions using a qualitative analysis method called GSInquire, to obtain deeper insight into crucial features related to COVID-19 infected patients, which can assist clinicians in efficient and precise diagnosis.

Han *et al.* [16] presented an attention scheme involving deep 3D multiple instance learning called AD3D-MIL to learn a detection model from 3D chest CTs. With the attention scheme AD3D-MIL, not only can it accurately predict an individual category of disease such as COVID-19, common pneumonia, or no pneumonia, but it also produces interpretability of results. During the learning process, users will not receive a set of labeled instances, but each bag contains many instances, and each bag has a label rather than separately labeled sets of instances. The idea behind AD3D-MIL is to treat all CT images of an individual patient as the instances of a labeled bag. Meanwhile, a fully 3D convolutional neural network is used to produce the feature map of each instance, and an attention-based MIL pooling is designed to select and combine the feature maps into a bag representation. Finally, the bag representation is fed into a typical fully-connected neural network to make the final predictions.

TABLE I Issues addressed by existing neural network methods.

	TPVP	CSVI	TRD	PD	PHPM
ROY *ET AL.*'S STUDY [14]		✓			
COVID-NET [15]		✓			
HAN *ET AL.*'S STUDY [16]		✓			
PANWAR *ET AL.*'S STUDY [17]		✓			
OH *ET AL.*'S STUDY [18]		✓			
APOSTOLOPOULOS *ET AL.*'S STUDY [19]		✓			
WANG *ET AL.*'S STUDY [20]		✓			
AYYOUBZADEH *ET AL.*'S STUDY [21]	✓		✓		
VAID *ET AL.*'S STUDY [22]		✓			

Panwar *et al.* [17] developed a deep learning-based COVID-19 detection model that can detect a COVID-19 positive patient within 5 seconds using X-ray images. The proposed model extends VGG-16 by adding five custom layers as the head layers, of which the first layer is an average pooling 2D layer. Unlike max pooling, this average pooling layer uses the average value of all the pixels with a pool size of (4, 4) to down-sample the images. The second layer is a flattened layer that transforms a two-dimensional tensor into a vector as an input of a fully dense connected layer (i.e., the third layer). Meanwhile, the activation function of the fully dense connected layer is ReLU. The fourth layer is a dropout layer that ignores half of the units of the fully dense connected layer. The fifth layer is the output layer, which uses two units to produce the confidence values for the infected and uninfected, respectively. Based on a pre-trained VGG-16 with the five layers added, the proposed model was able to achieve a 97.62% true-positive rate with a limited amount of data, consisting of 142 images of uninfected and 192 images of infected people.

The training of neural networks with limited training sample sizes is key to applying deep learning to address the issues regarding COVID-19. To deal with the limited data size, Oh *et al.* [18] developed a neural network for COVID-19 diagnosis that is suitable for training with limited X-ray images. An extended fully convolutional neural network called (FC)-DenseNet103 [28] was adopted for lung image segmentation. The results of the lung image segmentation from the segmentation networks are utilized for masking the pre-processed images. To classify the masked images, ResNet-18 [29] was adopted to build a classification model. Meanwhile, the classification model was implemented with two different contexts: global appearance and zooming in a partial area. To consider the view of global appearance, each masked image is resized to 224×224 so that each input is a complete X-ray image. Oh *et al.* utilized this approach as a baseline network for experimental evaluation. To consider zooming in a partial area of an X-ray image, each masked image is cropped randomly to produce several 224×224 images so that a masked image may produce several input images. Although the overall accuracy of this approach is 91.9%, slightly lower than that of COVID-Net [15] (92.4%), the model size of this approach (11.6 M parameters) is much smaller than that of COVID-Net (116.6 M parameters). In other words, this approach requires much less data to train the model.

Apostolopoulos *et al.* [19] examined the significance of the extracted features and utilized MobileNet V2 [30] to train a

classification model using 3,905 X-ray images for classification of six similar diseases, including COVID-19. As the symptoms shown in the X-ray images of the five diseases are very similar to those of COVID-19, the idea behind the training scheme is to use similar cases to extract reliable features. The proposed model, trained based on MobileNetV2, achieved 99.18% accuracy in detection of COVID-19, but the overall accuracy of the seven classes, including six diseases and one normal group, was about 88%. This phenomenon may suggest that vital biomarkers of COVID-19 can be brought to light by using the proposed model trained on MobileNetV2.

In addition to building a deep neural network model for characterizing viral infections using CT or X-ray images, Wang *et al.* [20] developed a deep learning system to fully automatically diagnose and prognose COVID-19. This system consists of three parts: 1) automatic lung segmentation, 2) non-lung area suppression, and 3) COVID-19 diagnostic and prognostic analysis. For automatic lung segmentation, DenseNet121 [31] was developed and combined with feature pyramid networks (FPN) [32] to produce a lung-ROI, which contains the whole lung and all inflammatory tissues and eliminates most areas outside the lung. Therefore, the lung-ROI would contain some non-lung tissues. A non-lung area suppression operation was proposed to decrease the luminance of non-lung areas inside the lung-ROI. Finally, a novel neural network called COVID-19-Net was proposed for diagnostic and prognostic analyses.

Google Trends has previously been used to accurately predict the outbreak of a new flu. Ayyoubzadeh *et al.* [21] imitated Google Trend's method of analysis for the prediction of incidence of COVID-19 in Iran. Two types of machine learning methods, i.e., linear regression and a recurrent neural network with long short-term memory (LSTM), were adopted to build the prediction model. The effectiveness of the linear regression model achieved 7.562 ± 6.492 in terms of RMSE, while the model utilizes factors such as previous day incidence, hand sanitizer, antiseptic topics, and the frequency of searches for handwashing. The effectiveness of the LSTM model only reached only 28.487 ± 20.705 in terms of RMSE. In addition, the LSTM model showed a fluctuating performance and a low training loss, which might have been caused by overfitting. The reason why a linear regression outperforms an LSTM might be that the data size of the daily incidence of COVID-19 in Iran is quite small, whereas the capacity of the deep learning model, such as the LSTM, is too high for such limited data. Therefore, the LSTM model overfits the limited number of training data easily.

To overcome the problem of a limited amount of training data, Vaid *et al.* [22] developed a transfer learning approach to build a deep learning model by transferring pre-trained CNNs. Structural abnormalities are key to uncover hidden patterns. Based on the transfer learning approach, a pre-trained detection model from anterior-posterior radiographs of the chest of patients was transferred to detect structural abnormalities and disease categorization. Publicly available datasets consisting of patient information from multiple countries were used to refine the pre-trained model and improve the accuracy. The experimental results showed an extremely high accuracy, of 96.3%, and a low loss, of 0.151, in terms of binary cross-entropy. Meanwhile, the proposed model was able to accurately identify 74 true negatives and 32 true positives while incorrectly identifying three false-positives and one false-negative.

According to the papers we surveyed in this section, several insightful findings can be gleaned:

1) Most existing studies on neural networks have focused on dealing with the problem of characterization of symptoms of virus infections. The primary reason for this might be that many LUS images have been produced and collected, and many pre-trained CNNs can be retrieved from open sources. Accordingly, many works straightforwardly utilized pre-trained CNNs to extract vital biomarkers of COVID-19 from LUS images. The variants of neural networks are numerous and varied, which can be utilized for addressing various issues. For example, recurrent neural network (RNN) and its variants are suitable for time series analysis. As more and more patients are cured, many treatment records will be produced. Such treatment records could be viewed as a set of time series data. Therefore, we believe a hybrid model that combines the aforementioned CNN-based works with RNN-based mechanisms is a promising research direction to address the issues on treatment design (TrD).
2) Since datasets are very limited, using a pretrained model is a promising way to quickly produce accurate results. Panwar *et al.* [17] achieved the highest performance using this approach (97.62% true-positive rate). However, using pretrained CNNs for the extraction of vital biomarkers of COVID-19 usually requires labeled LUS images, and this motivates [18], [19] to consider building a lightweight model from small datasets. However, the design for a lightweight model may limit the applicability of these existing works: they may be effective for only a small group of patients who have the same symptoms as shown in the training images.
3) More and more countries have been utilizing the IoT technologies for smart care in COVID-19. Massive data collected from the IoT devices may benefit the effectiveness and applicability of the artificial neural network model. Based on the data collected from IoT devices, integrations of intelligent computing mechanisms with these existing works, such as Ayyoubzadeh *et al.*'s study [21], may be applied to address the issue on precaution development (PD) and the prediction of virus propagation.

IV. Fuzzy Logic for Combating COVID-19

Fuzzy logic is one of the main principles of computational intelligence and enables measurements and process modeling for complex processes in real life. Unlike artificial intelligence, which requires precise knowledge, fuzzy logic may be used with incomplete and even incorrect data applied in a process model. That is, fuzzy logic can easily be used for an

uncertainty-based analysis of limited data. Most of the issues in the fight against COVID-19 can be inherently dealt with using fuzzy logic. For example, when characterizing viral infections, we can find that different patients may have different symptoms, which makes COVID-19 difficult to be diagnosed without virus testing. Table II shows the issues addressed by existing fuzzy logic methods.

Dhiman and Sharma [33] utilized six input factors to build a fuzzy inference system to diagnose COVID-19. By using the proposed fuzzy inference system, the severity level of the infected patients can be presented with three linguistic categories: low, medium, and high. Through a series of training and optimizations, the model learned the following three fuzzy rules:

1) When the atmospheric temperature is medium, if a patient takes a high volume of ethanol and has a slight body temperature, the patient may have a medium severity level.
2) When the atmospheric temperature is low, if a patient takes a low volume of ethanol, has a medium body temperature and suffers from a cough, then the patient may have a low severity level.
3) If a patient takes a low volume of ethanol and has high breath shortness and a sneezing problem, then the patient may have high severity level.

We can see that these rules are reasonable, and thus the proposed inference system can be utilized to accelerate the preliminary diagnosis of COVID-19-infected patients.

The transfusion of blood plasma of recovered COVID-19 patients has been recognized as one possible treatment method. The development of a donation system that can distinguish whether donors have undergone COVID-19 infection is extremely important. However, some infected people have no symptoms, and performing mass testing on all donors is unrealistic. Nazarov [34] used statistical data to construct a fuzzy model to evaluate the quality of blood from donor systems in the Sverdlovsk region in Russia. The evaluation of factors reflects not only the number of donors who have experienced COVID-19 but also the general statistics of donation based on region, age, gender, regularity of blood donation, and the number of donors per 1000 people. The system uses 12 input variables to estimate three output variables according to three rule blocks. By doing so, the problem of evaluating the quality of blood from donor systems in each region can be solved.

TABLE II Issues addressed by existing fuzzy logic methods.

	TPVP	CSVI	TRD	PD	PHPM
DHIMAN AND SHARMA'S STUDY [33]		✓			
NAZAROV'S STUDY [34]	✓		✓		
TINH'S STUDY [35]	✓				
YANG *ET AL.*'S STUDY [36]					✓
REN *ET AL.*'S STUDY [37]			✓		

Tinh [35] utilized a fuzzy time series model combined with particle swarm optimization to forecast the trend in the number of confirmed cases of COVID-19 in Vietnam. Unlike a conventional fuzzy logic model, the proposed fuzzy model uses fuzzy relationship groups, instead of a fuzzy relationship matrix, in the building of the fuzzy forecasting model. To learn the fuzzy rules for constructing fuzzy logical relationship groups, a particle swarm optimization algorithm was designed to determine the proper number of intervals and to refine the length of each interval. The basic idea behind fuzzy logical relationship groups is that the fuzzy logical relationships, which have the same precedence, can be grouped together into a fuzzy logical relationship group. This approach can deal with the problem of time-series forecasting based on limited data. Accordingly, the best performance achieved 2.85% MAPE based on setting a fifth-order fuzzy time series with an interval number of 16.

Yang *et al.* [36] defined the new form of a spherical normal fuzzy set (SpNoFS) that could be used to generate operational rules. Based on the operational rules, a decision support algorithm was designed for optimization of antivirus mask selection. Owing to the complementary use of the Bonferroni mean operator, the new information aggregation operators that can evaluate the utility of antivirus mask selection were formed via the operational rules of SpNoFS. Since the Bonferroni mean operator has two types (Bonferroni mean and weighted Bonferroni mean), there are two kinds of aggregating operators. One is the operator formed by the spherical normal fuzzy rules using the Bonferroni mean (SpNoFBM), and the other is the operator formed by the spherical normal fuzzy rules made using the weighted Bonferroni mean (SpNoFGBM). Based on the SpNoFBM and SpNoFGBM operators, a multi-criteria decision-making method can be realized to reasonably select suitable antivirus masks during the COVID-19 pandemic.

Ren *et al.* [37] adopted the Dempster–Shafer theory to design a multi-criterion decision-making method and the concept of generalized Z-numbers to select medicine for patients with mild symptoms of COVID-19. Meanwhile, the idea behind the medicine selection based on generalized Z-numbers was extended with the expression of human habits, inspired by the concept of a hesitant fuzzy linguistic term set. To avoid ambiguity in the expression form of the generalized Z-numbers, the identification framework in the Dempster–Shafer theory was employed to describe the expression form of generalized Z-numbers. For medicine selection, the basic probability assignment of the evidence could be derived by the expression form of the generalized Z-numbers, and all evaluations of each delivered medicine could be integrated by using the synthetic rules in the Dempster–Shafer theory.

Several insightful findings have been gleaned from the papers surveyed in this section:

1) Fuzzy logic is good at reasoning by analyzing uncertainty from limited data. Therefore, we can see that most kinds of issues we mentioned have been addressed by the fuzzy

logic models we surveyed. This may imply that we can combine fuzzy logic and neural networks so that the hybrid system can be applied for almost all issues relevant to COVID-19.

2) Since the main advantage of the fuzzy logic methods is to produce fuzzy rules that can deal with the uncertainty from limited data, the findings from fuzzy rules are expected to be shown in the literature. However, most relevant existing works [35]–[37] did not provide such discussions. One relevant work [34] only briefly introduced the fuzzy rules without deep discussion. This has led to the low applicability of these works. This is noteworthy for future studies on fuzzy logic.

V. Evolutionary Computation for Combating COVID-19

Evolutionary computation initially creates a set of candidate solutions and refines the set iteratively. The set of candidate solutions at each iteration is called the population. By stochastically removing the less-desired solutions and putting small random changes in the current generation, the next generation is produced. In biological terms, a set of solutions undergoes natural selection (or manual selection) and mutation. As a result, the population incrementally increases in fitness. The fitness function of the algorithm determines the goal of learning. Evolutionary computation techniques can produce highly optimized solutions for various problems. Many variants and extensions have been designed for group-based risk profiling, and they are suitable for analysis of the possible impacts of COVID-19 and forecasting how COVID-19 will behave in the future. Table III shows the issues addressed by existing evolutionary computation methods.

Yousefpour *et al.* [38] combined Susceptible, Exposed, Infectious, and Recovered (SEIR) [43] with a multi-objective genetic algorithm that focuses on epidemic prevention and economic concerns to estimate the early transmission dynamics of COVID-19. Besides, Yousefpour *et al.* utilized the estimation results to find the best decision rules. Two cost functions were designed and involved in the multi-objective genetic algorithm. The first cost function represents epidemic prevention:

$$J_1 = \sum E(t) + A(t), \tag{1}$$

where $E(t)$ indicates the number of exposed people at time t, and $A(t)$ indicates the number of asymptomatic infected people at time t. The second cost function represents economic concerns:

$$J_2 = -\eta_1 (c_0 + c_f) + \eta_2 q, \tag{2}$$

where c_0 denotes the contact rate at the initial time, c_f denotes the minimum contact rate under the current control strategies, and q denotes the quarantined rate of exposed individuals. Based on [38], the optimal policies were designed and showed that treating infection control as an optimization problem can protect countries against both disease outbreaks and economic breakdown.

Niazkar *et al.* [39] adopted the multi-gene genetic programming (MGGP) to predict COVID-19 outbreaks. Since the numbers of daily confirmed cases fluctuate, predicting a COVID-19 outbreak is a challenging task. MGGP was originally designed for behavioral modeling, which is suitable for modeling series with high fluctuations. The proposed method based on MGGP showed very promising results. More specifically, the predicted number of confirmed cases of COVID-19 approximated the observations in the seven countries considered in their study. Therefore, the MGGP-based approach has been suggested to be appropriate for the estimation of COVID-19 outbreaks.

Salgotra *et al.* [40] proposed a prediction model by developing genetic programming (GP) which analyzes the possible impact of COVID-19 in India and predicts the future behavior. The developed GP predicted the number of confirmed cases and numbers of death cases in the three most affected states in India. The fitness function was designed with respect to the mean squared error. To validate the evolved models, statistical parameters and metrics were used to evaluate the fitness. Furthermore, the proposed GP-based models were lined with each other by using simple linkage functions for gene size greater than 1. The experimental results showed that the proposed GP-based models are significantly reliable for predicting the numbers of confirmed and death cases in India.

To expand the contributions of GP for predicting the possible impact of COVID-19 in India, Salgotra *et al.* [41] further applied their GP to build a prediction model for forecasting the potential effects of COVID-19 in the 15 most affected countries in the world. The prediction model estimated that the daily confirmed cases and daily death count would result in a negative value in China. Besides the results in China, the overall prediction results are listed in Table IV. We can find that Brazil had the highest daily increase in the COVID-19 reproduction rate. This prediction was made at the end of May, and Brazil's situation did fall into its worst in June. This indirectly proves the applicability of Salgotra *et al.*'s study.

TABLE III Issues addressed by existing evolutionary computation methods.

	TPVP	CSVI	TRD	PD	PHPM
YOUSEFPOUR *ET AL.*'S STUDY [38]	✓				✓
NIAZKAR *ET AL.*'S STUDY [39]	✓			✓	
SALGOTRA *ET AL.*'S STUDY [40]	✓			✓	
SALGOTRA *ET AL.*'S STUDY [41] (EXTENSION OF [40])	✓			✓	
DILBAG *ET AL.*'S STUDY [42]		✓			

Dilbag *et al.* [42] proposed a multi-objective differential evolution algorithm to optimize the hyperparameters of the regular CNN that was trained from CT images for classification of COVID-19-infected patients. A multi-objective fitness function was designed according to both the sensitivity and specificity of classifications of COVID-19-infected patients. According to Dilbag *et al.*'s experiments, the proposed model slightly outperformed state-of-the-art models, such as a regular CNNs, an adaptive neuro-fuzzy inference system, and an artificial neural network. The overall improvement in terms of accuracy was 1.9789%.

According to the papers we surveyed in this section, several insightful observations can be made:

1) Since evolutionary computation was designed for the optimization of parameters, we can see that most works, like [38]–[40], utilize evolutionary computation to predict virus propagation. Meanwhile, some of these works [38], [39] adopted the concept of multi-objective genetic algorithm to estimate the number of confirmed cases and tackle other properties such as economic concerns. The prediction results can also address other issues such as precaution development.
2) Although multi-objective genetic algorithm could be utilized for solving multi-objective problems, the works [37], [38] only adopted them straightforwardly without any modifications. Since the application scenarios of the works of [37], [38] are different from that of multi-objective genetic algorithms, their effectiveness is not significant in supporting their reliability and applicability. We believe these works could be further improved. For example, the interaction among the multiple fitness functions could be included to adjust for the process of optimization.

VI. Computational Learning Theory for Combating COVID-19

Computational learning theory has many implementations. Based on different assumptions, various inference principles can be deduced. As a result, the deduced inference principles are utilized to design different computational learning theory approaches. These approaches can usually be categorized into six types: 1) exact learning; 2) probably approximately correct learning, which is a machine learning framework based on mathematical analysis; 3) Vapnik–Chervonenkis theory, which is a learning process explained by a statistical point of view; 4) Bayesian inference, which is a statistical inference based on Bayes' theorem; 5) algorithmic learning theory, which is a machine learning theory explained by an algorithmic point of view; and 6) online machine learning, which is a sort of machine learning method for continuously updating data. Although its primary goal is to understand learning in an abstract manner, through the development of learning theory, we can design various practical learning algorithms. For example, Bayesian inference is the foundation of the concept of belief networks. Because the concept of belief networks is the foundation of the deep neural networks introduced in the previous section, we will introduce the remaining approaches in this section, which are designed based on belief networks except deep neural networks. Table V shows the issues that have been addressed by existing computational learning theory methods.

Duffey and Zio [44] proposed a computational learning theory that can learn a prediction model from the prediction errors in the recovery time from the outbreak of the COVID-19 pandemic. This approach uses the exponential Universal Learning Curve to estimate the trend in the infection rates of the COVID-19 pandemic. The key to the proposed approach is to treat the infection rate as a measure of false prediction results and time as a measure of experience/knowledge or risk exposure to allow learning. The results of Universal Learning Curve, which was learned from China, South Korea, and other nations, show a decreasing trajectory after a peak. The reason might be that countermeasures are effective for controlling the spread of the virus.

Wang *et al.* [45] proposed a novel noise-robust learning framework called COPLE-Net based on the self-ensemble of convolutional neural networks [50], [51], a sort of semi-supervised learning mechanism. Unlike conventional semi-supervised learning mechanisms that use the exponential moving average of a model to adjust standard model, Wang *et al.* [45] developed two designs to address the issue on noisy labels. The first design is a dynamic adjustment that can reduce the impact of the exponential moving average of a model while the training loss is decreased. The second is an adaptive learner that enables the standard model to learn from the exponential moving average of a model. The proposed COPLE-Net outperforms state-of-the-art models in terms of the average Dice similarity (80.29%) and the average 95-th percentile of Hausdroff distance (18.72 mm).

Barstugan *et al.* [46] presented an early phase detection method for COVID-19 using a support vector machine classifier. The classifier was trained from four extensive datasets, which were produced by fetching patches with sizes of 16 × 16, 32 × 32, 48 × 48,

TABLE IV Prediction results of Salgotra *et al.*'s study [41].

COUNTRY	DAILY CONFIRMED CASES	DAILY DEATH COUNT	COUNTRY	DAILY CONFIRMED CASES	DAILY DEATH COUNT
USA	20,972	1358	TURKEY	1,071	17
BRAZIL	28,822	1076	CANADA	717	103
RUSSIA	6,928	270	SPAIN	321	148
MEXICO	4,121	466	GERMANY	271	23
UK	3,759	204	ITALY	247	178
IRAN	1,652	57	FRANCE	191	50
SOUTH AFRICA	1,895	60	SINGAPORE	68	0.05

and 64 × 64 from 150 CT images. To increase the classification performance, the feature extraction process was performed on each patch. Five computational learning theory algorithms were adopted and utilized as feature extraction methods: a gray level co-occurrence matrix (GLCM), a gray level run length matrix (GLRLM), a local directional pattern (LDP), a discrete wavelet transform (DWT), and a gray-level size zone matrix (GLSZM). To avoid an overfitting problem, k-fold cross validation was performed during training. With GLSZM and 10-fold cross-validation, the classifier achieved the best accuracy (99.68%).

Randhawa *et al.* [47] proposed a method of using computational learning theory for genome analyses. This method combines decision trees with digital signal processing to construct a model for classification of the COVID-19 virus sequences and can identify intrinsic viral genomic signatures. To validate the results of identifications, Spearman's rank correlation coefficient analysis was adopted. The proposed method can be used to analyze large datasets containing more than 5,000 unique viral genomic sequences. In this dataset, there are 29 COVID-19 viral sequences, implying an imbalanced data issue (29: 5000). The proposed method achieved a 100% accuracy. Furthermore, the proposed method uses only raw DNA sequence data to discover the most relevant relationships between more than 5,000 viral genomes within minutes from scratch. This shows that, for new viral and pathogen genome sequences, unmatched genome-wide machine learning methods can provide reliable real-time courses of action for taxonomic classification.

Mei *et al.* [48] developed an ensemble model to identify COVID-19 infections, which can allow early identification of COVID-19 patients at an early stage based on the initial chest CT scans and related clinical information. This model combines a deep convolutional neural network with three classifiers: random forest, support vector machine, and multilayer perceptron. The deep convolutional neural network is utilized for imaging the characteristics of COVID-19 patients, and the three classifiers form an ensemble model to classify COVID-19 patients based on extracted characteristics of COVID-19 and other clinical information. This ensemble model showed significant performance in terms of sensitivity (84.3%), specificity (82.8%), and AUC (0.92).

Apostolopoulos *et al.* [49] extended their previous work [19] by using transfer learning to train deep CNNs since there are many pre-trained models that can be retrieved from open sources, such as VGG-19 [52], MobileNets V2 [30], Inception V4 [53], and Xception [54]. Unlike their previous work [19], which straightforwardly utilized MobileNets V2 to build an image recognition model for classifying COVID-19 patients, Apostolopoulos *et al.* [49] applied transfer learning on the pre-trained models and used a dataset that consists of 224 chest CT images of patients with COVID-19, 700 chest CT images of confirmed common bacterial pneumonia, and 504 chest CT images of no diseases to fine-tune the pre-trained models.

Based on the papers surveyed in this section, some insightful findings can be made:

1) Although deep learning has become the most popular notion recently, some classical computational learning theory approaches, such as support vector machine, random forest, and decision tree, could still be useful while the amount of data is limited. The studies [46]–[48] reveal that the shallow learning method can be utilized as an initial model for building a classification model to distinguish COVID-19 patients.
2) With a bigger dataset, the concept of model ensembles can be used to combine initial models with some deep learning methods. The work in [45] and [47] provides possible solutions for model ensembles. Besides model ensemble, the concept of domain adaptation is a possible solution to combine two models. The transfer learning techniques utilized in Apostolopoulos *et al.*'s study [48] are also a possible solution.

VII. Probabilistic Methods for Combating COVID-19

In computational intelligence, a probabilistic method is applied by calculating the expected value of a random variable. The probabilistic method is typically used for analysis of the risk factors correlated with COVID-19 and explains why they are crucial. Table VI shows the issues that have been addressed by probabilistic methods.

Cássaro and Pires [55] assume that the number of infected patients grows exponentially over the time. As a result, the probabilistic model can be formulated as

$$I(t) = I(t_0)e^{rt}, \tag{3}$$

TABLE V Issues addressed by existing computational learning theory methods.

	TPVP	CSVI	TRD	PD	PHPM
DUFFEY AND ZIO'S STUDY [44]	✓				
WANG *ET AL.*'S STUDY [45]		✓			
BARSTUGAN *ET AL.*'S STUDY [46]		✓		✓	
RANDHAWA *ET AL.*'S STUDY [47]		✓			
MEI *ET AL.*'S STUDY [48]		✓			
APOSTOLOPOULOS *ET AL.*'S STUDY [49]		✓			

TABLE VI Issues addressed by existing probabilistic methods.

	TPVP	CSVI	TRD	PD	PHPM
CÁSSARO AND PIRES'S STUDY [55]	✓				
ZHANG *ET AL.*'S STUDY [57]	✓				
KUCHARSKI *ET AL.*'S STUDY [56]	✓				

where $I(t)$ is the number of diagnosed infections over time, t_0 is initial time, and r is the growth rate, which can be determined through learning from the time-series data of diagnosed infections by minimizing the mean absolute error. Cássaro and Pires [55] utilized the time-series data of diagnosed infections collected from eight countries (Greece, Italy, Spain, Germany, France, Netherlands, the UK, and the USA) to learn the growth rate, r. The prediction results show that the exponential model can accurately predict the number of confirmed cases within 14 days when the first infection is observed; however, it is unable to make long-term predictions.

Zhang *et al.* [57] proposed a logistic growth probabilistic model that considers both the power law and the exponential law to estimate the number of infected patients. Unlike the exponential model that can only deal with the estimation of uncontrolled prevalence, the logistic growth probabilistic model is initially approximated to the exponential law, but the upper bound of the model is set and used to reduce the growth rate. Accordingly, the logistic growth probabilistic model is formulated as

$$I(t) = \frac{N}{1 + e^{b - c(t - t_0)}}, \tag{4}$$

where $I(t)$ is the number of diagnosed infections over time, N is the predicted upper bound, b and c are the fitting coefficients that can be learned from the dataset, and t_0 is the time when the first infection is observed. The prediction results show that the logistic growth probabilistic model can make long-term predictions that have a low prediction error within 3 months.

Kucharski *et al.* [56] designed a stochastic transmission dynamic model to estimate the variation in transmission over time during January and February of 2020. A dataset that consists of the COVID-19 population in or from Wuhan was collected. The transmission was modelled as a geometric random walk. Based on the proposed stochastic transmission dynamic model, the probability of outbreak in other areas was estimated. To train the proposed stochastic transmission dynamic model, the model was fitted into four publicly available datasets: 1) daily numbers of new global confirmed cases beginning January 26, 2020; 2) daily numbers of new confirmed cases in Wuhan between December 1, 2019 and January 1, 2020; 3) daily numbers of new confirmed cases in China between December 29, 2019, and January 23, 2020; and 4) proportions of confirmed cases on evacuation flights between January 29, 2020 and February 4, 2020.

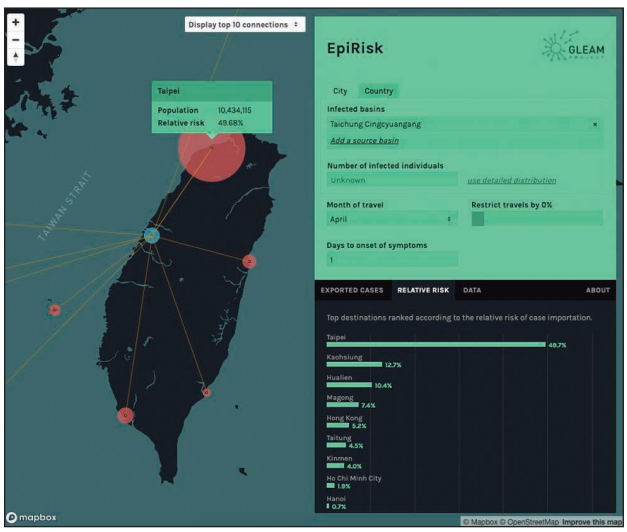


FIGURE 2 A snapshot of EpiRisk.

VIII. Real-World Systems and Tools Using Computational Intelligence for Combating COVID-19

Many industries and nonprofit organizations have been utilizing computational intelligence to develop systems or tools for combating COVID-19. According to the report published by the Organization for Economic Co-operation and Development (OECD) [58], these real-world systems and tools for combating COVID-19 can be utilized to support decision makers, the medical community, and society to manage every stage of the COVID-19 crisis; these stages consist of detection, prevention, response, and recovery. Based on the OECD's report, several AI-powered tools, including BlueDot [59], EpiRisk [60], CRUZR robot [61], Canada's COVID-19 chatbot [62] and Satellites Monitor [63], can be used for combating COVID-19. However, the details of these systems and tools were not stated in the report or other relevant literature. In fact, their results are still worth introducing and promoting to the community of computational intelligence. Instead of introducing how they work, in this section, we focus on what they have done.

BlueDot [59] is a software that evaluates the outbreak risk of infectious diseases caused by over 150 different pathogens, toxins, and syndromes. In fact, COVID-19 is the most crucial disease whose outbreak risk is detected by BlueDot. The main technique behind BlueDot is a crawler that can scan over 100,000 official and mass media sources in 65 languages per day. Based on the data crawled, natural language processing and text mining are applied to extract important information for the evaluation of outbreak risk of infectious diseases. Although BlueDot is recognized by the OECD's report to support decision makers at the detection stage, it provides a user-centric view that can also be utilized to calculate an individual's probability of infection (i.e., issues at the prevention stage). Unlike BlueDot, EpiRisk [60] is a web-based application that calculates an individual's probability of infection based on a topology structure of airline transportation networks. Since the data source of EpiRisk is quite narrow and might miss some crucial data, the evaluation results are doubtful, even if it is very convenient to use. For example, as shown in Fig. 2, EpiRisk shows that the probability of infection in Taipei is higher than 45%, which is completely untrue, with zero new confirmed domestic cases over the past 80 days in Taiwan.

When the outbreak began, how to effectively isolate infection was a key issue. The operating site might be divided into

two parts: hospital and home. To isolate suspected infections in hospitals, many hospitals have been utilizing robots to serve patients that arrive. For example, at Antwerp University Hospital, Cruzr Health [61] takes patients' body temperatures and checks whether they are wearing masks when patients arrive. If the patients are wearing masks and their body temperatures are within the normal range, Cruzr Health leads them to their appointment. To take care of home isolation, Canada developed a chatbot [62] that provides information for home isolation and reminders for those who are suspected to be infected. As shown in Fig. 3, we can see that the chatbot provides information on COVID-19 symptoms at users' requests.

At the end of the pandemic, the most important task for every country would be economic recovery. However, the timing of enforcing policies for economic recovery is a difficult problem. Researchers at WeBank [64], a private Chinese neobank, collected data from satellites, GPS, and social media to detect the hot spots of actual steel manufacturing inside the factories in China [63]. They believe that the detection results may reflect economic recovery in China. Although this system was built for internal use, the data from satellites, GPS, and social media can be crawled easily, implying that computational intelligence researchers can build a model according to this concept.

IX. Conclusions

In this survey, we reviewed several critical issues on combating COVID-19 that have been or can be resolved using computational intelligence techniques. Computational intelligence is classified into five different principles: neural networks, fuzzy logic, evolutionary computation, computational learning theory, and probabilistic methods. Our survey found that most research studies have been designed based on neural networks for addressing the issues on characterization of the symptoms of viral infections. Meanwhile, Panwar *et al.* [16]'s method achieved the highest performance (97.62% true-positive rate), which means that using deep neural networks to detect symptoms from CT images is well-developed, and we may devote our efforts to other issues.

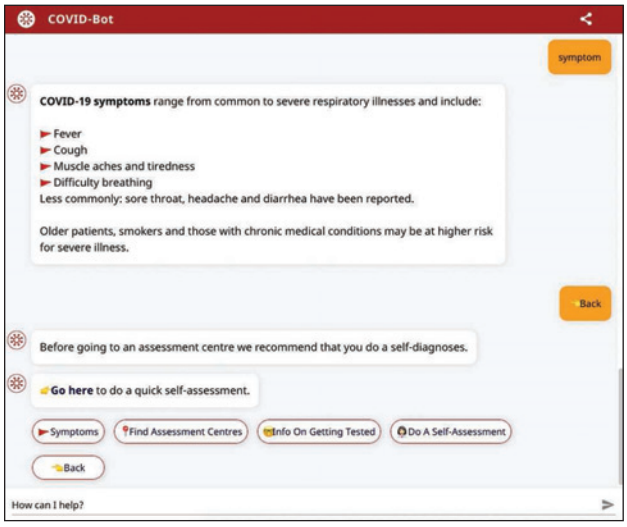


FIGURE 3 A snapshot of Canada's COVID-19 chatbot.

Theoretically, all issues we listed in Section II can be solved by at least one of the principles of computational intelligence. Unfortunately, based on our survey, many COVID-19 pandemic issues have not yet been addressed in computational intelligence studies. On the contrary, most reported studies have focused only on specific issues, such as the characterization of the symptoms of viral infection. This may be because computational intelligence is a data-driven technique that can work well mostly when the amount of data is sufficient. Currently, the data that we can most easily crawl is chest CT images. Therefore, existing works have focused on discovering the characteristics of COVID-19 patients based on their chest CT images to build classification models. As more and more patients are cured, many treatment records will be produced. Such treatment records could be viewed as a set of time series data. Many computational intelligence techniques could be then applied to analyze treatment records. To address the issues on TrD and PD, future works can combine time-series analysis mechanisms with previous works. For example, if we obtain COVID-19 patients' CT images for each stage, the characteristics at each stage can be modeled and utilized for treatment design.

Finally, we observe that some existing works, such as [42], [46]–[48], utilized more than two principles to design hybrid models that can balance the strengths and weaknesses of two principles so that the applicability of these works could be improved. For example, evolutionary computation could be used to optimize the hyperparameters of deep learning models so that some deep learning models might be built from limited data. We believe, in the near future, the computational intelligence community will invent new algorithms by combining multiple principles to address the critical issues described in this survey using limited data or under strict conditions, such as visual analytics techniques and applications for propagation modeling and monitoring, vaccine design or drug repositioning, as well as IoT for smart care in COVID-19.

Acknowledgment

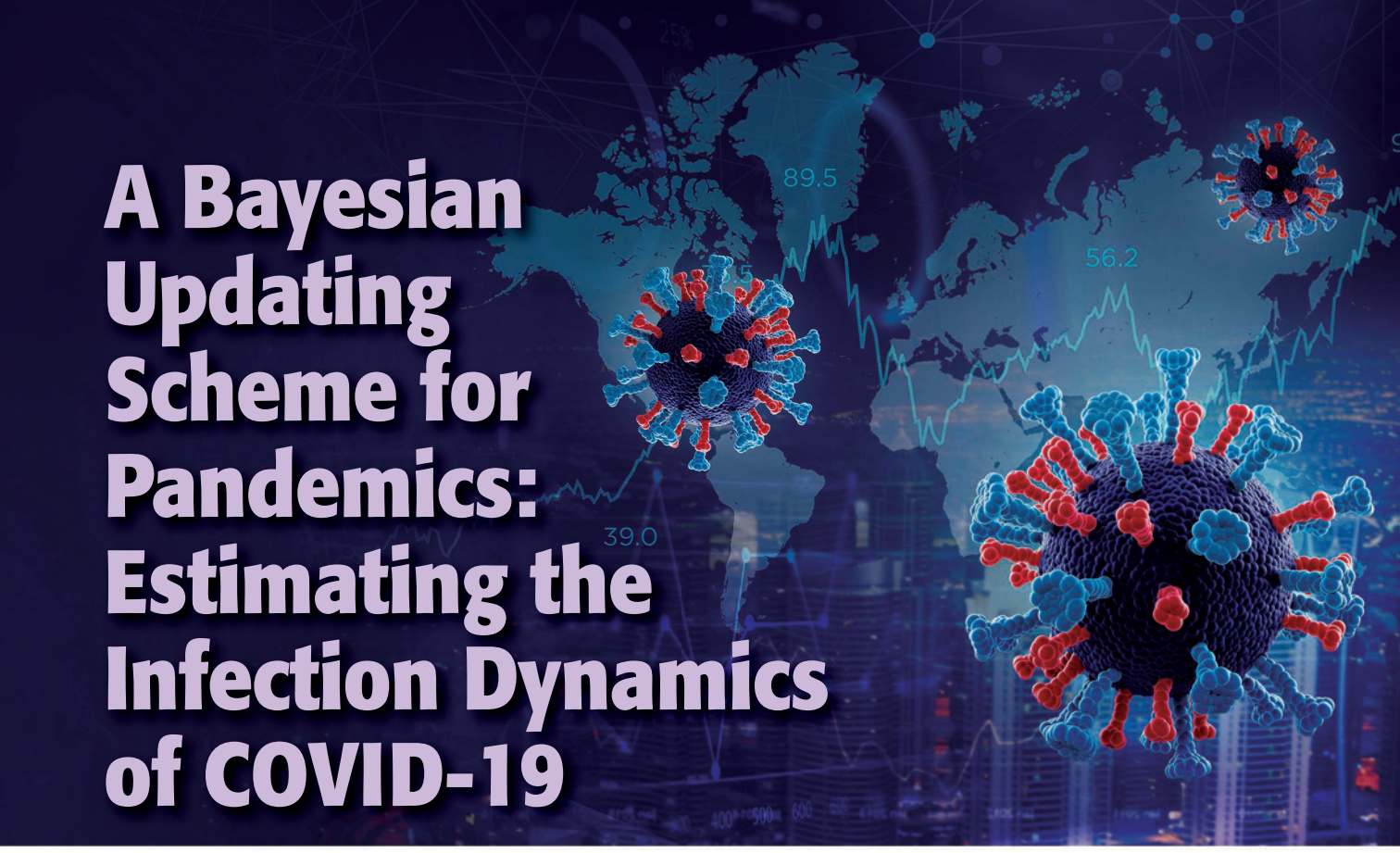
This research was partially supported by Ministry of Science and Technology Taiwan under grant no. MOST 109-2224-E-009-003 and by T.T. and W.F. Chao Foundation and John S Dunn Research Foundation at Houston, Texas, USA.

References

[1] F.S. Lu et al., "Accurate influenza monitoring and forecasting using novel internet data streams: A case study in the Boston metropolis," *JMIR Public Health Surveill.*, vol. 4, no. 1, p. e4, 2018. doi: 10.2196/publichealth.8950.
[2] S. Jang, S. Lee, S.M. Choi, J. Seo, H. Choi, and T. Yoon, "Comparison between SARS CoV and MERS CoV using Apriori algorithm, decision tree, SVM," in *Proc. MATEC Web Conf.*, 2016, vol. 49, p. 08001. doi: 10.1051/matecconf/20164908001.
[3] D. Kim, S. Hong, S. Choi and T. Yoon, "Analysis of transmission route of MERS coronavirus using decision tree and Apriori algorithm," in *Proc. 18th Int. Conf. Advanced Communication Technology (ICACT)*, Pyeongchang, 2016, pp. 559–565.
[4] "COVID-19 Open Research Dataset Challenge (CORD-19)." Accessed: Aug. 3, 2020. [Online]. Available: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
[5] J. C. Bezdek, "What is computational intelligence?" in *Computational Intelligence: Imitating Life*, J. Zurada, R. Marks, and C. Robinson, Eds. Piscataway, NJ: IEEE Press, 1994, pp. 1–12.

[6] N. Siddique, H. Adeli, *Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing*. Hoboken, NJ: Wiley, 2013.
[7] R. Marks, "Intelligence: Computational versus Artificial," *IEEE Trans. Neural Netw.*, vol. 4, no. 5, pp. 737–739, 1993.
[8] C. Stergiou and D. Siganos, "Neural networks," *Imperial Coll. London Surprise J.*, vol 4, p. 1.
[9] M. J. Somers, J. C. Casal, "Using artificial neural networks to model nonlinearity," *Organ. Res. Methods*, vol. 12, no. 3, pp. 403–417, 2009. doi: 10.1177/1094428107309326.
[10] V. Novák, I. Perfilieva, J. Močkoř, *Mathematical Principles of Fuzzy Logic*. Dordrecht: Kluwer, 1999.
[11] K. De Jong, *Evolutionary Computation: A Unified Approach*. Cambridge, MA: MIT Press, 2006.
[12] A. K. Palit and D. Popovic, *Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications*. New York: Springer-Verlag, 2006, p. 4.
[13] F. Gaillard, "Normal chest CT - Lung window." Accessed: June 28, 2020. [Online]. Available: https://radiopaedia.org/cases/normal-chest-ct-lung-window-1?lang=us
[14] S. Roy et al., "Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2676–2687, Aug. 2020. doi: 10.1109/TMI.2020.2994459.
[15] L. Wang and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," 2020, arXiv:2003.09871.
[16] Z. Han et al., "Accurate screening of COVID-19 using attention based deep 3D multiple instance learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2584–2594, Aug. 2020, doi: 10.1109/TMI.2020.2996256.
[17] H. Panwar, P.K. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection of COVID-19 in X-rays using nCOVnet," *Chaos, Solitons, Fractals*, vol. 138, p. 109,944, 2020. doi: 10.1016/j.chaos.2020.109944.
[18] Y. J. Oh, S. J. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020. doi: 10.1109/TMI.2020.2993291.
[19] I. Apostolopoulos, S. Aznaouridis, and M. Tzani, "Extracting possibly representative COVID-19 biomarkers from X-ray images with deep learning approach and image data related to pulmonary diseases," *J Med. Biol. Eng.*, 2020. doi: 10.1007/s40846-020-00529-4.
[20] S. Wang et al., "A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis," *Eur. Respir. J.*, vol. 56, no. 1, p. 2,000,775, 2020. doi: 10.1183/13993003.00775-2020.
[21] S. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalhori, "Predicting COVID-19 incidence through analysis of google trends data in Iran: Data mining and deep learning pilot study," *JMIR Public Health Surveill.*, vol. 6, no. 2, p. e18828, 2020. doi: 10.2196/18828.
[22] S. Vaid, R. Kalantar, and M. Bhandari, "Deep learning COVID-19 detection bias: Accuracy through artificial intelligence," *Int. Orthop.*, vol. 44, pp. 1539–1542, 2020. doi: 10.1007/s00264-020-04609-7.
[23] L. Huang, R. Han, T. Ai, P. Yu, H. Kang, Q. Tao, and L. Xia, "Serial quantitative chest CT assessment of COVID-19: Deep-learning approach," *Radiol., Cardiothorac. Imag.*, vol. 2, no. 2, p. e200075, 2020. doi: 10.1148/ryct.2020200075.
[24] B. Hurt, S. Kligerman, and A. Hsiao, "Deep learning localization of pneumonia: 2019 coronavirus (COVID-19) outbreak," *J. Thoracic Imag.*, vol. 35, no. 3, pp. 87–89, May 2020. doi: 10.1097/RTI.0000000000000512.
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks or biomedical image segmentation," in *Proc. 2015 Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, Munich, Oct. 5–9, 2015, pp. 234–241.
[26] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Proc. Workshop Deep Learning Medical Image Analysis and Multimodal Learning Clinical Decision Support*, Granada, Spain, Sept. 20, 2018, pp. 3–11.
[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. 2018 European Conf. Computer Vision (ECCV)*, Munich, Sept. 8–14, 2018, pp. 801–818.
[28] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition Workshops*, Honolulu, HI, pp. 11–19.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNet V2: Inverted residuals and linear bottlenecks," in *Proc. 2018 IEEE Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, pp. 4510–4520.
[31] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 4700–4708
[32] T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 936–944. doi: 10.1109/CVPR.2017.106.
[33] N. Dhiman, and M. Sharma, "Fuzzy logic inference system for identification and prevention of coronavirus (COVID-19)," *Int. J. Innov. Technol. Exploring Eng.*, vol. 9, pp. 2278–3075, 2020. doi: 10.35940/ijitee.F4642.049620.
[34] D. M. Nazarov, "Fuzzy model of digital assessment of donation systems' level in COVID-19," in *Proc. 2nd Int. Scientific and Practical Conf. "Modern Management Trends and Digital Economy: From Regional Development Global Economic Growth" (MTDE 2020)*, Yekaterinburg, Russia, 2020. doi: 10.2991/aebmr.k.200502.199.
[35] N. Van Tinh, "Forecasting of COVID-19 Confirmed cases in Vietnam using fuzzy time series model combined with particle swarm optimization," *Comput. Res. Progr. Appl. Sci. Eng.*, vol. 6, no. 2, pp. 114–120, 2020.
[36] Z. Yang, X. Li, H. Garg, and M. Qi, "Decision support algorithm for selecting an antivirus mask over COVID-19 pandemic under spherical normal fuzzy environment," *Int. J. Environ. Res. Public Health*, vol. 17, p. 3407, 2020. doi: 10.3390/ijerph17103407.
[37] Z. Ren, H. Liao, and Y. Liu, "Generalized Z-numbers with hesitant fuzzy linguistic information and its application to medicine selection for the patients with mild symptoms of the COVID-19," *Comput. Ind. Eng.*, vol. 145, p. 106,517, July 2020, doi: 10.1016/j.cie.2020.106517.
[38] A. Yousefpour, H. Jahanshahi, and S. Bekiros, "Optimal policies for control of the novel coronavirus COVID-19 outbreak," *Chaos, Solitons Fractals*, p. 109,883, 2020. doi: 10.1016/j.chaos.2020.109883.
[39] M. Niazkar, and H. R. Niazkar, "COVID-19 outbreak: Application of multi-gene genetic programming to country-based prediction models," *Electron. J. General Med.*, vol. 17, no. 5, p. em247, 2020.
[40] R. Salgotra, M. Gandomi, and A. Gandomi, "Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming," *Chaos, Solitons, Fractals*, vol. 138, p. 109,945, 2020. doi: 10.1016/j.chaos.2020.109945.
[41] R. Salgotra, M. Gandomi, and A. Gandomi, "Evolutionary modelling of the COVID-19 pandemic in fifteen most affected countries," *Chaos, Solitons Fractals*, vol. 140, p. 110,118, Nov. 2020. doi: 10.1016/j.chaos.2020.110118.
[42] S. Dilbag, Singh, V. Chahar, Vaishali, and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks," *Eur. J. Clin. Microbiol. Infect. Dis., Official Publ. Eur. Soc. Clin. Microbiol.*, vol. 39, no. 7, pp. 1379–1389, 2020. doi: 10.1007/s10096-020-03901-z.
[43] K. Price, R. M. Storn, and J. A. Lampinen, *Differential Evolution: A Practical Approach to Global Optimization*. New York: Springer-Verlag, 2006.
[44] R. Duffey, and E. Zio, "Analysing recovery from pandemics by learning theory: The case of COVID-19," *IEEE Access*, vol. 8, pp. 110,789–110,795, 2020. doi: 10.1109/ACCESS.2020.3001344.
[45] G. Wang et al., "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, Aug. 2020. doi: 10.1109/TMI.2020.3000314.
[46] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) classification using CT images by machine learning methods," 2020, arXiv:2003.094.
[47] G. S. Randhawa, M. P. M. Soltysiak, H. El Roz, C. P. E. de Souza, K. A. Hill, and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study," *PLoS ONE*, vol 15, no. 4, pp. 2653–2663, 2020. doi: 10.1371/journal.pone.0232391.
[48] X. Mei et al., "Artificial intelligence–enabled rapid diagnosis of patients with COVID-19," *Nature Med.*, pp. 1–5, 2020.
[49] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, pp. 635–640, 2020. doi: 10.1007/s13246-020-00865-4.
[50] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 31st Conf. Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, pp. 1195–1204.
[51] L. Yu, S. Wang, X. Li, C. W. Fu, and P. A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. 22nd Int. Conf. Medical Image Computing and Computer Assisted Intervention*, Shenzhen, China, 2019, pp. 605–613.
[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv:1409.1556.
[53] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-V4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artificial Intelligence*, San Francisco, 2017, pp. 4278–4284.
[54] F. Chollet, "Xception: Deep learning with depth-wise separable convolutions," in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 1251–1258. doi: 10.1109/CVPR.2017.195.
[55] F. A. Cássaro and L. F. Pires, "Can we predict the occurrence of COVID-19 cases? Considerations using a simple model of growth," *Sci. Total Environ.*, vol. 728, p. 138,834, 2020. doi: 10.1016/j.scitotenv.2020.138834.
[56] A. J. Kucharski et al., "Early dynamics of transmission and control of COVID-19: A mathematical modelling study," *Lancet Infect. Dis.*, vol. 20, no. 5, pp. 553–558, 2020. doi: 10.1016/S1473-3099(20)30144-4.
[57] X. Zhang, R. Ma, and L. Wang, "Predicting turning point, duration and attack rate of covid-19 outbreaks in major western countries," *Chaos, Solitons, Fractals*, vol. 135, p. 109,829, 2020.
[58] "Using artificial intelligence to help combat COVID-19." Accessed: June 28, 2020. [Online]. Available: https://www.oecd.org/coronavirus/policy-responses/using-artificial-intelligence-to-help-combat-covid-19-ae4c5c21/
[59] BlueDot. Accessed: June 28, 2020. [Online]. Available: https://bluedot.global/
[60] EpiRisk. Accessed: June 28, 2020. [Online]. Available: https://epirisk.net/
[61] J. Parrock, "Coronavirus: Belgium hospital employs robot to protect against COVID-19," *Euronews*, Accessed: June 28, 2020. [Online]. Available: https://www.euronews.com/2020/06/02/coronavirus-belgium-hospital-employs-robot-to-protect-against-covid-19
[62] "Canada's COVID-19 Chatbot." Accessed: June 28, 2020. [Online]. Available: https://covidchatbot.com/
[63] T. S. Perry, "Satellites and AI monitor Chinese economy's reaction to coronavirus." Accessed: June 28, 2020. [Online]. Available: https://spectrum.ieee.org/view-from-the-valley/artificial-intelligence/machine-learning/satellites-and-ai-monitor-chinese-economys-reaction-to-coronavirus/
[64] WeBank. Accessed: June 28, 2020. [Online]. Available: https://www.webank.com/#/home

A Bayesian Updating Scheme for Pandemics: Estimating the Infection Dynamics of COVID-19



©SHUTTERSTOCK/MILA SUPINSKAYA GLASHCHENKO

Shuo Wang, Philip Nadler, and Rossella Arcucci
Imperial College London, UK

Xian Yang
Hong Kong Baptist University, China and Imperial College London, UK

Ling Li
University of Kent, UK

Yuan Huang
University of Cambridge, UK

Zhongzhao Teng
University of Cambridge, UK

Yike Guo
Hong Kong Baptist University, China and Imperial College London, UK

***Abstract*—**Epidemic models play a key role in understanding and responding to the emerging COVID-19 pandemic. Widely used compartmental models are static and are of limited use to evaluate intervention strategies of combatting the pandemic. Applying the technology of data assimilation, we propose a Bayesian updating approach for estimating epidemiological parameters using observable information to assess the impacts of different intervention strategies. We adopt a concise renewal model and propose new parameters by disentangling the reduction of instantaneous reproduction number R_t into mitigation and suppression factors to quantify intervention impacts at a finer granularity. A data assimilation framework is developed to estimate these parameters including constructing an observation function and developing a Bayesian updating scheme. A statistical analysis framework is built to quantify the impacts of intervention strategies by monitoring the evolution of the estimated parameters. We reveal the intervention impacts in European countries and Wuhan and the resurgence risk in the United States.

I. Introduction

In response to the COVID-19 pandemic, governments have taken non-pharmaceutical intervention measures. Common measures include travel restriction, school and non-essential business closure and social distancing, as well as early isolation of confirmed patients. Recently, as the first-wave epidemic peak faded away in many countries, the accumulated observations of epidemic growth [1] and corresponding intervention policies [2] shed more insights on how the interventions worked. Meanwhile, many governments have switched into the phase to reopen economic and social activities with attention on tamping down possible resurgences. However, recent second-wave outbreaks in some countries

Digital Object Identifier 10.1109/MCI.2020.3019874
Date of current version: 14 October 2020

Corresponding author: Yike Guo (email: yikeguo@hkbu.edu.hk).
Shuo Wang and Xian Yang had equal contribution.

and regions (e.g. the United States, Hong Kong) alert us to monitor the epidemic evolution carefully while intervention measures are being relaxed.

Mathematical models play a key role in understanding and responding to the emerging COVID-19 pandemic [3]–[5]. Compartmental models (e.g. SIR, SIER) and time-since-infection models (i.e. renewal process-based models) are the two well-known approaches describing the underlying transmission dynamics [6], [7]. The compartmental models describe the transmission among sub-populations while the renewal process-based approach starts from the inter-individual transmission. Despite different nomenclatures and applications, each model contains parameters characterizing the epidemic dynamics. One of the most well-known parameters is the reproduction number R, which represents the average number of secondary cases that would be induced by an infected primary case [8]. This key parameter is related to the final epidemic size of an infectious disease [9]. Intervention measures aim to maintain the reproduction number under one so that the epidemic can be contained along with time. Thus, the estimation of time-varying R will reflect the impacts of an intervention.

The basic reproduction number R_0 is the reproduction number at the beginning of an epidemic outbreak, when the susceptible population is approximately infinite and without intervention measures. When various intervention measures are being introduced, the instantaneous reproduction number R_t (also called effective reproduction number) is of greater interest. To gain insights into epidemic evolution, most existing studies such as [3], [10] focus on estimating time-varying instantaneous reproduction number R_t.

However, the nowcasting of R_t from reported data is not an easy task. Several approaches have been proposed to estimate R_t with different advantages [11]–[13], but the timeliness and accuracy are still of concern. Nowcasting results are affected by different factors, such as assumptions of the epidemic models, statistical inference methods and uncertainty of data resources. Inappropriate interpretation or imprecise estimation of R_t are criticized for providing misleading information [14]. For example, the nowcasting from reported confirmed cases will fall behind the nowcasting from onset data because there is a delay from symptom onset to case report. We hypothesize that more detailed characteristics of the time-varying infectiousness profile could be estimated from the publicly available reports (e.g., death data, confirmed data, onset data and laboratory data) and help to better understand and evaluate the efficiency of interventions.

In this study, we propose a comprehensive Bayesian updating scheme for reliable and timely estimation of parameters in epidemic models. The transmission dynamics are modeled as a concise renewal process with time-varying parameters. To monitor the evolving impacts, more fine-grained modeling of the transmission dynamics is required. Instead of the well-known R_t, we introduce two complementary parameters: the mitigation factor (p_t) captures the effect of shielding susceptible population (e.g. through social distancing), and the suppression factor (D_t) captures the effect of isolating the infected population (e.g. through quarantine) to stop virus transmission. We propose a novel method to estimate these parameters by taking the data assimilation approach with Bayesian updating methods. We use daily reports of confirmed cases as the observation. A deconvolution method is used to build an observation function to estimate the infection cases by taking into account the incubation time and report delay. The evolution of the time-varying infectiousness profile (i.e. p_t and D_t) is estimated from the adjusted epidemic curve through a Bayesian approach of data assimilation. Such a fine-grained infectiousness profile enables us to quantify the impacts of various intervention measures in a comprehensive way.

The paper is structured as follows: We introduce the related work in Section II. In III, we present the overview of a time-varying renewal process model where the two parameters p_t and D_t are proposed. In IV, we present in detail the Bayesian updating scheme for estimating the dynamic parameters. In V, we develop a statistical analysis method of assessing the intervention impacts based on the estimated results and the report of intervention policies. In VI, as applications of our approach, we investigate the impacts of intervention measures in European countries, the United States and Wuhan to illustrate the importance of this development.

II. Related Work

At the beginning of the COVID-19 outbreak in Wuhan, China, compartmental models (e.g. SIR, SEIR model) have been used to investigate the epidemic dynamics [15]–[17], where the basic reproductive number was estimated from the models with static parameters. With the spread of COVID-19 worldwide, renewal process-based models (i.e. time-since-infection model) are also widely used in the study of COVID-19. The R package 'EpiEstim' [11], [12] is the most widely used in estimating the time-varying R_t with a sliding window. In [10], 'EpiEstim' was applied to infer R_t via the discrete renewal process for policy impact assessment. Similar work has been done in [3] to infer R_t using 'EpiEstim' from laboratory-confirmed cases in Wuhan and hence evaluated the impact of non-pharmaceutical public health interventions. The work in [18] has pointed out that the infection data is usually not available and death data were used as observation for R_t estimation. Instead of simply applying 'EpiEstim' to reported data, they estimated R_t by employing the renewal equation as a latent process to model infections and connecting the infections to death data via a generative mechanism. However, the estimated R_t is in a piecewise form and the number of changing points was assumed to be determined by the imposed interventions. [19] estimates R_t from the death data as well while linking the disease transmissibility to mobility using the renewal equation. In general, [18] and [19] explicitly

formulated the R_t's updating function by introducing external factors (e.g. interventions and mobility). Thus, the estimated R_t curve is largely constrained by the factors that are considered in the model.

Data Assimilation [20] lends itself naturally to this problem since it provides a framework to enable dynamically updating the model states and parameters when new observations become available while also taking into account model and observation uncertainty. Data assimilation technologies, such as Kalman filter and variational method [21], have been widely used in signal tracking, oceanology, environment monitoring and weather forecasting where physical models and observation data are assimilated to produce accurate predictions. Data assimilation for epidemiological modeling was first proposed in [22] where compartment models were used as the underlying model for assimilation. In [25] and [26], estimating time-varying parameters in the compartment models was further investigated. To the authors' best knowledge, our work is the first study to apply data assimilation to the renewal process-based model.

III. Epidemic Modeling of COVID-19 Transmission

In this section, we propose a time-varying renewal process with two complementary parameters p_t and D_t to model the evolving infectiousness profile. We adopted a time-varying renewal process for epidemic modeling. The renewal process [8] of infectious disease transmission is:

$$I(t) = \int_0^{\infty} I(t-\tau)\beta(\tau)d\tau \tag{1}$$

where $I(t)$ is the incident infection on time t and $\beta(\tau)$ is the infectiousness profile. The infectiousness profile means that a primary case infected τ time ago (i.e. with the infection-age τ) can now generate new secondary cases at a rate of $\beta(\tau)$, describing a homogenous mixing process. $\beta(\tau)$ is related to biological, behavioral and environmental factors. We can calculate the reproduction number R as the area under the curve of $\beta(\tau)$, which is the overall number of secondary cases infected by a primary case. Further, $\beta(\tau)$ can be rewritten as:

$$\beta(\tau) = R \cdot w(\tau) \tag{2}$$

where the unit-normalized transmission rate $w(\tau)$ is the probability density function of generation time, i.e. the interval between the primary infection and the secondary infection. In the early stage without intervention, the infectiousness profile remains time-independent as the baseline $\beta_0(\tau)$ which describes the transmission dynamics when the susceptible population is infinite. The corresponding R is the well-known basic reproduction number R_0. In reality, the infectiousness profile $\beta(\tau)$ will evolve with time t, therefore we introduce $\beta_t(\tau)$ to address the change in its distribution caused by intervention measures.

To quantify the impacts of intervention measures to the evolution of R_t, we propose two factors: **mitigation** and **suppression** to disentangle the intervention effects. As illustrated in Figure 1, we uste two complementary metrics p_t and D_t to model these factors, respectively.

The suppression effects mainly shorten the infectious period of the infected population, corresponding to the truncation of $\beta(\tau)$ along the horizontal axis. We use a time-varying parameter D_t to denote the effective infectious window induced by suppression. The mitigation effects attenuate the overall infectiousness by shielding the susceptible population, corresponding to the scaling in the vertical direction. We introduce another time-varying parameter p_t to describe this attenuation effect induced by mitigation. Formally, we parameterize the evolution of the infectiousness profile as:

$$\beta_t(\tau) = \begin{cases} \beta_0(\tau) \cdot p_t & \tau < D_t \\ 0 & \tau \geq D_t \end{cases} \tag{3}$$

Accordingly, the instantaneous reproductive number R_t can be derived:

$$R_t = p_t \cdot \int_0^{D_t} \beta_0(\tau)d\tau \tag{4}$$

Therefore, the impact of intervention measures on R_t reduction is disentangled: mitigation factor p_t attenuates the overall infectiousness through shielding the susceptible population, and suppression factor D_t shortens the infectious period

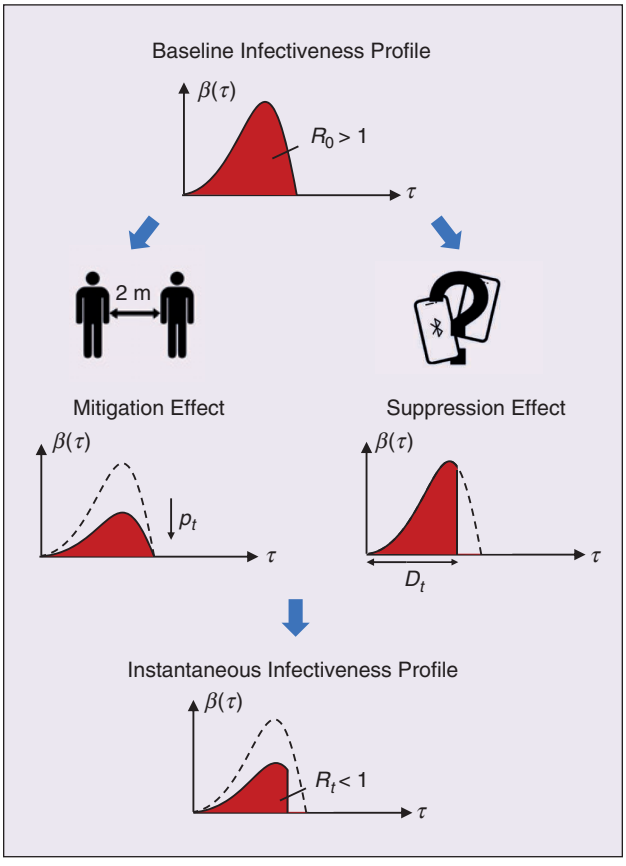


FIGURE 1 Disentangling the reduction of reproduction number into mitigation and suppression factors.

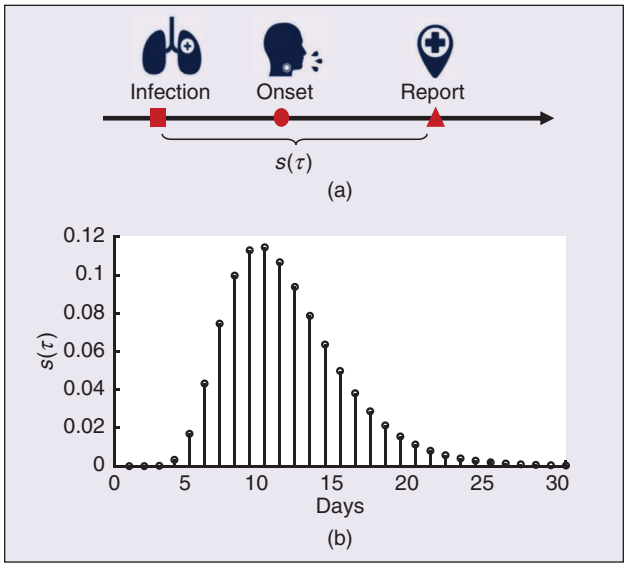


FIGURE 2 Reconstruction of daily infection from the confirmed cases using deconvolution algorithms. The time delay between the infection and onset and report is demonstrated (a). The estimated distribution between infection and report is presented which is used for deconvolution (b).

through isolating the infected population. It is noted that the R_t can be derived from p_t and D_t, both of which provide more mechanistic details about the evolution of the infectiousness profile.

IV. Adaptive Parameter Estimation

We aim to develop a comprehensive framework to estimate parameters of renewal process models using the Bayesian updating approach of data assimilation, especially three key parameters: $<R_t,\ p_t,\ D_t>$. The impacts of different interventions can be quantified through monitoring the evolution of $<R_t,\ p_t,\ D_t>$. This framework contains all components of a data assimilation system: building an observation function to map observations to model state, modeling and Bayesian updating as shown in Figures 2 and 3. By applying the observation function, we reconstruct the number of daily infections from reports of confirmed cases, taking into account the incubation time and report delay with a deconvolution algorithm. Then $<R_t,\ p_t,\ D_t>$ is estimated through a Bayesian updating approach of data assimilation.

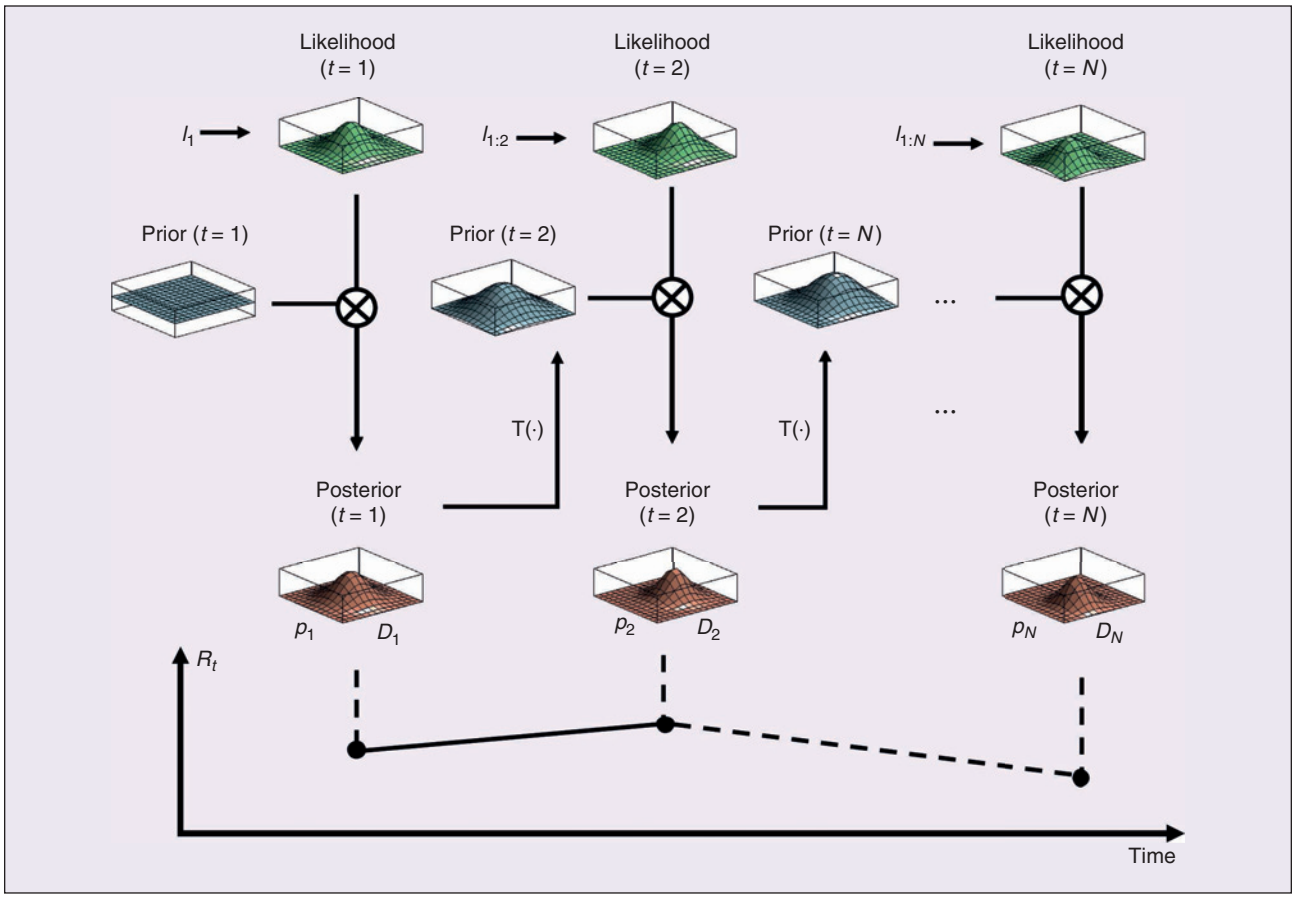


FIGURE 3 Illustration of the Bayesian updating framework for estimating suppression and mitigation factors. We employ a two-level hierarchical model: For each time step, the low-level model (i.e. renewal process) provides the likelihood of p_t, D_t (green). The posterior (orange) is calculated through the element product of the likelihood and the prior (blue) from the previous time step. To generate the prior for next time step, we use the high-level model (i.e. the transformation T) to induce the evolution of parameters. The high-level model is a piecewise gaussian random walk process where the fluctuations of p_t and D_t differ before and after an intervention time. The instantaneous reproduction number R_t can be derived from the posterior distribution of p_t and D_t.

A. Reconstruction of Daily Infection from Reported Cases

Data assimilation updates model states and parameters using new observation data. It is important for parameter estimation that proper observation is chosen, and an observation function can be built to map observations to a state variable (usually regarded as the output of the model).

In our study, observations are obtained from the reported number of confirmed cases. The model output is the daily infection incidence through the renewal process. However, such observations experience an inevitable time delay between the actual infection time and the reporting date (Figure 2A). This includes an incubation time (i.e. the period between infection and onset of symptoms) and confirmation period (i.e. the period between onset and officially reported after being tested). The confirmed cases reported on time t were actually infected within a past period and the reported number is the convolution result of the historical daily infection numbers (Figure 2B).

Here, we define an observation function to reconstruct the daily infection instances from the confirmed cases using the deconvolution technique with the Richardson-Lucy (RL) iteration method [25]. We use the incubation period calculated by Ferretti et al. [5], which is a lognormal distribution with a mean of 5.5 days and a standard deviation of 2.1 days. We use the confirmation period previously reported by Leung et al. [10], which is a gamma distribution with a mean of 4.9 days and a standard deviation of 3.3 days. Sampling from these two sequential distributions, we estimated the discrete interval distribution $s(\tau)$ for $\tau \in \{0, d\}$ from infection to report (Figure 2). Denoting the epidemic curve of reported infection cases $\hat{I}_{1:t} = \{\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_t\}$ and the epidemic curve of confirmed cases $C_{1:t} = \{C_1, C_2, \ldots, C_t\}$, the reported infection with an observation process of past infections is modeled as a Poisson process:

$$C_t \sim Poisson\left(mean = \sum_{k<t} s(t-k)\hat{I}_k\right) \tag{5}$$

To estimate the daily reported infection curve $\hat{I}_{1:t}$ given the daily confirmed cases curve $C_{1:t}$ and infection-to-confirmed time distribution $s_{1:d}$ is an ill-posed deconvolution problem and can be solved using the Richardson-Lucy (RL) iteration method [25]. The initial guess $\hat{I}^0_{1:t}$ is obtained from the curve of the confirmed case $C_{1:t}$ shifted back by the mode of the infection-to-confirmed time distribution. Let $\hat{C}^n_t = \Sigma_{k<t} s(t-k)\hat{I}^n_k$ be the expected number of confirmed cases on day t of iteration n, and q_t be the probability that a reported case resulting from infection on day t will be observed as defined in [25]. Then the iteration of $\hat{I}_t$ is computed by an expectation-maximization (EM) algorithm as:

$$\hat{I}^{n+1}_t = \frac{\hat{I}^n_t}{q_t} \sum_{i>t} \frac{s(i-t)C_i}{\hat{C}^n_i} \tag{6}$$

A normalized χ^2 statistics is used as the stop criterion of the iteration:

$$\chi^2 = \frac{1}{N} \sum_{t=1}^{N} \frac{(\hat{C}^n_t - C_t)}{\hat{C}^n_t} < 1 \tag{7}$$

where N is the number of days being considered. It is of note that the reported number of confirmed cases constitute the lower bound of the real infection due to the lack of mass test and the existence of asymptomatic cases. However, as long as the detection rate remains consistent, the scaling of reconstructed data does not affect the following inference of transmission dynamics.

B. Bayesian Updating for Parameter Estimation

Following the Bayesian updating approach of data assimilation, we propose an instantaneous estimation method. For the defined epidemiology renewal process, the daily incident infection I_t is the state variable and can be assimilated from the reconstructed infection data from observation. The evolution of the state I_t is governed by the renewal process with the time-varying infectiousness profile $\beta_t(\tau)$, parameterized with p_t and D_t. Here we present a Bayesian framework to monitor the evolution of p_t and D_t using the daily reports of confirmed cases (Figure 3).

Our updating scheme employs a two-level hierarchical model for the inference of time-varying parameters [26]. Let us denote the observed daily incidence of infection till time step t as $\hat{I}_{1:t} = \{\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_t\}$. Suppose $\mathrm{p}(\boldsymbol{\theta}_{t-1} \mid \hat{I}_{1:t-1})$ is the estimated distribution of $\boldsymbol{\theta} = [p, D]^T$ at time step $t-1$. Under the assumption of consistent detection rates, the observed daily incidence $\hat{I}_t$ also satisfies the renewal process. The low-level model predicts the observation (i.e. reconstructed daily infection) given a parameter set through the renewal process:

$$p(\hat{I}_t \mid \boldsymbol{\theta}_t, \hat{I}_{1:t-1}) \sim Poisson\left(mean = \sum_{k=1}^{t-1} \beta_t(k; \boldsymbol{\theta}_t)\hat{I}_{t-k}\right) \tag{8}$$

where a Poisson process of observing the infected cases is assumed. This describes the likelihood of observing the new incidence data given history observations and parameter value $\boldsymbol{\theta}_t$. The high-level model describes the evolution of the model parameters p_t and D_t through transforming the joint distribution:

$$p(\boldsymbol{\theta}_t \mid \hat{I}_{1:t-1}) = \mathrm{T} \circ \mathrm{p}(\boldsymbol{\theta}_{t-1} \mid \hat{I}_{1:t-1}) \tag{9}$$

where $\mathrm{T}(\cdot)$ is a transformation function defining the temporal variations of the $\boldsymbol{\theta}$. The prior knowledge of parameter distribution is transferred to the next time step t by the high-level model p_t. Under the scenario without interventions, the parameters p_t and D_t fluctuate around the baseline values. Therefore, we can assume a random walk of $\boldsymbol{\theta}$ in the parameter space as the high-level model. The joint parameter distribution is updated by convoluting with a Gaussian kernel with variance σ_1. When the intervention is introduced on time d, the random walk of $\boldsymbol{\theta}$ is altered where the variance of the Gaussian kernel will become σ_2. The transformation $\mathrm{T}(\cdot)$ is defined as:

$$\mathrm{T} \circ \mathrm{p}(\boldsymbol{\theta}) = \begin{cases} \mathrm{p}(\boldsymbol{\theta}) * \mathrm{K}_{\sigma_1}(\boldsymbol{\theta}) & t < d \\ \mathrm{p}(\boldsymbol{\theta}) * \mathrm{K}_{\sigma_2}(\boldsymbol{\theta}) & t \geq d \end{cases} \tag{10}$$

where $K_{\sigma_1}(\boldsymbol{\theta})$ and $K_{\sigma_2}(\boldsymbol{\theta})$ are the Gaussian kernels before and after the deployment of intervention at time d. This high-level model includes three hyperparameters: variances before and after intervention: σ_1 and σ_2, and the change-point time d. Let us denote the hyperparameters $\boldsymbol{\eta} = [\sigma_1, \sigma_2, d]^T$. After the latest observation $\hat{I}_t$, the posterior estimation of $\boldsymbol{\theta}$ is updated by the Bayes rule:

$$\mathrm{p}(\boldsymbol{\theta}_t \mid \hat{I}_{1:t}) = \frac{\mathrm{T} \circ \mathrm{p}(\boldsymbol{\theta}_{t-1} \mid \hat{I}_{1:t-1}) \cdot \mathrm{p}(\hat{I}_t \mid \boldsymbol{\theta}_t, \hat{I}_{1:t-1})}{\mathrm{p}(\hat{I}_t \mid \hat{I}_{1:t-1})} \tag{11}$$

This step reflects the Bayesian principle in the key updating step in Kalman filtering [21]. Unlike the Kalman filtering method where uncertainty is explicitly modeled through a covariance matrix under the Gaussian assumption, we directly use posterior probability to capture the uncertainty of estimation. The posterior is usually intractable but can be approximated through grid-based methods. Given a set of hyperparameters $\boldsymbol{\eta}_i$, the hybrid model evidence can be calculated as [26]:

$$\mathrm{p}(\hat{I}_{1:t} \mid \boldsymbol{\eta}_i) = \int p(\hat{I}_{1:t}, \boldsymbol{\theta}_t \mid \boldsymbol{\eta}_i) d\boldsymbol{\theta}_t \tag{12}$$

Finally, the posterior estimation $\mathrm{p}(\boldsymbol{\theta}_t \mid \hat{I}_{1:t})$ can be averaged across the hyperparameter grids weighted by the hybrid model evidence. The posterior mean and confidence intervals of p_t and D_t as well as the corresponding R_t are obtained in a dynamic manner. The prior of R_0 at the first timestep is set uninformative as a uniform distribution with the pre-set lower and upper limits (e.g., the upper limit for the European countries is set to 8 in the experiment). The shape of $\beta_0(\tau)$ is adapted from the distribution of generation time interval $w(\tau)$ reported by Ferretti et al. [5]. We applied the above framework to infer the epidemic evolution in 14 European countries, states in the US and Wuhan city, China in Section VI.

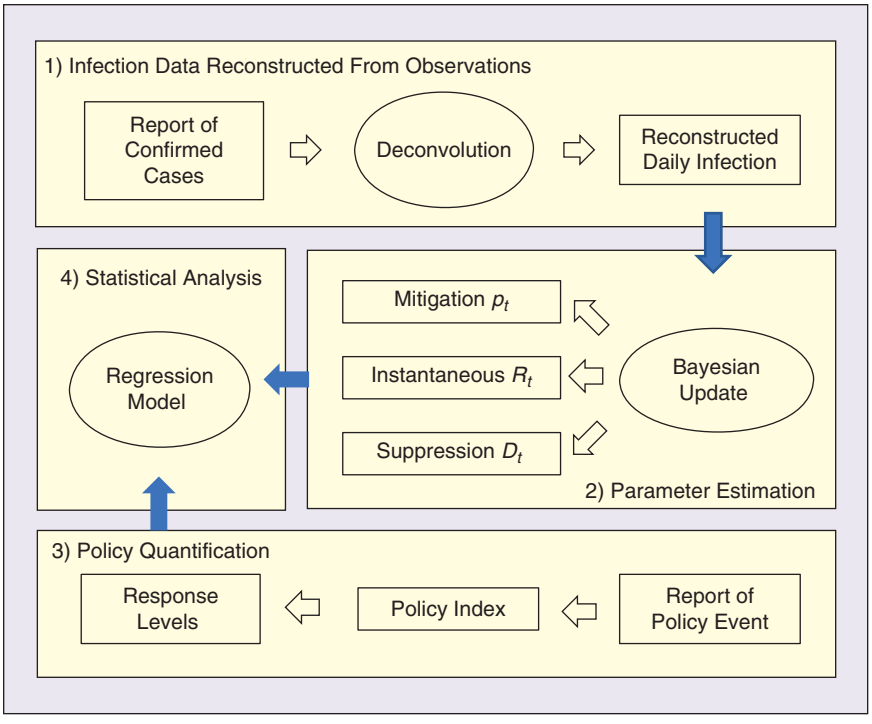


FIGURE 4 Components of the quantification framework. The evolution of mitigation and suppression factors are estimated using the infection data reconstructed from the daily reported confirmed cases. Given the history of government responses, the impacts of intervention measures are quantified by correlating the inferred epidemic parameters to response levels.

V. Evaluation of Intervention Measures

With the estimated results from the above Bayesian updating scheme, now we can perform statistical analysis between the evolution of the transmission dynamics and the implementation of intervention measures. The whole framework containing data reconstruction, dynamic modeling, Bayesian updating, and statistical analysis is presented in Figure 4. In this section, we introduce the quantification of intervention measures and the statistical method.

A. Data Source

For the observations, we use the aggregated data of publicly available daily confirmed cases of 14 Europe countries (Austria, Belgium, Denmark, France, Germany, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the United Kingdom) and 52 states of the United States from John Hopkins University database [1]. The data include the time series of confirmed cases from 22nd January to 8th June 2020 (accessed on 9th June 2020). Six states with accumulated confirmed cases less than 1,000 are excluded from the analysis. The daily number of onset patients in Wuhan is adopted from the retrospective study by Pan et al. [3].

The data of intervention measures in European countries are collected from the Oxford Coronavirus Government Response Tracker [2], reporting the overall stringency index S_t of intervention measures during the analysis period (accessed on 9th June 2020). This overall stringency index is calculated based on the policy quantification of eight intervention measures (i.e. school closing, workplace closing, cancel public events, restrictions on gatherings, close public transport, stay-at-home requirements, restrictions on internal movement and international travel controls) and one health measure (i.e. public info campaigns) to indicate the government response level of intervention.

According to the normalized stringency index by Oxford report [2], we categorized the dates into five response levels (Level 0: $S_t \leq 20\%$, minimal response for reference; Level 1: $20\% < S_t \leq 40\%$, soft response; Level 2: 40%

$< S_t \leq 60\%$, strong response; Level 3: $60\% < S_t \leq 80\%$ and Level 4: $80\% < S_t \leq 100\%$, emergent responses). The representative intervention measures for each response level were identified based on the contribution to the stringency index S_t.

B. Calculation of Intervention Policy Indices

To identify the representative measures of each response level, we calculate the quantification indices of the eight intervention measures. Descriptions of the eight intervention measures and the quantification methods are provided in [2]. For each intervention measure, the Oxford report provides an ordinal scale quantification $v_{j,t}$ of the strength of j-th policy implementation and a binary flag $f_{j,t}$ representing whether it is implemented in the whole country at time t. Following similar practice use in the Oxford report, we normalize the implementation of each intervention measure as

$$P_{j,t} = \frac{\max(0,\ v_{j,t} + 0.5 f_{j,t} - 0.5)}{N_j} \times 100\% \tag{13}$$

where N_j is the maximum value of the indicator P_j. To assign a label of response level to each measure, we calculate the change of mean policy indices across different response levels. The response level with the largest increase is considered as the level that the measure belongs to (i.e. the measure is a representative measure of this response level). For example, the mean index of school closure showed the largest increase from Level 0 to Level 1, so we consider this is a representative measure of Level 1. The representative measures of each response level are listed in Table 1.

C. Regression Analysis of the Intervention Impacts

We performed a retrospective analysis of the time-varying transmission dynamics during different response levels in European countries. First, the evolution history of R_t and the overall stringency index S_t are obtained using the above framework. The stringency index S_t is categorized into five response levels. We fit a log-linear mixed-effect model, where the logarithm of R_t is the outcome variable and the categorical stringency index is the predictor. The logarithm is used to obtain the intervention impacts on the relative change of R_t [27]. We performed a partial-pool analysis by assuming the impacts of intervention measure (slopes) shared across all selected European countries while the basic reproduction number R_0 (intercept) varies due to environmental and social factors. The regression formula is written as:

$$\ln R_{j,t} = b_0 + \sum_{k=1}^{4} b_k * D_{j,k} + \gamma_j + \epsilon \quad j = 1, 2, \ldots, 14 \tag{14}$$

where $R_{j,t}$ is the estimated reproduction number of j-th country, b_0 is the fixed effect term of $\ln R_0$ and b_k is the fixed effects of interventions in response level k. $D_{j,k}$ is the dummy variable that takes the value 1 if and only if the response status is at Level k. γ_j is the random effect term following zero-mean Gaussian which explains the difference of $\ln R_0$ across countries, and ϵ is the Gaussian error term. Equation 14 associates the relative changes in R to the fixed effects of response levels, and can be rewritten into its marginal form as:

$$\ln\left(1 + \frac{R - R_0}{R_0}\right) = \sum_{k=1}^{4} b_k * D_k \tag{15}$$

Therefore, the relative change of R due to the intervention measures in k-th response level can be derived from b_k (i.e. $\Delta R/R_0 = \exp(b_k) - 1$). Country-specific $\ln R_0$ can be estimated as $b_0 + \gamma_j$ at the Level 0. The statistical analysis is performed using the R package 'lme4.' The fixed effect is considered significant with P value < 0.05. The 95% confidence intervals (CI) are estimated using the bootstrap method. The assumption of normality is checked by inspecting the quantile-quantile plot of the residuals. The same procedure is also applied to the analysis of D_t and p_t to quantify the

TABLE 1 The relative reduction of mitigation factor and suppression factor under different response levels of 14 European countries.

RESPONSE	REPRESENTATIVE MEASURES	IMPACT OF MEASURES R_t RELATIVE REDUCTION	SUPPRESSION EFFECT D_t RELATIVE REDUCTION	MITIGATION EFFECT P_t RELATIVE REDUCTION
LEVEL 0 MINIMAL RESPONSE	NO MANDATORY RESTRICTIONS	0	0	0
LEVEL 1 SOFT RESPONSE	CLOSING SCHOOLS, INTERNATIONAL TRAVEL CONTROLS.	35% CI: [25%, 45%]	22% CI: [17%, 27%]	29% CI: [18%, 38%]
LEVEL 2 STRONG RESPONSE	CANCEL PUBLIC EVENTS, RESTRICTIONS ON GATHERING, RESTRICTIONS ON INTERNAL MOVEMENT.	60% CI: [54%, 65%]	26% CI: [21%, 30%]	56% CI: [50%, 61%]
LEVEL 3	CLOSE WORKPLACE, CLOSE PUBLIC TRANSPORT,	71% CI: [68%, 74%]	37% CI: [35%, 40%]	67% CI: [64%, 70%]
LEVEL 4 EMERGENT RESPONSE	STAY-AT-HOME REQUIREMENTS.	74% CI: [71%, 77%]	40% CI: [37%, 42%]	70% CI: [66%, 73%]

We observed that the ground-truth R_t is well estimated within our confidence interval. In particular, the sharp change of R_t caused by the intervention is captured immediately by our approach while there is a lag using the sliding window-based method.

suppression and mitigation factors, respectively. The results are demonstrated in Table 1.

VI. Results

A. Validation with Simulated Data

We simulated an artificial epidemic outbreak with a time-varying infectiousness profile using the renewal process. The generation time intervals were adapted from Ferretti et al. [5]. The simulation period includes 50 days, and an intensive intervention measure is induced on day 35 altering the transmission dynamics. Before the intervention, the ground-truth R_t followed Gaussian random walk with a mean of 2.5. After the intervention (50% p_t reduction and 67% D_t reduction), the mean of R_t was reduced to 0.5 (black line).

We validate the effectiveness of our approach in capturing the sudden change of R_t evolution induced by interventions, which is difficult to detect using traditional sliding window-based methods (Figure 5). We compared the results using our approach (red line with 95% confidence intervals) to the results computed by the R package 'EpiEstim v2.2' [11] (blue) which is a sliding window-based method widely used for R_t estimation. We observed that the ground-truth R_t is well estimated within our confidence interval. In particular, the sharp change of R_t caused by the intervention is captured immediately by our approach while there is a lag using the sliding window-based method.

B. Evaluation of Intervention Measures in Europe

In this part, we applied the proposed framework to analyze the epidemic evolution in the 14 European Countries and also Wuhan. With the inferred $<R_t, p_t, D_t>$, we can then assess the impacts of intervention measures.

Figure 6 demonstrates the reconstruction of daily infections in the UK from the reported confirmed cases. The infected-to-report delay between report and infected time is composed of the incubation period (a lognormal distribution with a mean of 5.5 days and a standard deviation of 2.1 days [5]) and the onset-to-report period (a gamma distribution with a mean of 4.9 days and a standard deviation of 3.3 days [10]). The blue bars in Figure 6 indicate the numbers of confirmed cases. After deconvolving the confirmed numbers using infected-to-report delay, we obtained the infection curve (curve of estimated daily infected instances), which is colored in red in Figure 6. To check the reliability of the deconvolution results, we convolve the inferred infection curve (in red) with the infected-to-report delay to recover the confirmed curve (in black). We can see that the black curve matches well to the original blue bars, and is much smoother. With the above observation, we can see the effectiveness of the infection curve inference. Figure 7 shows the results of estimating R_t of the UK from the infection curve. The missing values in the infection curve are replaced by the average mean of the neighboring numbers. The green bar is the posterior mean of estimated R_t.

To quantitatively show the impacts of different strength levels of interventions, Table 1 summarizes the statistical analysis

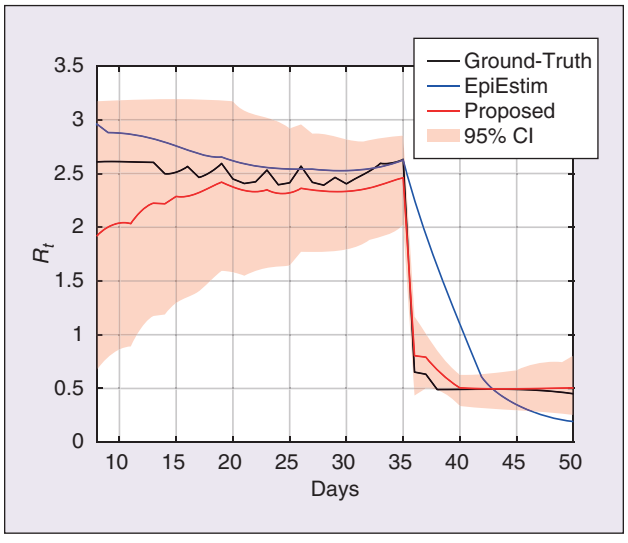


FIGURE 5 Validation of the proposed Bayesian updating scheme on simulated data with intensive intervention measure.

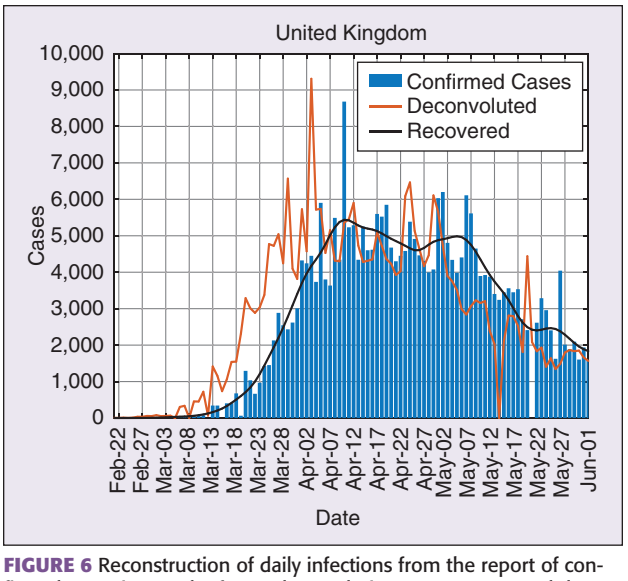


FIGURE 6 Reconstruction of daily infections from the report of confirmed cases in UK. The forward convolution on reconstructed data (black line) matches well with actual reported data (blue bars), validating the correctness of the deconvolution method.

results of 14 European countries. It shows different reduction rates of $<R_t, p_t, D_t>$ for different response levels. The relative reduction of $<R_t, p_t, D_t>$ compared to the minimal response (Level 0 where R_t is set to R_0) was estimated for each response level. With soft response (Level 1), the corresponding intervention measures (e.g. school closure, quarantine of international arrivals from high-risk regions) are correlated with a relative reduction of R_t by 35%, showing both strong suppression effect (D_t shortening 22%) and mitigation effect (p_t reduction 29%). With strong response (Level 2), the relative reduction of R_t increases to 60% with a strong mitigation effect (p_t reduction 56%). But the suppression effect (D_t shortening 26%) is similar to that of Level 1, indicating marginal incremental suppression effect. This observation shows a consistency with the aim of representative intervention measures on this level (e.g. canceling public events, restrictions on gathering and internal movements) to reduce the contact rates among the population.

The emergent response (Level 3) shows substantial relative reduction of reproductive number (R_t reduction 71%) with suppression (D_t shortening 37%) and mitigation (p_t reduction 67%) effects, correlated to the intensive measures (e.g. workplace closure and stay-at-home requirements). A similar degree of reductions is found for Level 4 (R_t reduction 74%; D_t shortening 40%; p_t reduction 70%) while the stringency of intervention measures is higher. We find that our estimated evolving patterns of p_t and D_t correspond well to the serial strategies taken by some European countries, such as the 'contain-delay-lockdown' route taken in the UK.

In addition to the results of 14 European Countries, Figure 8 shows the results of applying our method to the Wuhan data, where the greens bars indicate the posterior mean of R_t during the outbreak of COVID-19. We can see that at the early stage of the pandemic, the R_t levels are above 1. After the lockdown intervention has taken effect, R_t experienced a sharp decrease from 23rd Jan. When the centralized quarantine policy has been enforced from the beginning of February, the R_t values then largely remain below zero (the spike around 14th Feb is due to misreporting).

Figure 9 compares the reductions in $<R_t, p_t, D_t>$ for different response levels between European Countries and Wuhan. From the analysis of Wuhan data, the strong impact of lockdown is clearly demonstrated with the immediate relative reduction of R_t by 58%. We also observed that the combination of lockdown, centralized quarantine and immediate admission of confirmed patients starting from Feb 2nd in Wuhan was associated with a more substantial relative reduction of R_t with strong suppression and mitigation effects.

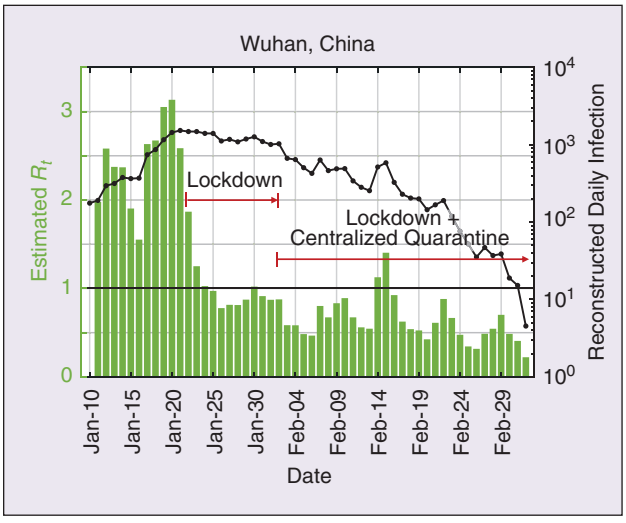


FIGURE 8 Estimated evolution of transmission dynamics in Wuhan. The black line represents the reconstructed daily infection number and the green bar is the posterior mean of estimated R_t. Two major events (city lockdown measure from 23rd Jan and centralized quarantine from 2nd Feb) are annotated with red arrows.

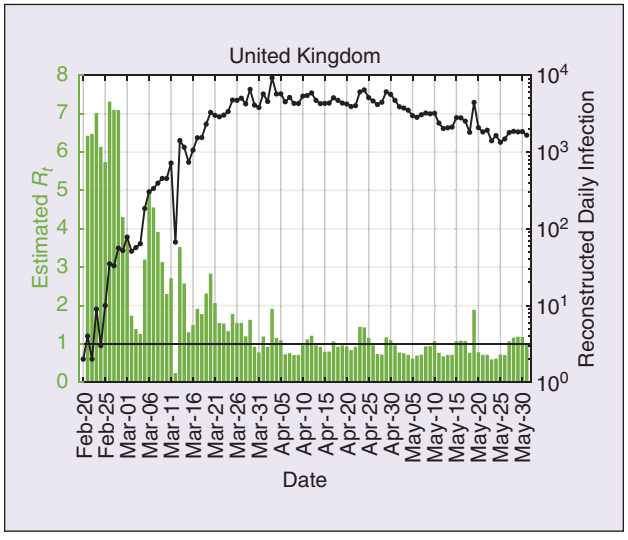


FIGURE 7 Estimated evolution of transmission dynamics in UK. The black line represents the reconstructed daily infection number and the green bar is the posterior mean of estimated R_t.

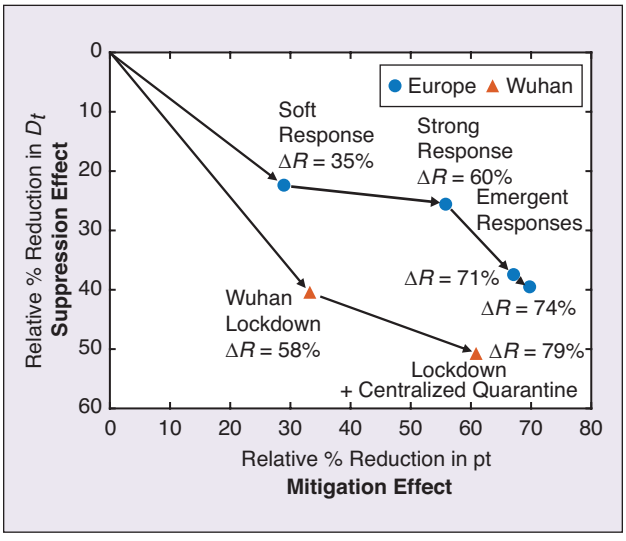


FIGURE 9 The relative reduction of mitigation factor p_t and suppression factor D_t under different response levels compared to minimal response level.

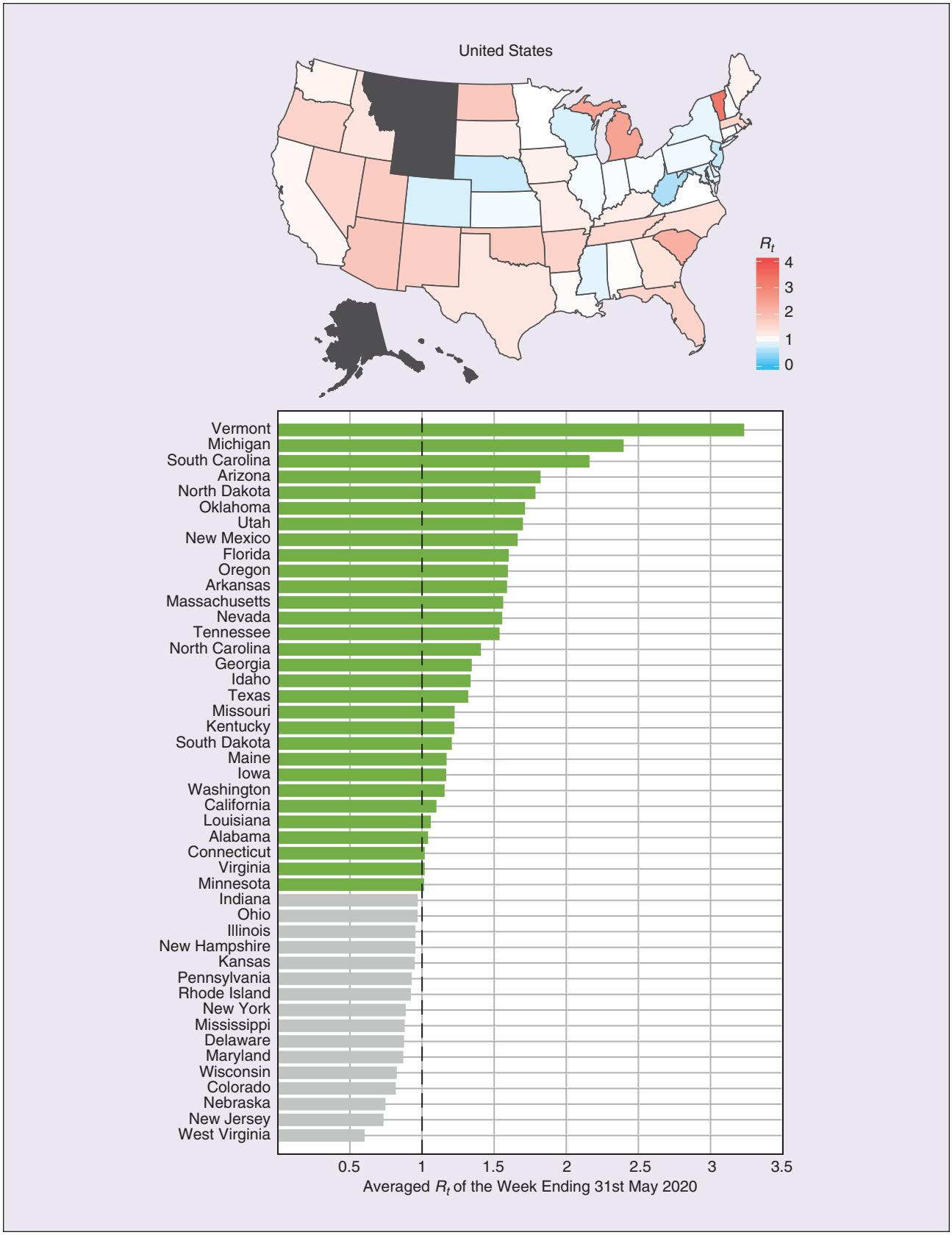


FIGURE 10 The averaged R_t values in different states of the United States. We report the result of averaged R_t in the US during the week ending 31st May 2020, which is ranked by the averaged R_t value (annotated with green if above 1, bottom). States with total confirmed cases less than 1,000 are excluded from the analysis.

C. Resurgence Risks in the United States

We also used the proposed framework to estimate the epidemic evolution in different states of the United States. We observed that, as of the week ending 31st May, the averaged reproduction number R_t in 30 states exceeds 1 (Figure 10). These could be related to the recent lift of government restrictions and alert us to take a close monitoring on the epidemic evolution.

At the time of preparing this paper (18th June 2020), 29 out of the 30 states we alerted on 9th June 2020 have experienced an increased number of daily confirmed cases compared to that of 31st May, and 14 states have recorded all-time high after 31st May. When we prepared the final version in early August, this alarming prediction of a second wave outbreak is unfortunately proven true for all the states listed.

So far, the application of the framework to many countries and the retrospective impact analysis of intervention measures in European countries indicate the effectiveness of our approach in monitoring R_t. This can be further validated by predicting the evolution of $<R_t, p_t, D_t>$ and projected infections in the future study. Our current study has several limitations. Firstly, the reporting protocols and standards of confirmed cases, as well as the detection rates, vary among countries. However, as long as the reporting bias is consistent over time, the inference results of p_t, D_t and R_t should also be consistent under the protocol. Since the impacts of interventions are assessed by measuring the evolution of these parameters, the framework can be generally applicable to assess the policy impacts among different reporting protocols. We also note that the implementation of multiple intervention measures within a short interval makes it challenging to quantify the impact of a single measure which needs further statistical analysis.

VII. Conclusions

In conclusion, we propose a comprehensive data assimilation approach of using Bayesian updating to timely estimate parameters of COVID-19 epidemic models. The disease transmission dynamics is modelled by renewal equations with time-varying parameters. Instead of purely focusing on estimating instantaneous reproduction number R_t, we introduce two complementary parameters, the mitigation factor (p_t) and the suppression factor (D_t), to quantify intervention impacts at a finer granularity. A Bayesian updating scheme is adopted to dynamically infer model parameters. By monitoring and analyzing the evolution of the estimated parameters, the impacts of intervention measures in different response levels can be quantitatively assessed. We have applied our method to European countries, the United States and Wuhan, and reveal the effects of interventions in these countries and the resurgence risk in the United States. Our work opens a promising venue to inform policy for better decision-making in response to a possible second-wave outbreak.

Acknowledgment

We express our sincere thanks to all members of the joint analysis team between Imperial College London, University of Cambridge, University of Kent, and Hong Kong Baptist University. We thank Yuting Xing for helping collect epidemic data in Wuhan and the United States. We thank Siyao Wang and Liqun Wu for their efforts on developing a digital tracing app for validation and visualization.

References

[1] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *Lancet Infect. Dis.*, vol. 20, no. 5, pp. 533–534, May 2020. doi: 10.1016/S1473-3099(20)30120-1.

[2] T. Hale, A. Petherick, T. Phillips, and S. Webster, "Variation in government responses to COVID-19," 2020.

[3] A. Pan et al., "Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China," *JAMA*, vol. 323, no. 19, p. 1915, May 2020. doi: 10.1001/jama.2020.6130.

[4] R. Li et al., "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2)," *Science*, vol. 3221, p. eabb3221, Mar. 2020.

[5] L. Ferretti et al., "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing.," *Science*, vol. 6936, pp. 1–13, Mar. 2020.

[6] E. Vynnycky and R. White, *An Introduction to Infectious Disease Modelling*. New York: Oxford Univ. Press, 2010.

[7] N. C. Grassly and C. Fraser, "Mathematical models of infectious disease transmission," *Nat. Rev. Microbiol.*, vol. 6, no. 6, pp. 477–487, 2008. doi: 10.1038/nrmicro1845.

[8] C. Fraser, "Estimating individual and household reproduction numbers in an emerging epidemic," *PLoS One*, vol. 2, no. 8, 2007. doi: 10.1371/journal.pone.0000758.

[9] J. Ma and D. J. D. Earn, "Generality of the final size formula for an epidemic of a newly invading infectious disease," *Bull. Math. Biol.*, vol. 68, no. 3, pp. 679–702, 2006. doi: 10.1007/s11538-005-9047-7.

[10] K. Leung, J. T. Wu, D. Liu, and G. M. Leung, "First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment," *Lancet*, vol. 395, no. 10233, pp. 1382–1393, Apr. 2020. doi: 10.1016/S0140-6736(20)30746-7.

[11] R. N. Thompson et al., "Improved inference of time-varying reproduction numbers during infectious disease outbreaks," *Epidemics*, vol. 29, Aug. 2019. doi: 10.1016/j.epidem.2019.100356.

[12] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez, "A new framework and software to estimate time-varying reproduction numbers during epidemics," *Am. J. Epidemiol.*, vol. 178, no. 9, pp. 1505–1512, 2013. doi: 10.1093/aje/kwt133.

[13] J. Wallinga and P. Teunis, "Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures," *Am. J. Epidemiol.*, vol. 160, no. 6, pp. 509–516, 2004. doi: 10.1093/aje/kwh255.

[14] D. Adam, "A guide to R-the pandemic's misunderstood metric," *Nature*, vol. 583, no. 7816, pp. 346–348, 2020. doi: 10.1038/d41586-020-02009-w.

[15] N. Imai, I. Dorigatti, A. Cori, C. Donnelly, S. Riley, and N. Ferguson, "Report 2: Estimating the potential total number of novel Coronavirus cases in Wuhan City, China," 2020.

[16] Q. Li et al., "Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia," *N. Engl. J. Med.*, vol. 382, no. 13, pp. 1199–1207, 2020.

[17] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study," *Lancet*, vol. 395, no. 10225, pp. 689–697, 2020. doi: 10.1016/S0140-6736(20)30260-9.

[18] S. Flaxman et al., "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe," *Nature*, pp. 1–5, 2020.

[19] P. Nouvellet et al., "Report 26: Reduction in mobility and COVID-19 transmission."

[20] M. Asch, M. Bocquet, and M. Nodet, *Data Assimilation: Methods, Algorithms, and Applications*. 2016.

[21] Z. Chen, "Bayesian filtering: From Kalman filters to particle filters, and beyond," *Statistics*, vol. 182, no. 1, pp. 1–69, 2003.

[22] C. J. Rhodes and T. D. Hollingsworth, "Variational data assimilation with epidemic models," *J. Theor. Biol.*, vol. 258, no. 4, pp. 591–602, 2009. doi: 10.1016/j.jtbi.2009.02.017.

[23] L. M. A. Bettencourt and R. M. Ribeiro, "Real time bayesian estimation of the epidemic potential of emerging infectious diseases," *PLoS One*, vol. 3, no. 5, p. e2185, 2008. doi: 10.1371/journal.pone.0002185.

[24] L. Cobb, A. Krishnamurthy, J. Mandel, and J. D. Beezley, "Bayesian tracking of emerging epidemics using ensemble optimal statistical interpolation," *Spat. Spatiotemporal. Epidemiol.*, vol. 10, pp. 39–48, 2014. doi: 10.1016/j.sste.2014.06.004.

[25] E. Goldstein, J. Dushoff, M. Junling, J. B. Plotkin, D. J. D. Earn, and M. Lipsitch, "Reconstructing influenza incidence by deconvolution of daily mortality time series," *Proc. Nat. Acad. Sci. U. S. A.*, vol. 106, no. 51, pp. 21,825–21,829, 2009. doi: 10.1073/pnas.0902958106.

[26] C. Mark, C. Metzner, L. Lautscham, P. L. Strissel, R. Strick, and B. Fabry, "Bayesian model selection for complex dynamic systems," *Nat. Commun.*, vol. 9, no. 1, 2018.

[27] A. Agresti, *An Introduction to Categorical Data Analysis*. Hoboken, NJ: Wiley, 2018.

COVID-19 Time Series Forecast Using Transmission Rate and Meteorological Parameters as Features

Mohsen Mousavi
University of Tasmania, AUSTRALIA and University of Technology Sydney, AUSTRALIA

Rohit Salgotra
Thapar Institute of Engineering & Technology, INDIA

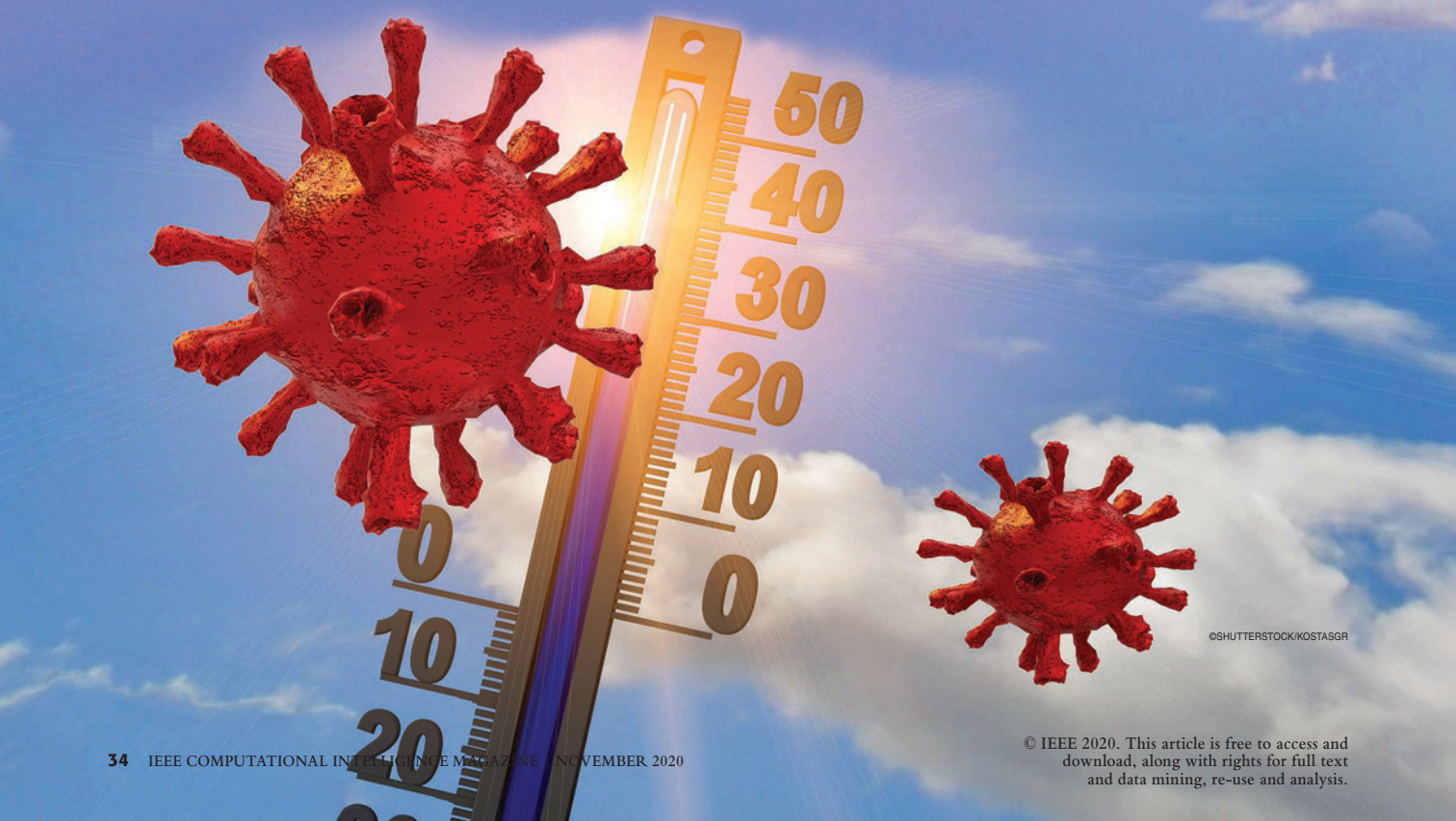
Damien Holloway
University of Tasmania, AUSTRALIA

Amir H. Gandomi
University of Technology Sydney, AUSTRALIA

***Abstract*—**The number of confirmed cases of COVID-19 has been ever increasing worldwide since its outbreak in Wuhan, China. As such, many researchers have sought to predict the dynamics of the virus spread in different parts of the globe. In this paper, a novel systematic platform for prediction of the future number of confirmed cases of COVID-19 is proposed, based on several factors such as transmission rate, temperature, and humidity. The proposed strategy derives systematically a set of appropriate features for training Recurrent Neural Networks (RNN). To that end, the number of confirmed cases (CC) of COVID-19 in three states of India (Maharashtra, Tamil Nadu and Gujarat) is taken as a case study. It has been noted that stationary and non-stationary parts of the features improved the prediction of the stationary and non-stationary trends of the number of confirmed cases, respectively. The new platform has general application and can be used for pandemic time series forecasting.

Digital Object Identifier 10.1109/MCI.2020.3019895
Date of current version: 14 October 2020

Corresponding Author: Amir H. Gandomi (gandomi@uts.edu.au).



©SHUTTERSTOCK/KOSTASGR

I. Introduction

... in this section we propose a novel framework for optimum training of an RNN using time series analysis of the signals used in training.

The novel coronavirus (SARS-CoV-2) has plunged the world into severe disaster recently. The virus made its way to many countries around the globe soon after the first case was reported in Wuhan, Hubei Province, People's Republic of China (PRC) in late December [1]. As such, the World Health Organization (WHO) declared the situation as a public health emergency of international concern on 30 January, 2020 [2]. WHO officially named the disease COVID-19 when PRC Center for Disease Control and Prevention (CDC) recognized the virus as a new type of coronavirus. Ever since, many countries have experienced disasters due to the widespreading infectious virus. This has put a huge burden on medical centers in different countries and many different measures have been put in place by jurisdictions to control the spread of the virus in different countries. These measures are mainly in the form of lockdowns enforced in several stages, where people are banned from congregating *en messe*. Physical distancing measures can have a huge impact on the virus transmission rate [3], and in one such study the time taken for the daily number of new cases to double was reported to increase from 2 to 4 days [4]. The optimal lockdown policy depends on the fraction of infected and susceptible in the population. As a result, a severe lockdown beginning two weeks after the outbreak was prescribed where it can be gradually relaxed to cover 60% of the population after a month, and 20% of the population after three months [5]. It was also recommended that the intensity of the lockdown should depend on the gradient of the fatality rate as a proportion of the infected, and on the assumed value of a statistical life [5].

The effect of the Meteorological parameters on the spread of the COVID-19 disease has also been investigated [6], [7]. It has been reported that the virus favors low temperature and low humidity [8]–[11]. Mortality is also shown to be affected by temperature and humidity variation [12]: one unit increase of temperature and absolute humidity was associated with a decreased COVID-19 death rate. Accordingly, temperature and humidity are suggested as important factors to be considered in modelling of rates.

Some studies have focused on prediction of the number of future cases with different lockdown policies in different countries [13]–[15]. This will facilitate the investigation of the effect of different measures on the future spread of the virus by administrators and health officials.

This paper uses both transmission rate and meteorological parameters (temperature and humidity) as features for training a set of Recurrent Neural Networks (RNN) to forecast the number of future cases of COVID-19. A systematic procedure is proposed in this paper which decomposes each signal (all features as well as the signal to be predicted) into its stationary and non-stationary modes. All the stationary modes that are similar in center frequency are used to train a separate RNN. Similarly, all the non-stationary modes are used to train another RNN. The results of all of the predictions are summed as the final forecast number of COVID-19 cases.

India is one of the highly impacted countries, and has been severely hit by the spread of the COVID-19 virus in many of its states. Some researchers have sought to predict the effect of lockdown measures on the spread of the virus in India and have suggested some policies to be followed by the jurisdictions to fight further spread of the virus in the country [16]–[18]. In [16], [17], an evolutionary data analytical method called genetic programming was used to predict the possible impact of COVID-19 in India. Here only two parameters, namely confirmed cases and total death count, were taken into consideration to analyze and predict the total rise in the coming ten days. The present work extends the former basic parametric analysis, adding transmission rate from outbreak and the local meteorological temperature and humidity data. In this paper, the data from the outbreak in different parts of India have been taken as the case study. The source of dataset for comparison is available at [19].

II. Calculation of the Transmission Rate

The number of the confirmed cases has continually increased since 24 March 2020 when an outbreak was declared in different states of India. Figure 1 shows the number of daily new confirmed cases in two of the severely affected states, namely Maharashtra (Figure 1(a)) and Tamil Nadu (Figure 1(b)), since 24 of March. These two states are taken as the case studies in this paper.

There have been overall, five lockdown periods in India as of 24 March, 2020, followed by an unlock phase. The information about each lockdown phase is outlined in the Table I. Using the number of confirmed cases corresponding to each lockdown phase, the value of the transmission rate in that phase has been calculated through the following formula [20],

$$\beta = -\frac{1}{T}\log\left(1 - I_N\left(\frac{1}{I} + \frac{1}{S}\right)\right) \tag{1}$$

where β is the mean estimated transmission rate for each lockdown phase, I_N is the number of new infections since the previous lockdown, I and S represent respectively the number of infected and susceptible individuals, and T is the sampling interval.

The transmission rates corresponding to the outbreak in Maharashtra and Tamil Nadu have been calculated using (1) for all lockdown phases of Table I. Note that the number of susceptible cases S in each state has been considered to be the entire population of that state. Figures 2(a) and 2(b) show respectively the calculated transmission rates in Maharashtra

Before constructing an RNN, we propose to pre-analyse the data to explore the nature of the signals used for training.

and Tamil Nadu per day. As can be seen from the figures, the transmission rate graphs resemble step functions. It is hypothesised that the effect of lockdown does not have immediate effects on the transmission rate as it takes time for the entire population to adapt their behaviour to the new set of rules. A robust spline based smoothing technique is exploited to slightly smooth these graphs. The so-called smoothing technique aims at balancing the fidelity in the data by minimising the following goal function [21],

$$F(\hat{S}(t)) = \left\| \hat{S}(t) - S(t) \right\|^2 + rP(\hat{S}(t)), \tag{2}$$

where $S(t)$ and $\hat{S}(t)$ represent respectively the original and smoothed signals, and $P(\hat{S}(t))$ is a penalty term that reflects the roughness of the obtained smoothed signal ($\hat{S}(t)$). The real positive scalar parameter r is the smoothing factor that controls the degree of smoothness in $\hat{S}(t)$. A smoothing factor of $r = 10$ has been used to this end. Figures 2(c) and 2(d) show the smoothed graphs of transmission rates for Maharashtra and Tamil Nadu, respectively.

III. Signal Pre-Processing

This section presents the procedure of the proposed method, aiming to obtain a computational model that can forecast the number of confirmed cases of the COVID-19 in Maharashtra and Tamil Nadu. RNN has been widely used for time series forecasting; in this section we propose a novel framework for optimum training of an RNN using time series analysis of the signals used in training.

To train a supervised Artificial Neural Network (ANN), one needs to decide the features and labels to be used for training. Here we discuss how these features and labels can be selected systematically. As stated earlier, we hypothesise that the number of confirmed cases of COVID-19 is a function of the variability of environmental conditions (temperature and humidity), and the measures put in place by the jurisdictions to control the spread of the virus (transmission rates). The effectiveness of such measures is usually reflected by the transmission rates varying with different lockdown phases. As a result, there are four different time series introduced to the training process in this paper: (1) temperature (T), (2) humidity (H), (3) the number of confirmed cases (CC), and (4) the transmission rates (TR). Restated, this paper aims to construct an RNN to predict the future number of CC signal based on its previous observed numbers and the other aforementioned signals (T, H, and TR).

A. Time Series Analysis of Features

Before constructing an RNN, we propose to pre-analyse the data to explore the nature of the signals used for training. Since the signals used in this paper have a stochastic nature, their stationary or non-stationary behaviour is first processed in this section. This will result in more accurate training and ensure much better prediction results. First, a brief definition of the stationary and non-stationary time series is presented.

The first order autoregressive process $AR(1)$ of a signal $S(t)$ is shown as

$$s_t = \phi\, s_{t-1} + \epsilon_t \tag{3}$$

where ϵ_t is a stationary white Gaussian noise process. Three different scenarios can occur for the above $AR(1)$ model: 1)

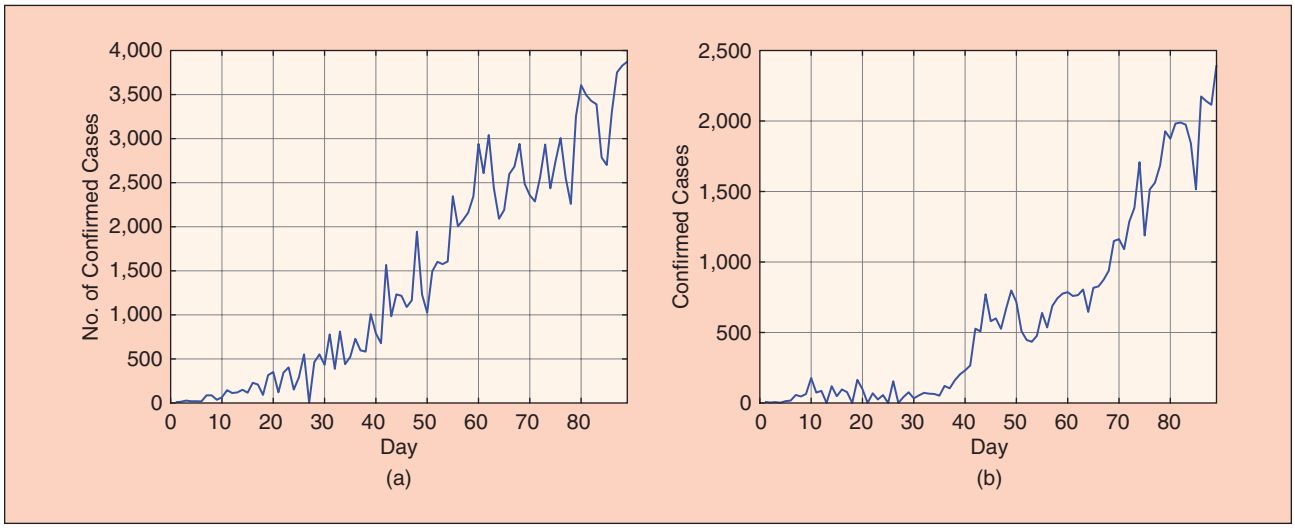


FIGURE 1 The number of daily new confirmed cases of COVID-19 in Maharashtra and Tamil Nadu as of 24 March, 2020. (a) Maharashtra. (b) Tamil Nadu.

$|\phi|<1$ implies the signal is stationary, 2) $|\phi|>1$ shows that the signal is non-stationary, and 3) $|\phi|=1$, represents a random walk model [22], [23].

In this paper, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test is run on each signal to explore the stationary and non-stationary nature of the signal. The KPSS test is used for testing a null hypothesis of stationary time series (no unit root) around a deterministic trend (i.e. trend-stationary) against the alternative of non-stationary (unit root) [24]. The null hypothesis of trend stationary of the signal is tested against the alternative hypothesis of trend non-stationary. The test can be conducted on several autocovariance lags in the Newey-West estimator [25] of the long-run variance, each conducted at 0.1 significance level using MATLAB.

However, before running a KPSS test, one needs to select an appropriate lag length for the time series. Care must be taken to ensure that an appropriate lag length is chosen. For instance, if the lag length is too short, the test will be biased; if the lag length is too large, the power of the test will suffer. A common rule of thumb for determining the maximum lag ($L_{\max}$) can be obtained from the following equation [24],

$$L_{\max} = \left[12 \times \left(\frac{n}{100}\right)^{\frac{1}{4}}\right] \tag{4}$$

where n is the sample size and $[\cdot]$ indicates the integer part of a number. Regarding the examples of this paper $n = 80$, a maximum lag ($L_{\max}$) 11 is obtained. Three different values of 7, 9, and 11 for lags are considered for the KPSS test.

TABLE I The date of the start and end of lockdown phases in India as of 24 March, 2020.

LOCKDOWN PHASE	TIME PERIOD
I	24/03/2020–13/04/2020
II	14/04/2020–3/05/2020
III	3/05/2020–17/05/2020
IV	17/05/2020–31/05/2020
V	31/05/2020–8/06/2020
UNLOCK	8/06/2020–21/06/2020 (ONGOING)

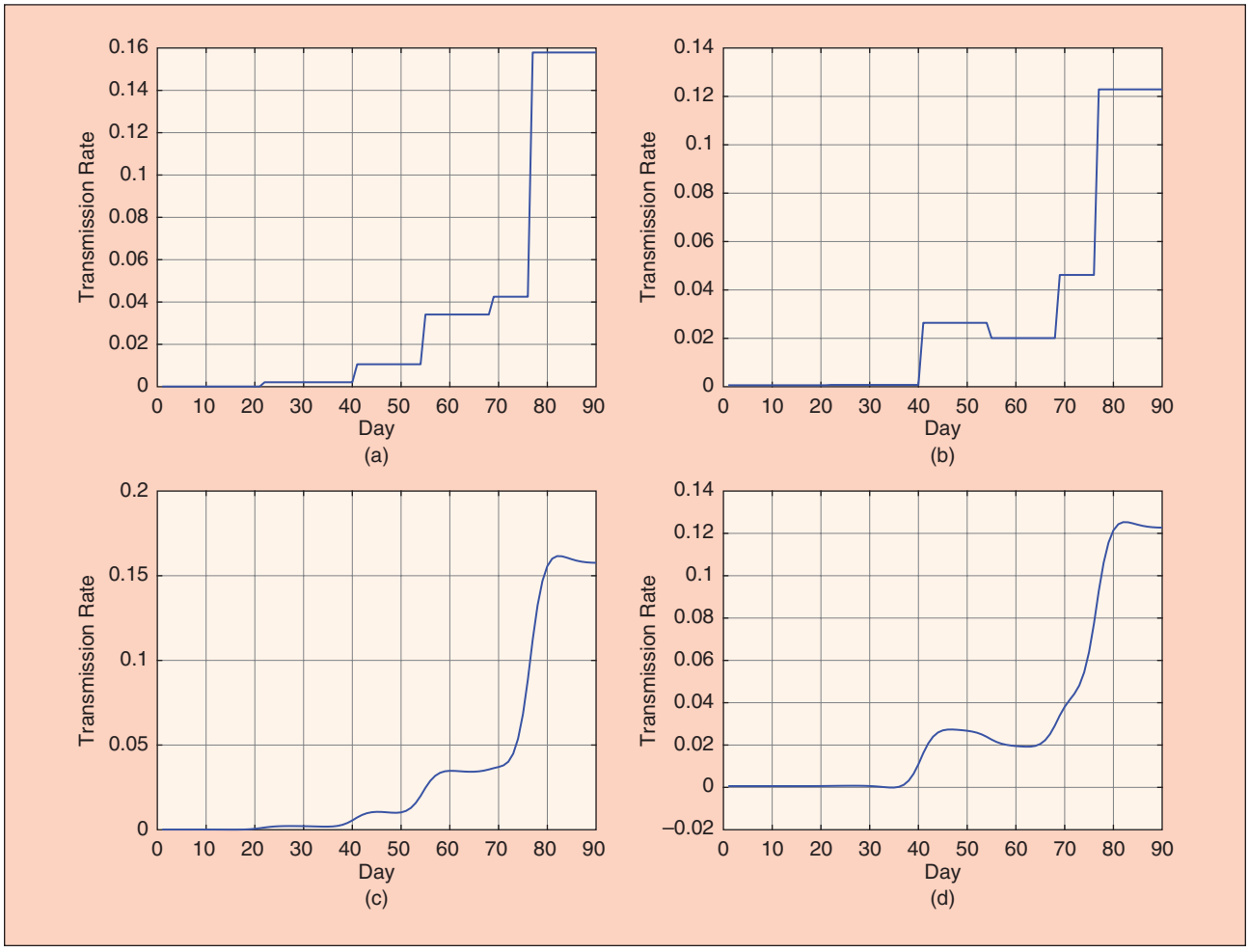


FIGURE 2 Calculated rough (a, b) and smoothed (c, d) transmission rates of COVID-19 corresponding to Maharashtra and Tamil Nadu for different lockdown phases as of 24 March, 2020. (a) Maharashtra. (b) Tamil Nadu. (c) Maharashtra, smoothed. (d) Tamil Nadu, smoothed.

Figures 3 and 4 show the temperature (T) and humidity (H) for Maharashtra and Tamil Nadu, respectively. Table II shows the results of the KPSS test run on CC, TR, T and H signals of Maharashtra.

As can be seen from the results, the KPSS test rejects the null hypothesis in favor of the alternative for the signals CC and the TR with a relatively small P-value (compared to the significance level 0.1) in all forms of the signals associated with the specified auto-covariance lags 7, 9, and 11. These signals therefore are considered non-stationary. The opposite results are obtained for the signals T and H, as can be seen from the table.

Likewise, Table III shows the results of the KPSS test run on each signal CC, TR, T, and H of Tamil Nadu. The KPSS test rejects the null hypothesis in favor of the alternative for the signals CC, TR, and H with a relatively small P-value (compared to the significance level 0.1) in all forms of the signals associated with the specified auto-covariance lags 7, 9, and 11. These signals are thus considered non-stationary. The opposite results are obtained for the signal T as seen from the table.

In the next section, more complicated signals, i.e. all signals except TR, are decomposed into some stationary and non-stationary modes using an advanced signal decomposition technique. This will further help in using features with low level of irregularities in the training process, which can further improve training results.

B. Signal Decomposition Using VMD

This section proposes to decompose complex signals (CC, T, and H) into their stationary and non-stationary oscillatory modes using an advanced decomposition technique called Variational Mode Decomposition (VMD). We further use the non-stationary part of the signals along with the signal TR

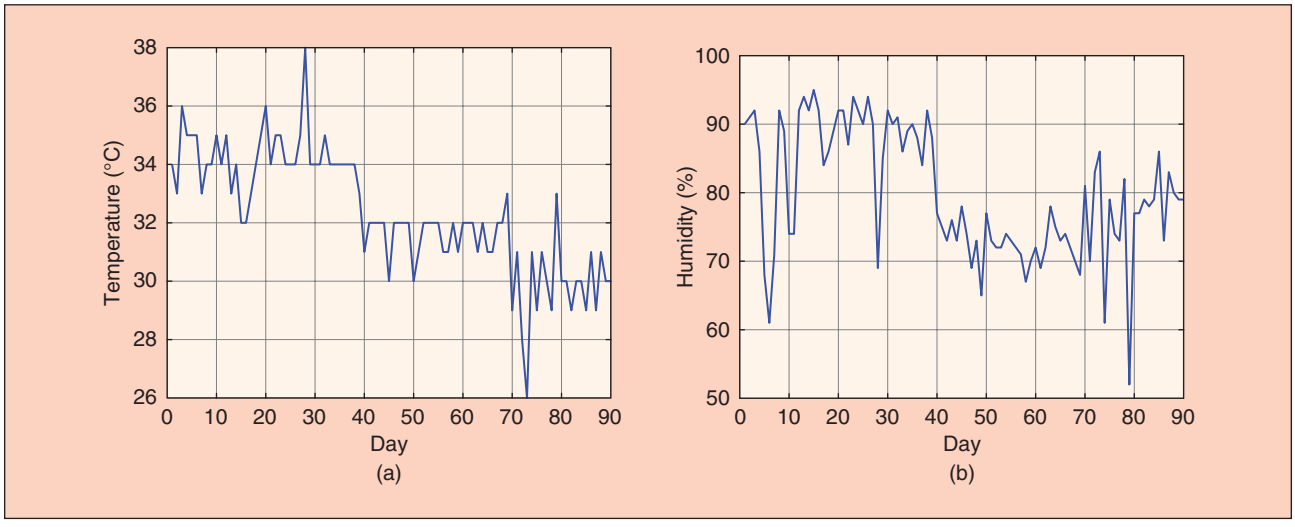


FIGURE 3 (a) Temperature, and (b) humidity time series corresponding to Maharashtra as of 24/3/2020 to 21/06/2020.

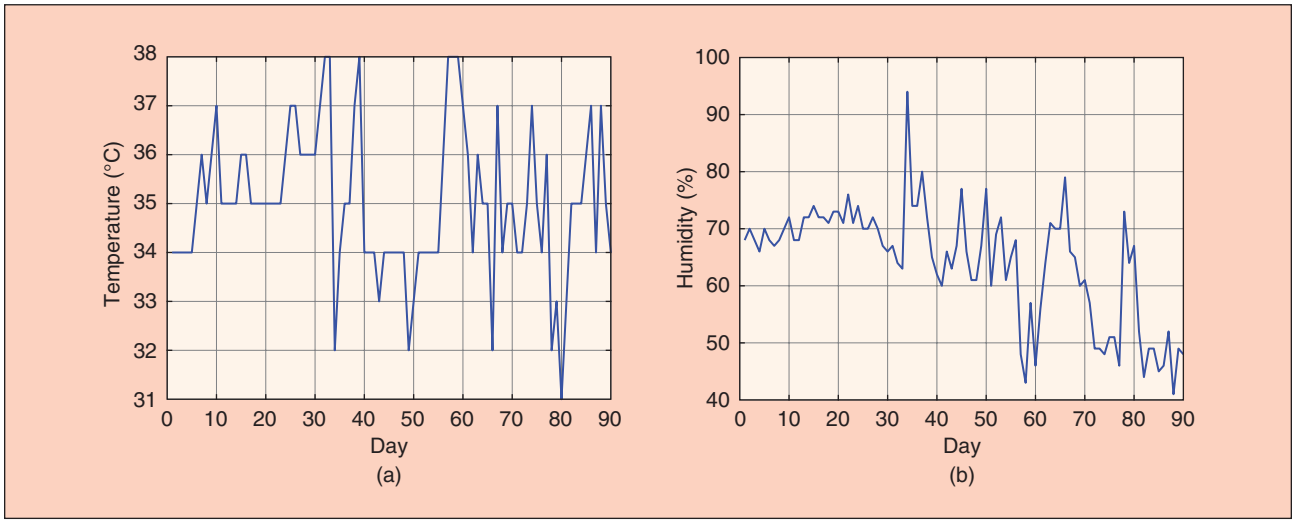


FIGURE 4 (a) Temperature, and (b) humidity time series corresponding to Tamil Nadu as of 24/3/2020 to 21/06/2020.

(non-stationary feature) for training. Similarly, the stationary modes are used to train another set of RNNs (Figure 5).

VMD is an adaptive decomposition algorithm that aims to decompose a non-linear non-stationary signal into its constructive modes [26]. These modes, which are known as Intrinsic Mode Functions (IMF), are frequency and/or amplitude modulated signals. The sum of which constructs the original signal (minus some noise, depending on settings).

VMD is an adaptive algorithm which solves a variational optimisation problem for a given signal $S(t)$ on k IMFs $\{u_k\} = \{u_1, u_2, \ldots, u_k\}$. It is assumed that each IMF is narrow-band and, therefore, has a center frequency $\{\omega_i\}$ where $i \in \{1, 2, \ldots, k\}$. The aforementioned variational optimisation problem follows,

$$\min_{\{u_k\}\&\{\omega_k\}} \sum_k \left\| \partial_t \left(\delta(t) + \frac{j}{\pi t} * u_k(t) \right) e^{-j\omega_k t} \right\|^2 \tag{5}$$

where in the above equation, $*$ is the convolution operator. The proposers of VMD argue that the solution to the minimization problem of (5) is the saddle point of the augmented Lagrangian in a sequence of iterative sub-optimizations called alternate direction method of multipliers (ADMM) [26]. The readers are referred to the original paper for further details.

There are some critical parameters that need to be determined when using VMD for signal decomposition:

1) The number of modes (k) into which the signal is chosen to be decomposed.
2) The weight of the quadratic penalty term α, which is a denoising factor, a larger value of which admits less noise into the decomposition process. Note that in this paper, α is set to a relatively small value of 10 since denoising is not a concern [27].

TABLE II KPSS test results run on the signals corresponding to Maharashtra. Note that ST stands for stationary.

SIGNAL	LAG	P-VALUE	H	ST
CC	7, 9, 11	0.02, 0.03, 0.04	1, 1, 1	×
TR	7, 9, 11	0.01, 0.01, 0.02	1, 1, 1	×
T	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
H	7, 9, 11	0.07 0.10 0.10	0, 0, 0	✓

TABLE III KPSS test results run on the signals corresponding to Tamil Nadu.

SIGNAL	LAG	P-VALUE	H	ST
CC	7, 9, 11	0.01, 0.01, 0.02	1, 1, 1	×
TR	7, 9, 11	0.01, 0.02, 0.03	1, 1, 1	×
T	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
H	7, 9, 11	0.05, 0.03, 0.02	1, 1, 1	×

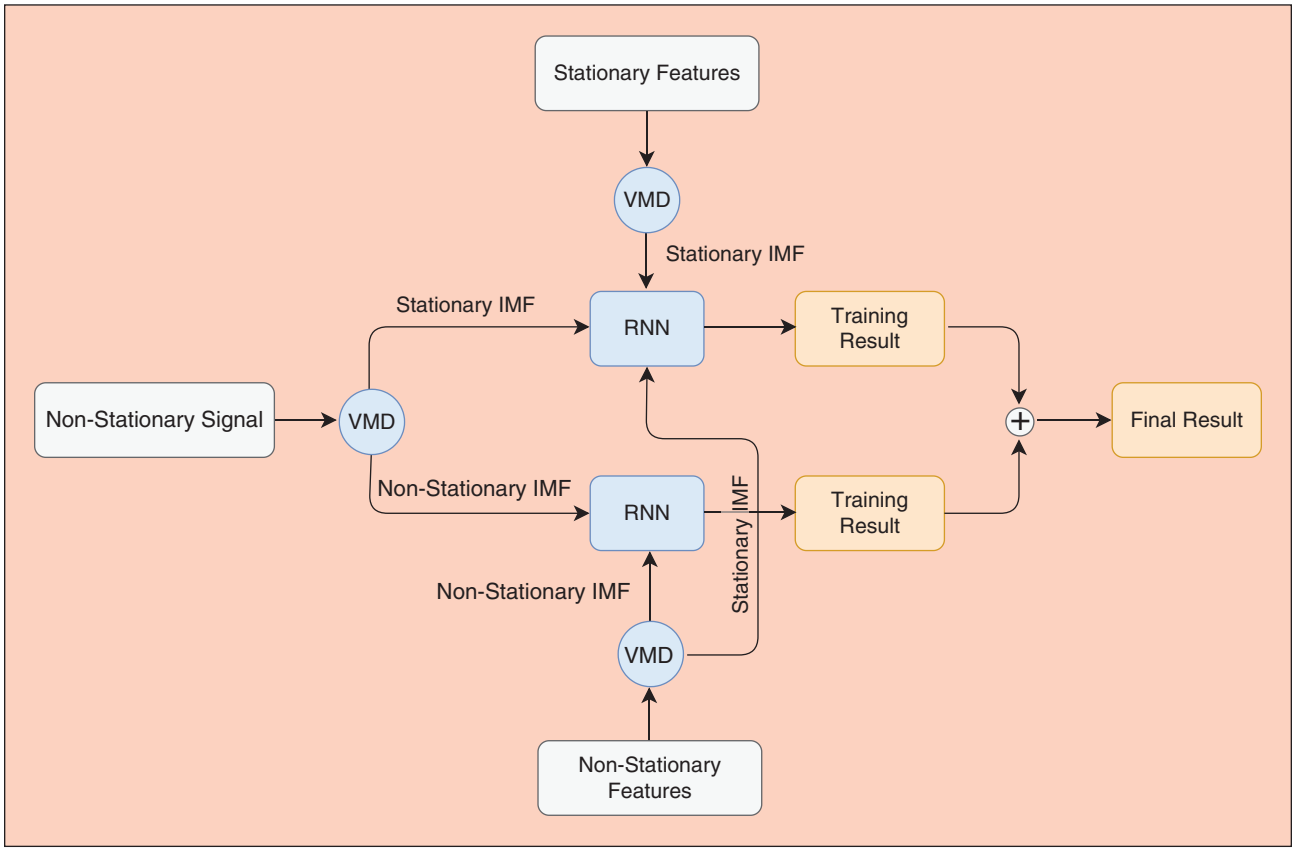


FIGURE 5 Flowchart of the proposed methodology.

3) The tolerance parameter ϵ, which controls the convergence of the algorithm. This value is set to 10^{-7} in this paper.

All signals have been decomposed into three modes. Figures 6 and 7 show respectively the IMFs corresponding to the CC signals of Maharashtra and Tamil Nadu along with their corresponding center frequencies.

We further run the KPSS test on IMFs corresponding to the VMD decomposition of the CC signals for both states using the same lags used in Section III-A, i.e. 7, 9, and 11. As is

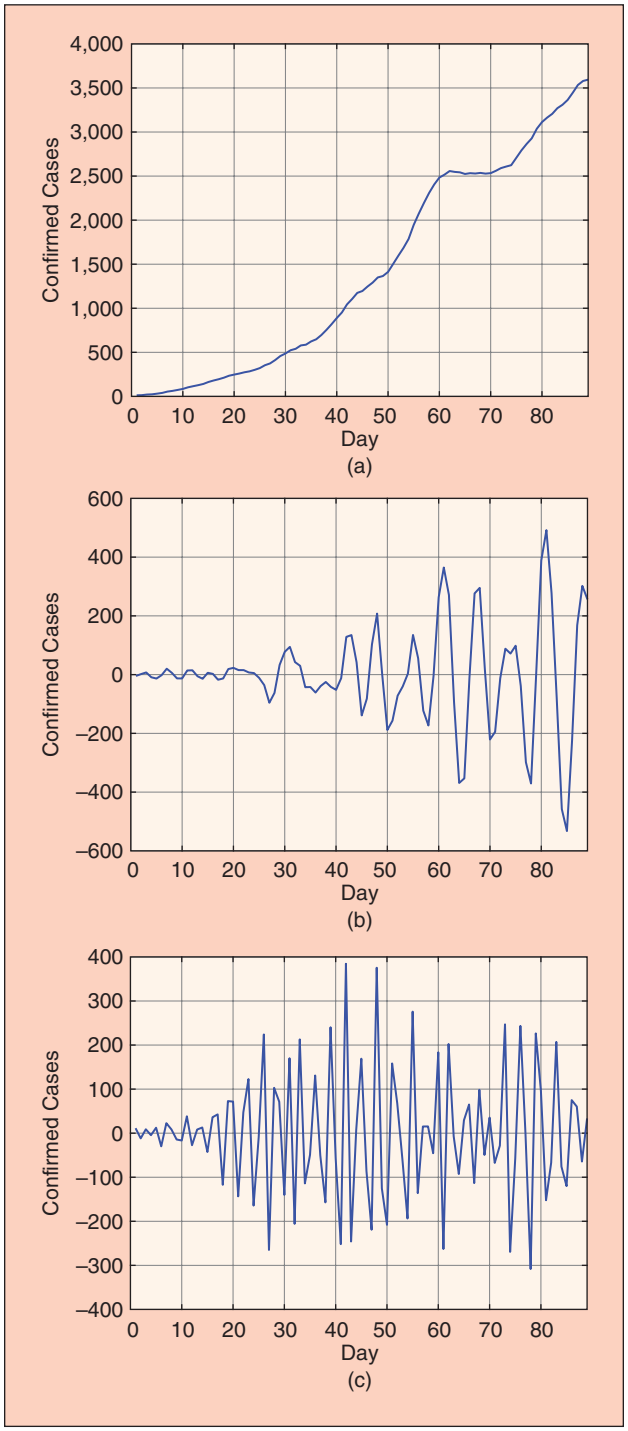


FIGURE 6 IMFs corresponding to the decomposition of CC time series of Maharashtra along with their center frequencies, (a) CC-IMF$_1$, $\omega_1 = 0.0015$. (b) CC-IMF$_2$, $\omega_2 = 0.1423$. (c) CC-IMF$_3$, $\omega_3 = 0.3480$.

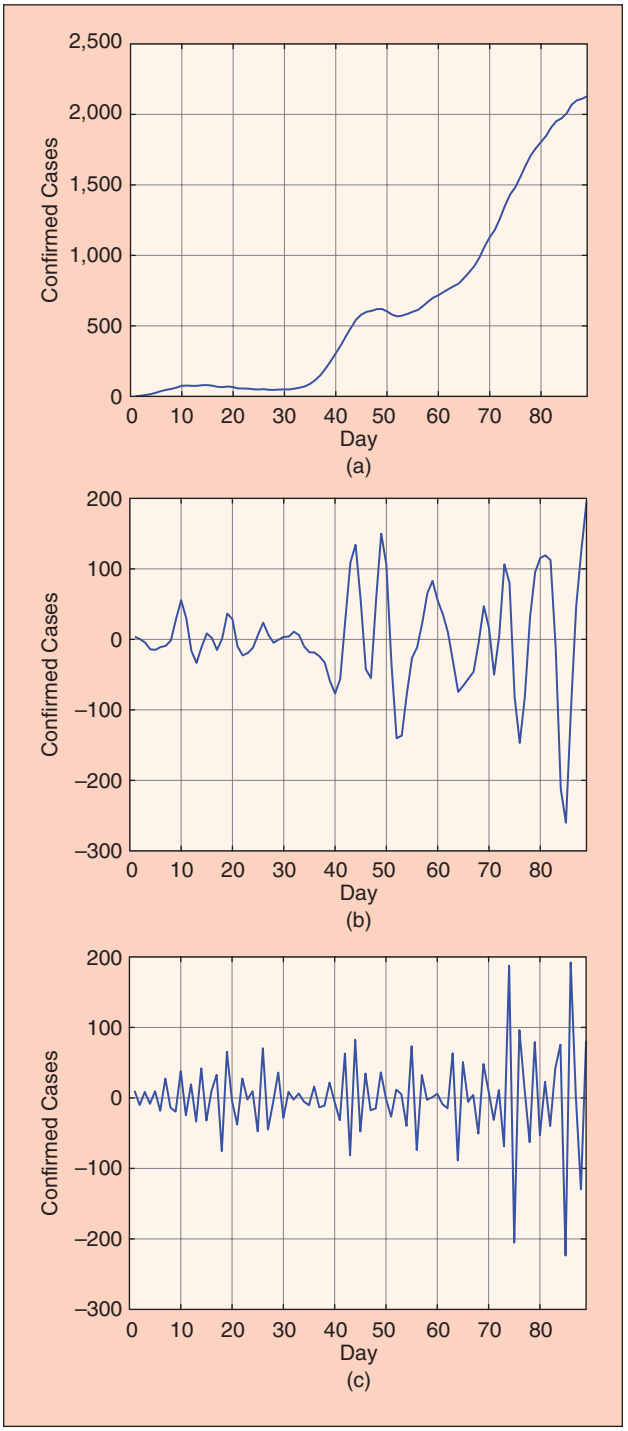


FIGURE 7 IMFs corresponding to the decomposition of CC time series of Tamil Nadu along with their center frequencies, (a) CC-IMF$_1$, $\omega_1 = 0.0025$. (b) CC-IMF$_2$, $\omega_2 = 0.1264$. (c) CC-IMF$_3$, $\omega_3 = 0.3866$.

evident from the KPSS test results, the first IMF of this decomposition in both cases is non-stationary while the remainder are stationary (Tables IV and V).

The same procedure is followed for signals T and H corresponding to Maharashtra (Figures 8 and 9) and Tamil Nadu (Figures 10 and 11). These signals are first decomposed into three IMFs, then the KPSS test is run on each IMF.

A set of multivariate stacked RNNs is developed to forecast the future values of each CC-IMF signals using the results of the previous section.

Tables VI and VII show respectively the results of KPSS tests run on IMFs of signals T and H corresponding to Maharashtra. There is no non-stationary trend in the IMFs of these signals. Likewise, Tables VIII and IX show respectively the results of KPSS tests run on IMFs of signals T and H corresponding to Tamil Nadu. As expected, regarding the H signal, H-IMF$_1$ has a non-stationary trend whereas other IMFs are stationary. As for the T signal, all IMFs show stationary trends again as expected.

The following conclusions can be made from the decomposition and KPSS test results. Regarding the decomposition of signals corresponding to Maharashtra:

1) Signals CC-IMF$_1$ (Figure 6(a)) and TR (Figure 2(c)) are trend non-stationary and, therefore, are used to train a separate RNN.
2) Signals CC-IMF$_2$ (Figure 6(b)), T-IMF$_2$ (Figure 8(b)) and H-IMF$_2$ (Figure 9(b)) are stationary and have similar center frequencies of 0.1423, 0.1507, and 0.1858 respectively and, therefore, are used to train a separate RNN.
3) Signals CC-IMF$_3$ (Figure 6(c)), T-IMF$_3$ (Figure 8(c)) and H-IMF$_3$ (Figure 9(c)) are stationary and have similar center frequencies of 0.3480, 0.3986, and 0.3860 respectively and, therefore, are used to train a separate RNN.
4) Signals T-IMF$_1$ (Figure 8(a)) and H-IMF$_1$ (Figure 9(a)) are stationary but are excluded from training process as they have no similarity to any other stationary IMFs in terms of center frequency.

Regarding decomposition of signals corresponding to Tamil Nadu:

TABLE IV KPSS test for the IMFs corresponding to the signal CC of Maharashtra.

SIGNAL	LAG	P-VALUE	H	ST
CC-IMF$_1$	7, 9, 11	0.02, 0.03, 0.04	1, 1, 1	×
CC-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
CC-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

TABLE V KPSS test for the IMFs corresponding to the signal CC of Tamil Nadu.

SIGNAL	LAG	P-VALUE	H	ST
CC-IMF$_1$	7, 9, 11	0.01, 0.01, 0.02	1, 1, 1	×
CC-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
CC-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

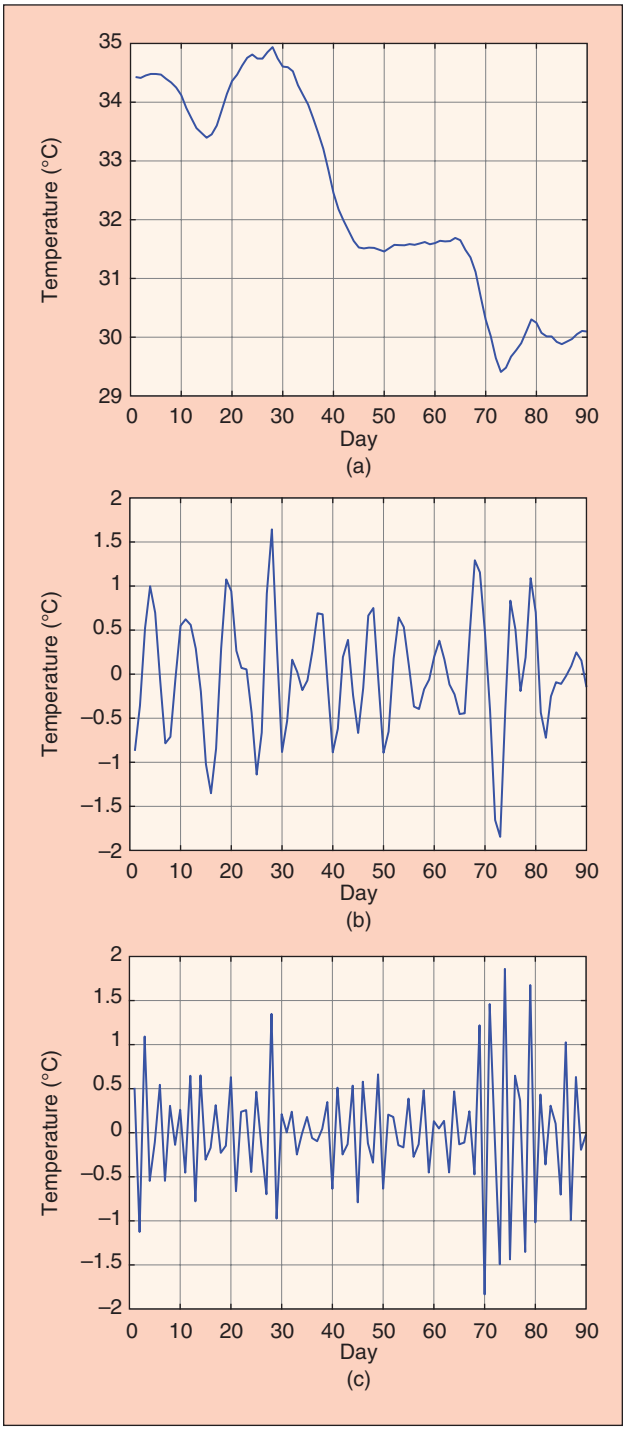


FIGURE 8 IMFs corresponding to the decomposition of temperature (T) time series of Maharashtra along with their center frequencies, (a) T-IMF$_1$, $\omega_1 = 10^{-5}$. (b) T-IMF$_2$, $\omega_2 = 0.1507$. (c) T-IMF$_3$, $\omega_3 = 0.3986$.

1) Signals CC-IMF$_1$ (Figure 7(a)), TR (Figure 2(d)), and H-IMF$_1$ (Figure 11(a)) are trend non-stationary and, therefore, are used to train a separate RNN.
2) Signals CC-IMF$_2$ (Figure 7(b)), T-IMF$_2$ (Figure 10(b)) and H-IMF$_2$ (Figure 11(b)) are stationary and have relatively similar center frequencies of 0.1264, 0.1312, and 0.0887 respectively and, therefore, are used to train a separate RNN.
3) Signals CC-IMF$_3$ (Figure 7(c)), T-IMF$_3$ (Figure 10(c)) and H-IMF$_3$ (Figure 11(c)) are stationary and have similar

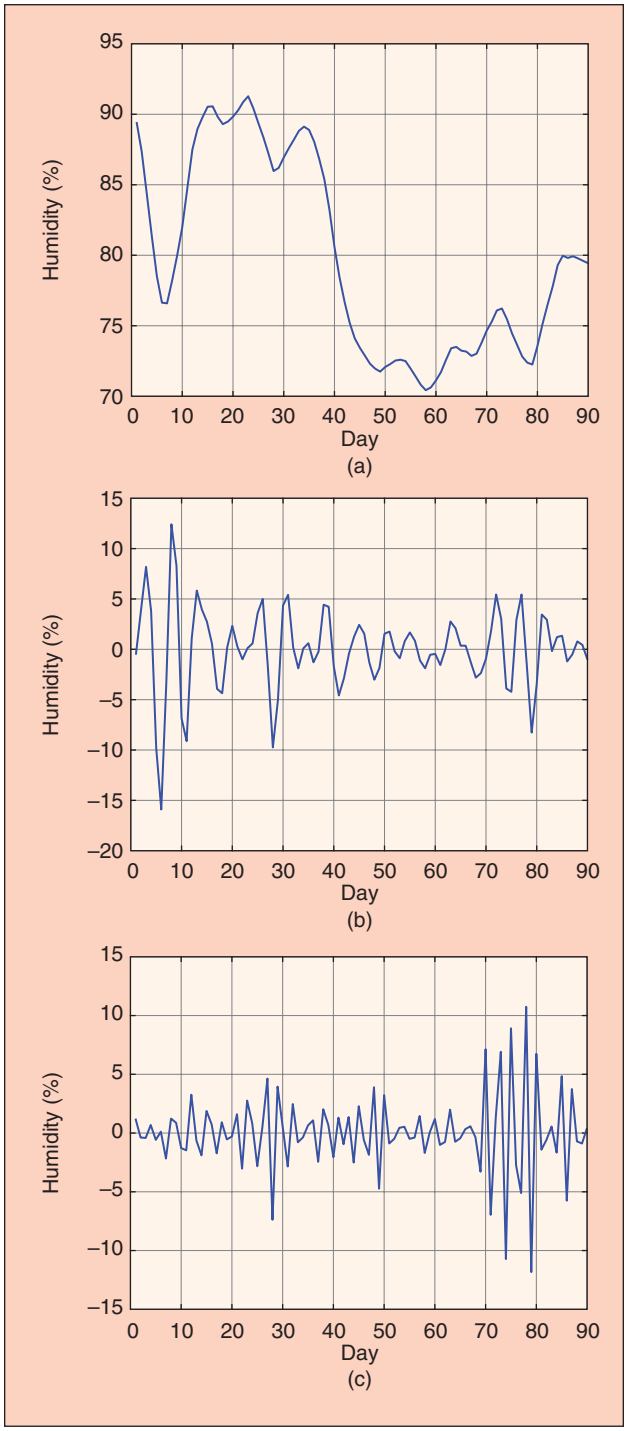


FIGURE 9 IMFs corresponding to the decomposition of humidity (H) time series of Maharashtra along with their center frequencies, (a) H-IMF$_1$, $\omega_1 = 0.0001$. (b) H-IMF$_2$, $\omega_2 = 0.1858$. (c) H-IMF$_3$, $\omega_3 = 0.3860$.

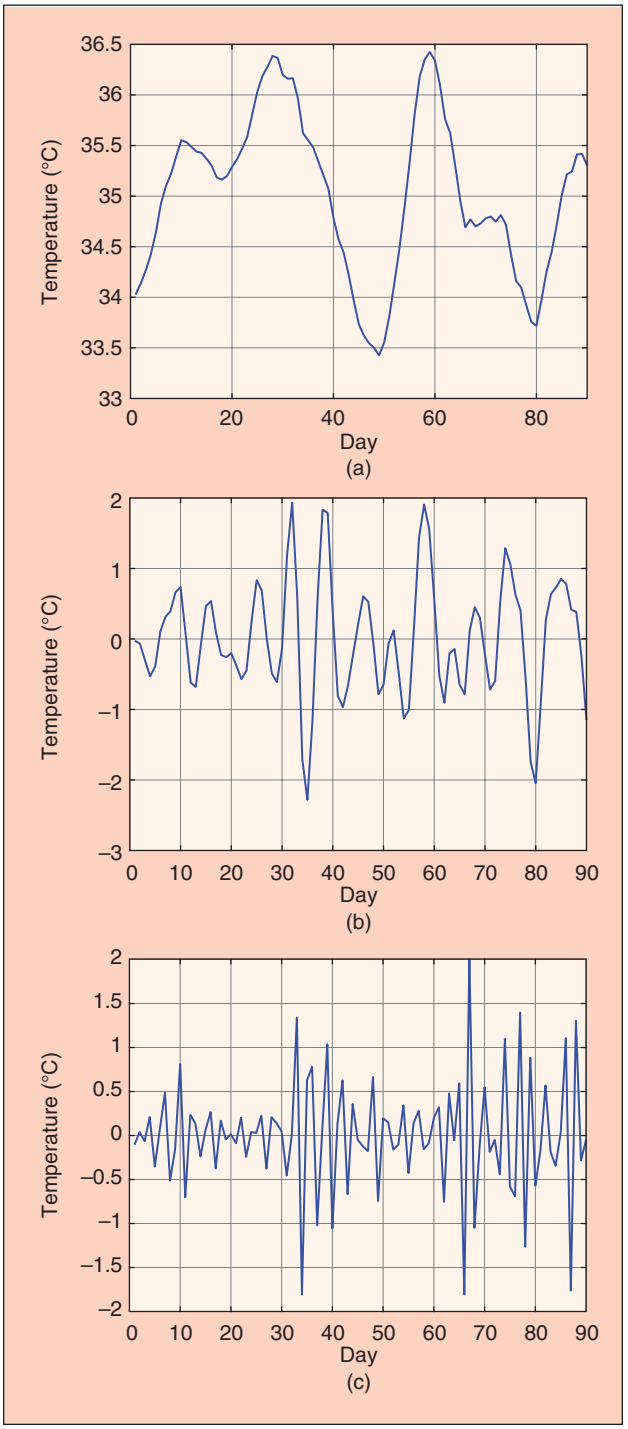


FIGURE 10 IMFs corresponding to the decomposition of temperature (T) time series of Tamil Nadu along with their center frequencies, (a) T-IMF$_1$, $\omega_1 = 9 \times 10^{-6}$. (b) T-IMF$_2$, $\omega_2 = 0.1312$. (c) T-IMF$_3$, $\omega_3 = 0.3804$.

center frequencies of 0.3866, 0.3804, and 0.3525 respectively and, therefore, are used to train a separate RNN.

4) Signal T-IMF$_1$ (Figure 10(a)) is stationary but is excluded from training process as it has no similarity to any other stationary IMFs in terms of center frequency.[1]

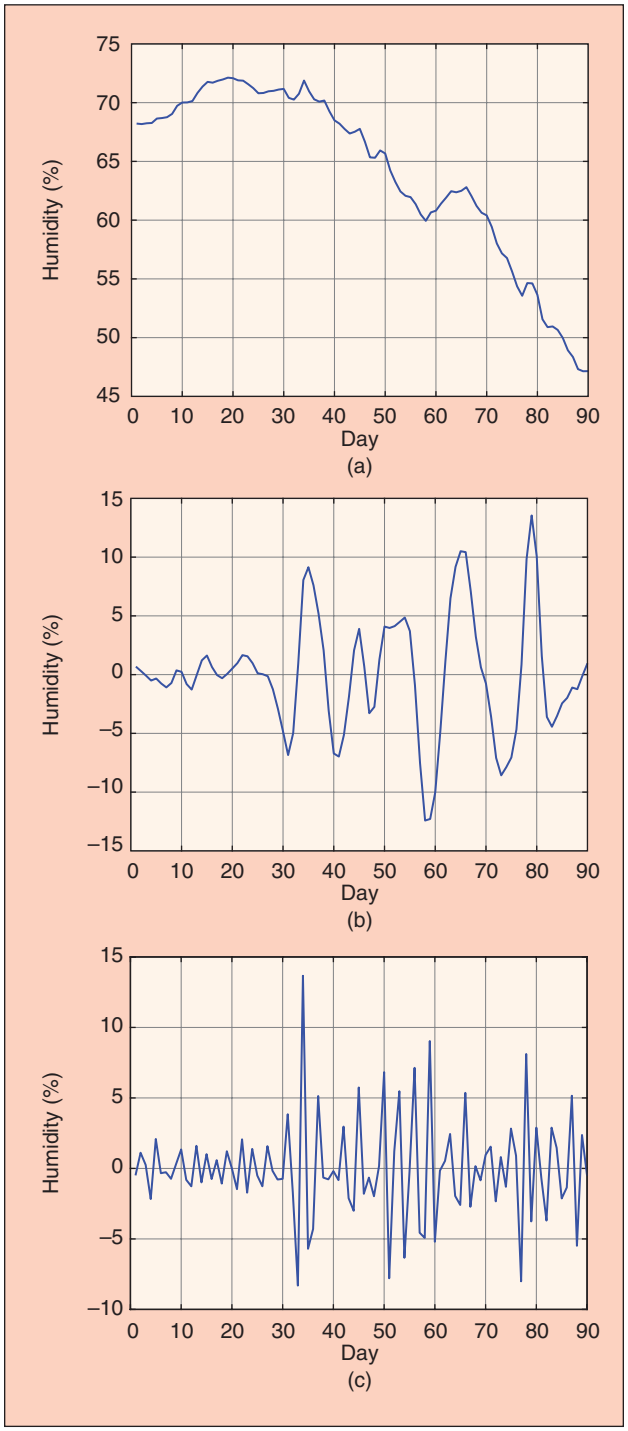


FIGURE 11 IMFs corresponding to the decomposition of humidity (H) time series of Tamil Nadu along with their center frequencies, (a) H-IMF$_1$, $\omega_1 = 0.0001$. (b) H-IMF$_2$, $\omega_2 = 0.0887$. (c) H-IMF$_3$, $\omega_3 = 0.3525$.

IV. Training Sequence Models

A set of multivariate stacked RNNs is developed to forecast the future values of each CC-IMF signals using the results of the previous section. The results of all predicted values of CC-IMFs are summed to obtain the forecast value of the CC signal one step forward in the future (Figure 5). A set of Recurrent Neural Networks (RNNs) with Long Short Term Memory (LSTM) cells is used because RNNs have been proven to be effective for forecasting time series [28], [29].

LSTM cells were initially designed to deal with vanishing and exploding gradient problems in sequence models [30]. The structure of LSTM is briefly explained in the next section.

[1]Note that one may argue that the first IMF of the temperature and humidity signals in any cases has a non-stationary trend in the long run and, therefore, suggest to consider them as features for training CC-IMF$_1$. The authors decided to ignore them to avoid masking the effect of non-stationary features which are believed to have more impact on CC-IMF$_1$.

TABLE VI KPSS test for the IMFs corresponding to the signal T of Maharashtra.

SIGNAL	LAG	P-VALUE	H	ST
T-IMF$_1$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
T-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
T-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

TABLE VII KPSS test for the IMFs corresponding to the signal H of Maharashtra.

SIGNAL	LAG	P-VALUE	H	ST
H-IMF$_1$	7, 9, 11	0.05, 0.09, 0.10	0, 0, 0	✓
H-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
H-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

TABLE VIII KPSS test for the IMFs corresponding to the signal T of Tamil Nadu.

SIGNAL	LAG	P-VALUE	H	ST
T-IMF$_1$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
T-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
T-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

TABLE IX KPSS test for the IMFs corresponding to the signal H of Tamil Nadu.

SIGNAL	LAG	P-VALUE	H	ST
H-IMF$_1$	7, 9, 11	0.01, 0.01, 0.02	1, 1, 1	×
H-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
H-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

A. Long Short Term Memory (LSTM) Cells

An LSTM unit consists of three gates (i.e., update, forget, and output gates) and three cells (i.e., input, memory, and update cells). The memory cell at time t is updated using a candidate value $\tilde{c}^{<t>}$, which is calculated using the output value at time $t-1$, i.e., $a^{<t-1>}$, and input value at time t, i.e., $x^{<t>}$, through the equation

$$\tilde{c}^{<t>} = \tanh\left(W_c\left[a^{<t-1>}, x^{<t>}\right] + b_c\right) \tag{6}$$

where $\tanh(.)$ is the hyperbolic tangent activation function, and W_c and b_c represent the matrix of parameters and biased vector of the memory cell, respectively. The value of the memory cell $c^{<t>}$ is then updated using the candidate value $\tilde{c}^{<t>}$ and the previous value $c^{<t-1>}$ through

$$c^{<t>} = \Gamma_u \odot \tilde{c}^{<t>} + \Gamma_f \odot c^{<t-1>} \tag{7}$$

where $\odot$ indicates element-wise multiplication. Γ_u and Γ_f are the values of the update and forget gates which are obtained from

$$\Gamma_u = \sigma\left(W_u\left[a^{<t-1>}, x^{<t>}\right] + b_u\right) \tag{8}$$

and

$$\Gamma_f = \sigma\left(W_f\left[a^{<t-1>}, x^{<t>}\right] + b_f\right) \tag{9}$$

in which $\sigma(.)$ is the sigmoid activation function, W_u and b_u are respectively the matrix of parameters and the bias vector corresponding to the update gate, and W_f and b_f are respectively the matrix of parameters and the bias vector corresponding to the forget gate.

The output value of the LSTM unit at time t is

$$a^{<t>} = \Gamma_o \odot \tanh\left(c^{<t>}\right) \tag{10}$$

where Γ_o is the value of the output gate which itself is

$$\Gamma_o = \sigma\left(W_o\left[a^{<t-1>}, x^{<t>}\right] + b_o\right) \tag{11}$$

in which W_o and b_o are respectively the matrix of parameters and the bias vector corresponding to the output gate. Figure 12 shows an LSTM unit.

A multivariate RNN architecture is used in this paper, which takes multiple features as input, and outputs the predicted value. Two different architectures are used, one for training CC-IMF$_1$, and another for training CC-IMF$_2$ and CC-IMF$_3$ separately. The architecture of the stacked RNN corresponding to the forecasting future value of CC-IMF$_1$ is as follows:

1) a sequence input layer which accepts the number of inputs equal to the number of features. Regarding Maharashtra, there are two features for training CC-IMF$_1$: the signals CC-IMF$_1$ and TR at time $t-1$. For Tamil Nadu there are three features for training CC-IMF$_1$: the signals CC-IMF$_1$, H-IMF$_1$ and TR at time $t-1$. The value of the CC-IMF$_1$ at time t is the target value which needs to be predicted in both cases.
2) an LSTM layer with 50 units.
3) a dropout layer with the factor 0.6.
4) a fully connected layer with one output unit.

The architecture used for training CC-IMF$_2$ and CC-IMF$_3$ is as follows:

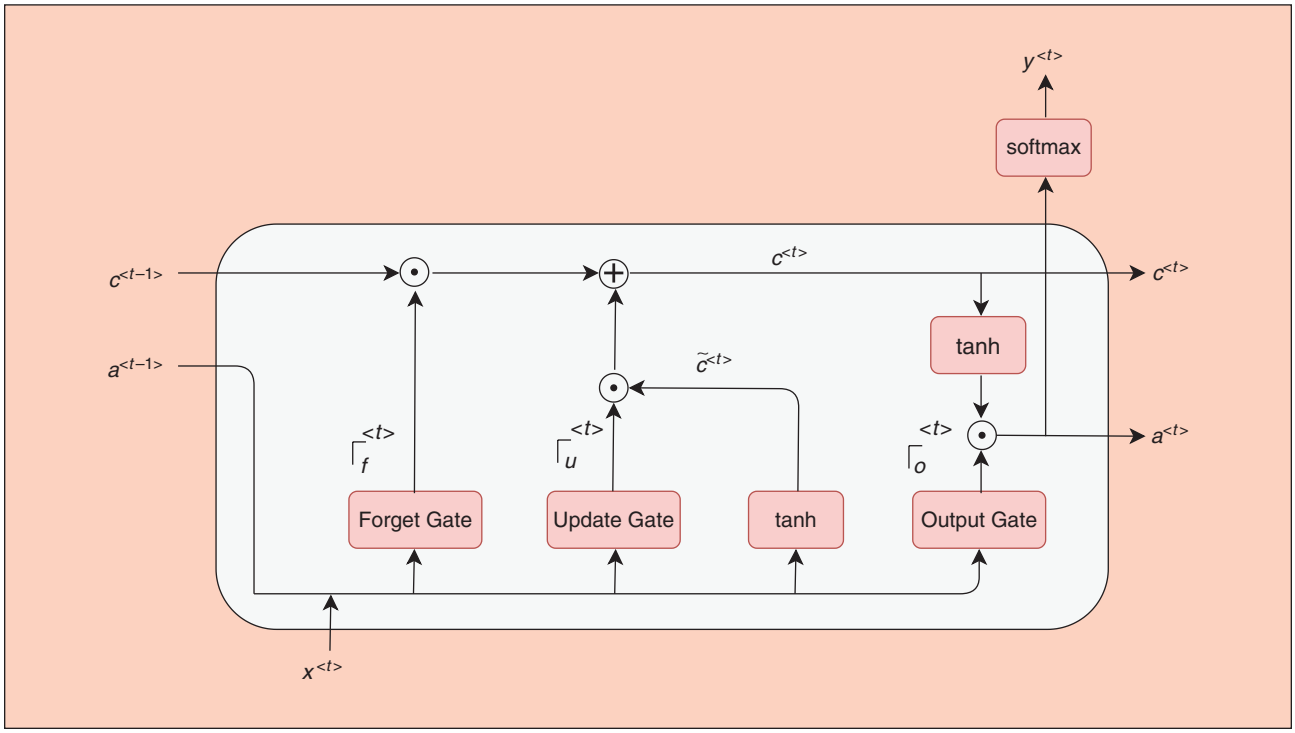


FIGURE 12 Visualisation of an LSTM unit. $y^{<t>}$ is the final output of an LSTM unit at time t which is computed by a softmax activation function.

1) a sequence input layer which accepts the number of inputs equal to the number of features. Note that there are three features for training CC-IMF$_2$ in both cases: the signals CC-IMF$_2$, T-IMF$_2$ and H-IMF$_2$ at time $t-1$. The target for the network in this case is the value of the signal CC-IMF$_2$ at time t. Similarly, there are three features for training CC-IMF$_3$ in both cases: the signals CC-IMF$_3$, T-IMF$_3$ and H-IMF$_3$ at time $t-1$. The target for the network is then the value of CC-IMF$_3$ at time t.
2) an LSTM layer with 200 units.
3) a fully connected layer with 200 units.
4) a dropout layer with the factor 0.6.

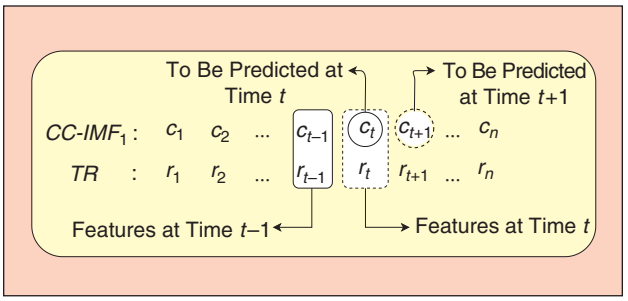


FIGURE 13 The value of the signal CC-IMF$_1$ at time t (c_t) is predicted using its value at time $t-1$ (c_{t-1}) and the value of the signal TR at time $t-1$ (r_{t-1}) as features.

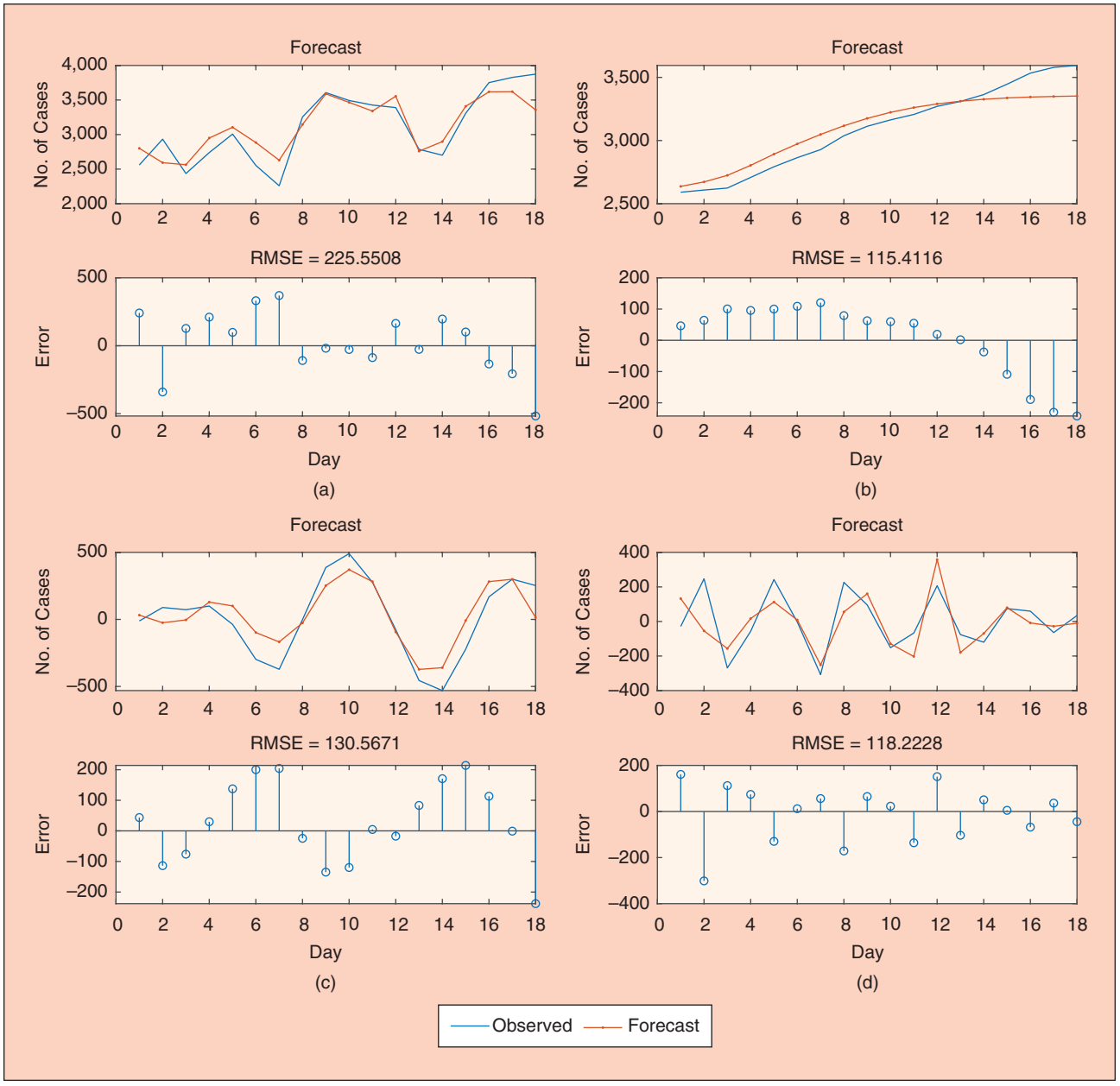


FIGURE 14 Predicted value and root mean square error (RMSE) of the number of COVID-19 cases forecast corresponding to each IMF of Maharashtra CC signal conducted on the test set. (a) CC; (b) CC-IMF$_1$; (c) CC-IMF$_2$; and (d) CC-IMF$_3$.

5) an LSTM layer with 50 units.
6) a fully connected layer with one output unit.

The dropout layers were adapted to prevent over-fitting on the training set. In fact, the dropout hyper-parameter indicates the probability of training a given node in a layer. It has the regularisation effect and prevents over-fitting on the training set [31].

B. Construction of the Training Set and the Test Set

Consider signals CC-IMF$_1$ and TR corresponding to Maharashtra. The proportion of 80% of data in both signals is used in the training set and the remainder (20%) is considered as the test set for validation purpose. To construct the training set, the value of the signal CC-IMF$_1$ and TR at time $t-1$ are considered as features to be fed into the constructed RNN, and the value of the signal CC-IMF$_1$ at time t is considered as the label or expected output of the RNN (Figure 13). The size of the training set is equal to 80% of the number of elements of signal CC-IMF$_1$, rounded to an integer (or equivalently the integer part of 80% of the number of elements of signal TR). The same procedure is followed in all other cases.

C. Setting the Options for the RNNs

Adam optimisation has been set in options as the optimisation method to update network weights in each iteration,

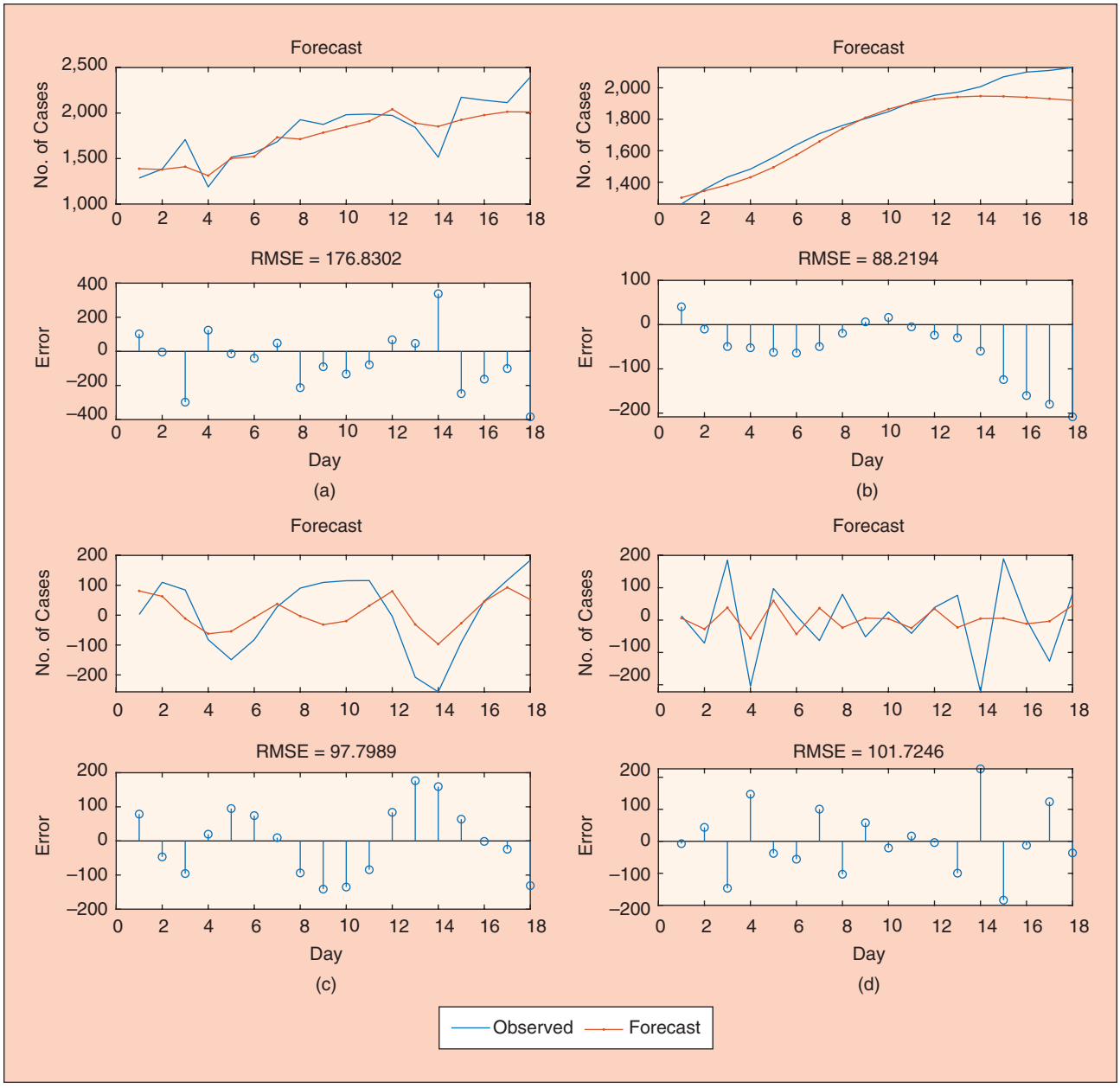


FIGURE 15 Predicted value and Root mean square error (RMSE) of the number of COVID-19 cases forecast corresponding to each IMF of Tamil Nadu CC signal conducted on the test set. (a) CC; (b) CC-IMF$_1$; (c) CC-IMF$_2$; and (d) CC-IMF$_3$.

as it is known to be an adaptive learning rate optimization algorithm designed specifically for training deep neural networks [32]. The learning rate was set initially at 0.005 and was decreased by a factor of 0.2 at every 200 epochs. The number of maximum epochs was chosen to be 1000. In order to avoid exploding gradients effect, a threshold 1 was set as the gradient threshold.

... the decomposed IMFs with similar center frequencies are used to train separate RNNs. Here we further work out the phase of each decomposed IMF corresponding to the CC, T, and H signals for both states using Gabor's complex analytical signal.

V. Results and Discussion

Figure 14 shows the predicted number of cases of COVID-19 for CC-IMF_1 (Figure 14(b)), CC-IMF_2 (Figure 14(c)) and CC-IMF_3 (Figure 14(d)) and the sum of all of them (Figure 14(a)) for the state of Maharashtra conducted on the test set. The figures show the Root Mean Square Error (RMSE) corresponding to each case. The results show that the model can predict the future number of cases within an acceptable range of error. However, Figure 14(b) shows that the predicted value of the signal deviates from its expected value at the right end of the signal. This is likely due to the end effect arising from the spline method used in the smoothing procedure to smooth the feature TR. This effect is also evident in Figure 2(c). As can be seen in the figure, the right end of the signal is slightly tilted downward whereas this is not the case in the original signal of Figure 2(a). Therefore, one may argue that smoothing the

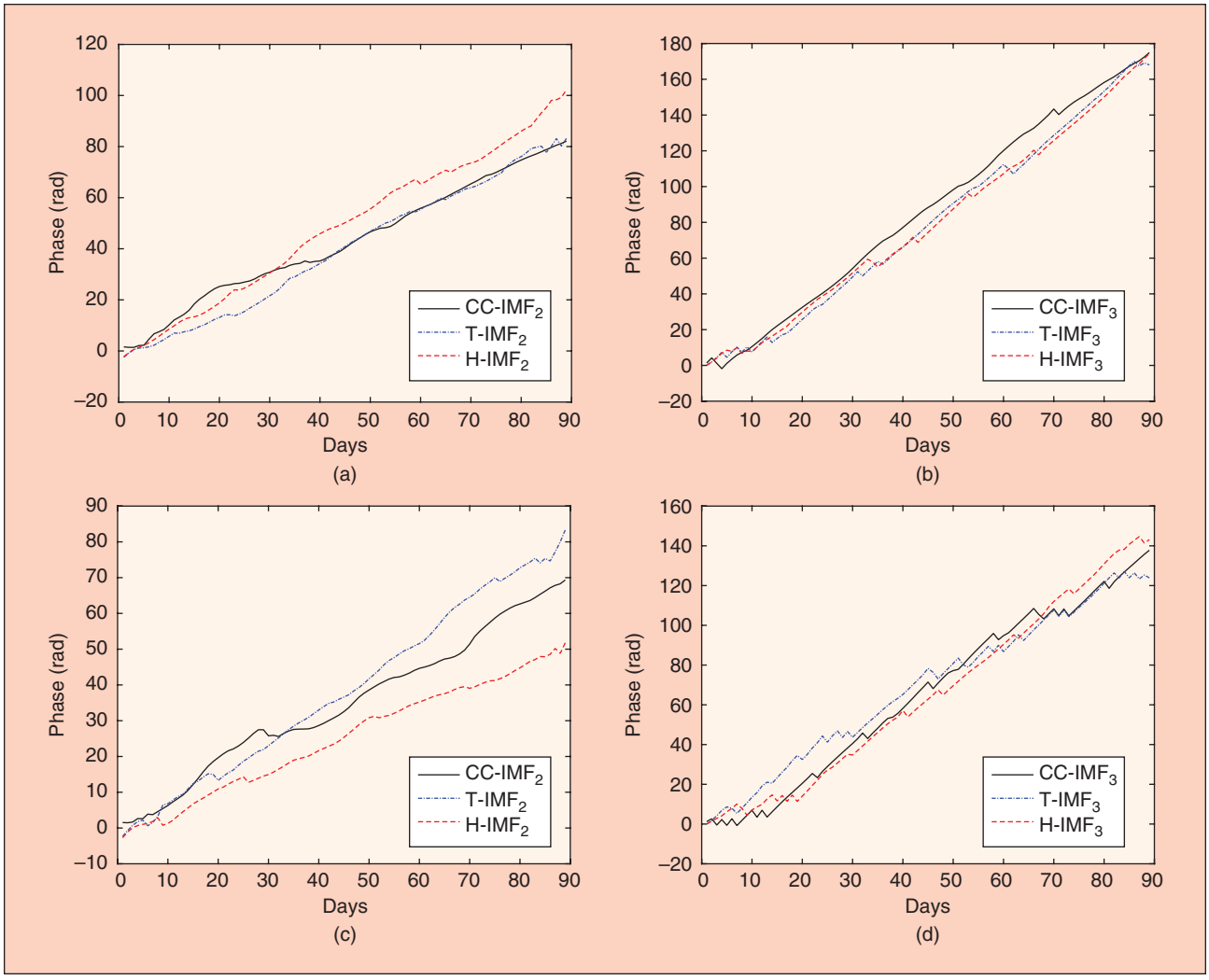


FIGURE 16 Unwraped phase corresponding to the IMFs of CC, T and H signals of states Maharashtra and Tamil Nadu. (a) Maharashtra (IMF_2). (b) Maharashtra (IMF_3). (c) Tamil Nadu (IMF_2). (d) Tamil Nadu (IMF_3).

signal TR is not beneficial. However, the prediction results have been boosted when the TR signal was smoothed.

The same procedure has been followed to train RNNs to predict the CC signal corresponding to Tamil Nadu. The results of the trained RNN on the test set is presented in Figure 15. The same effect of smoothing the TR signal is evident in Figure 15(a). The second and third modes of the CC signal are not as accurate as those of Maharashtra. The reason is that we conformed to the same architecture which was developed initially for Maharashtra. In the following we discuss this in more details.

We first look into the mean absolute percentage error (MAPE) corresponding to predictions for both cases, in order to compare the precision of the two different forecast problems with one another. The MAPE is calculated as

$$\text{MAPE} = \text{mean}\left(100 \times \left|\frac{p_t - y_t}{y_t}\right|\right) \tag{12}$$

where p_t and y_t represent respectively the predicted and observed values of the time series. The MAPEs for Maharashtra and Tamil Nadu are 6.23% and 7.77%, respectively.

As explained in Section III-B, the decomposed IMFs with similar center frequencies are used to train separate RNNs. Here we further work out the phase of each decomposed IMF corresponding to the CC, T, and H signals for both states using Gabor's complex analytical signal $X_a(t)$ [33] which is defined as

$$X_a(t) = X(t) + j\hat{X}(t), \tag{13}$$

where $X(t)$ and $\hat{X}(t)$ are respectively the original signal and its Hilbert transform. One can obtain the instantaneous phase of each band IMF as follows,

$$\phi(t) = \tan^{-1}\left(\frac{\hat{X}(t)}{X(t)}\right). \tag{14}$$

Figure 16 shows the obtained unwrapped phase of the IMFs corresponding to the aforementioned signals for both states (cf. `unwrap()` in Matlab). From Figures 16(a) and 16(b), the phase of the IMFs corresponding to the CC, H, and T signals of Maharashtra are more synchronised compared to those of

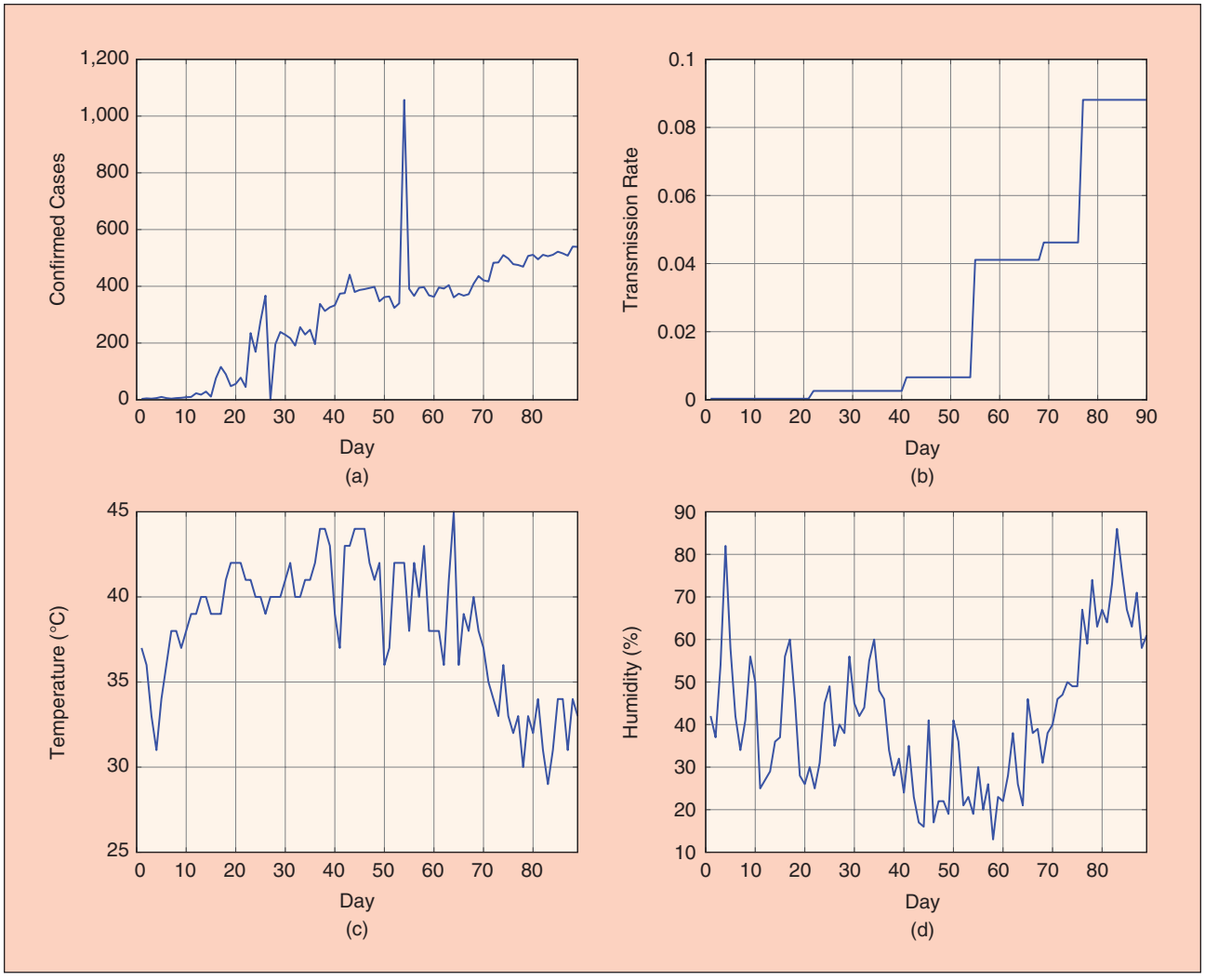


FIGURE 17 Signals (a) CC; (b) TR; (c) T; and (d) H corresponding to the state Gujarat.

Tamil Nadu (Figures 16(c) and 16(d)). This suggests a more complex dependency among IMFs of these signals corresponding to Tamil Nadu compared with Maharashtra. One way of achieving more accuracy in prediction in the case of Tamil Nadu is to use a deeper RNN architecture. However, in order to avoid over-fitting, either more data or a more severe regularisation strategy has to be exploited.

A further example is now investigated, corresponding to the state Gujarat in India where the number of cases is smaller. The data from this state is of interest particularly due to an outlier presenting at around day 54 (Figure 17(a)). The transmission rate TR of Figure 17(b) has been smoothed using the technique proposed in Section II. Also, all the signals CC, T, and H are decomposed using VMD (Section III-B), and their stationary and non-stationary parts are grouped and used for training the RNNs as discussed respectively in Sections III-B and IV. Figure 18 shows the final results of the prediction process. A satisfactory value of MAPE = 4.68% is obtained, which further confirms the applicability of the proposed technique.

VI. Conclusion

A systematic procedure to derive features for training RNNs to forecast the future number of confirmed cases of COVID-19

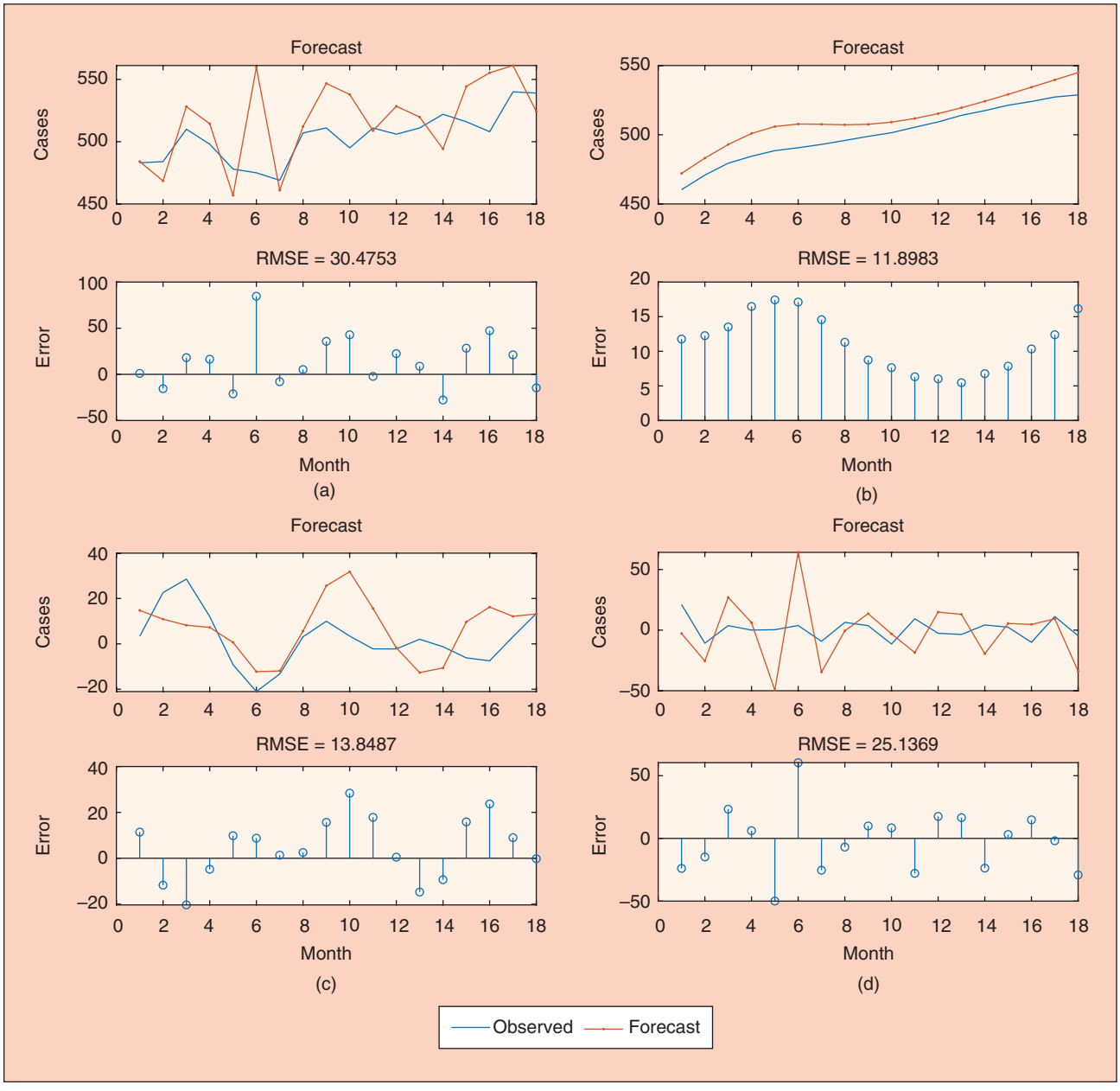


FIGURE 18 Predicted value and Root mean square error (RMSE) of the number of COVID-19 cases forecast corresponding to each IMF of Gujarat CC signal conducted on the test set (MAPE = 4.68%). (a) CC; (b) CC-IMF$_1$; (c) CC-IMF$_2$; and (d) CC-IMF$_3$.

in three states of India is proposed. Based on the literature review, the number of confirmed cases of COVID-19 is correlated with both temperature and humidity [8]–[11]. Therefore, both of these meteorological parameters are considered as features in training RNNs. Also, an equation proposed in [20] is used to calculate the transmission rates corresponding to each lockdown phase. As such, temperature, humidity and transmission rate have been used as features in this paper.

We conclude that specifying a soft transmission rate by smoothing the obtained step function can improve the prediction results. Moreover, compatible modes of signals were systematically derived, and it was found that training those with similar center frequency in separate RNNs improved the predictions. We collected the information from both outbreaks and available meteorological parameters to construct a model for predicting the future number of confirmed cases of COVID-19. However, one needs to take the following into account when predicting the future occurrence of COVID-19 using the proposed model:

1) The future value for transmission rates corresponding to a set of plausible lockdown phases may be approximated as those obtained from the previous lockdown stages.
2) The forecast value of temperature and humidity are usually available for some successive following days and can be used as features in the trained RNNs.

We have also shown through decomposing CC, T, and H signals into their modes using VMD that there are similar modes with close center frequencies in all of these signals. Although this confirms the effect of the temperature and humidity on the number of confirmed cases, one needs to look more carefully into the phase of the similar modes to unfold these dependencies more systematically. This issue contains sufficient merit to warrant independent research and can be a subject of future work.

Finally, the proposed procedure can provide insight into systematically forecasting the future number of COVID-19 cases, considering other factors affecting its spread in the community, which may include health policy, mask usage rate, and wind speed. As the method has been shown to be successful when applied to different Indian states that have quite different meteorological dynamics, it could be applied to other countries, especially if extended to additional relevant factors. The method proposed in this paper can also be used for other time series forecasting problems when complex signals are used as features.

References

[1] C. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet,* vol. 395, no. 10223, pp. 497–506, 2020. doi: 10.1016/S0140-6736(20)30183-5.

[2] L. Wang, Y. Wang, D. Ye, and Q. Liu, "A review of the 2019 novel coronavirus (COVID-19) based on current evidence," *Int. J. Antimicrob. Agent*, p. 105,948, 2020. doi: 10.1016/j.ijantimicag.2020.105948.

[3] C. I. Jarvis et al., "Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK," *BMC Med.*, vol. 18, p. 124, 2020. doi: 10.1186/s12916-020-01597-8.

[4] H. Lau et al., "The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China," *J. Travel Med.*, vol. 27, no. 3, p. taaa037, 2020. doi: 10.1093/jtm/taaa037.

[5] F. E. Alvarez, D. Argente, and F. Lippi, "A simple planning problem for covid-19 lockdown," National Bureau of Economic Research, Tech. Rep., 2020.

[6] S. Gupta, G. S. Raghuwanshi, and A. Chanda, "Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020," *Sci. Total Environ.*, p. 138,860, 2020. doi: 10.1016/j.scitotenv.2020.138860.

[7] H. Qi et al., "COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis," *Sci. Total Environ.*, p. 138,778, 2020. doi: 10.1016/j.scitotenv.2020.138778.

[8] B. Oliveiros, L. Caramelo, N. C. Ferreira, and F. Caramelo, "Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases," *medRxiv*, 2020.

[9] J. Wang, K. Tang, K. Feng, and W. Lv, "High temperature and high humidity reduce the transmission of COVID-19," 2020.

[10] A. Auler, F. Cássaro, V. da Silva, and L. Pires, "Evidence that high temperatures and intermediate relative humidity might favor the spread of COVID-19 in tropical climate: A case study for the most affected Brazilian cities," *Sci. Total Environ.*, p. 139,090, 2020. doi: 10.1016/j.scitotenv.2020.139090.

[11] J. Demongeot, Y. Flet-Berliac, and H. Seligmann, "Temperature decreases spread parameters of the new COVID-19 case dynamics," *Biology*, vol. 9, no. 5, p. 94, 2020. doi: 10.3390/biology9050094.

[12] Y. Ma et al., "Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China," *Sci. Total Environ.*, p. 138,226, 2020. doi: 10.1016/j.scitotenv.2020.138226.

[13] A. Tomar and N. Gupta, "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures," *Sci. Total Environ.*, p. 138,762, 2020. doi: 10.1016/j.scitotenv.2020.138762.

[14] D. Fanelli and F. Piazza, "Analysis and forecast of COVID-19 spreading in China, Italy and France," *Chaos, Solitons Fractal*, vol. 134, p. 109,761, 2020. doi: 10.1016/j.chaos.2020.109761.

[15] S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, and M. Prokopenko, "Modelling transmission and control of the COVID-19 pandemic in Australia," 2020, arXiv:2003.10218.

[16] R. Salgotra, G. Mostafa, and A. H. Gandomi, "Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming," *Chaos, Solitons Fractal*, p. 109,945, 2020. doi: 10.1016/j.chaos.2020.109945.

[17] R. Salgotra and A. H. Gandomi, "Time series analysis of the COVID-19 pandemic in Australia using genetic programming," in *Data Science for COVID-19*. Amsterdam, The Netherlands: Elsevier, 2020.

[18] T. Sardar, S. S. Nadim, and J. Chattopadhyay, "Assessment of 21 days lockdown effect in some states and overall India: A predictive mathematical study on COVID-19 outbreak," 2020, arXiv:2004.03487.

[19] R. Salgotra, S. Singh, U. Singh, S. Saha, and A. H. Gandomi, "COVID-19: Time series datasets India versus World," 2020.

[20] C. Kirkeby, T. Halasa, M. Gussmann, N. Toft, and K. Græsbøll, "Methods for estimating disease transmission rates: Evaluating the precision of Poisson regression and two novel methods," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, 2017. doi: 10.1038/s41598-017-09209-x.

[21] D. Garcia, "Robust smoothing of gridded data in one and higher dimensions with missing values," *Comput. Statist. Data Anal.*, vol. 54, no. 4, pp. 1167–1178, 2010. doi: 10.1016/j.csda.2009.09.020.

[22] K. Zolna, P. B. Dao, W. J. Staszewski, and T. Barszcz, "Towards homoscedastic nonlinear cointegration for structural health monitoring," *Mech. Syst. Signal Process.*, vol. 75, pp. 94–108, 2016. doi: 10.1016/j.ymssp.2015.12.014.

[23] P. B. Dao and W. J. Staszewski, "Data normalisation for Lamb wave–based damage detection using cointegration: A case study with single-and multiple-temperature trends," *J. Intell. Mater. Syst. Struct.*, vol. 25, no. 7, pp. 845–857, 2014. doi: 10.1177/1045389X13512186.

[24] D. Kwiatkowski et al., "Testing the null hypothesis of stationarity against the alternative of a unit root," *J. Econom.*, vol. 54, nos. 1–3, pp. 159–178, 1992. doi: 10.1016/0304-4076(92)90104-Y.

[25] W. K. Newey and K. D. West, "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," National Bureau of Economic Research, Tech. Rep., 1986.

[26] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 531–544, 2014. doi: 10.1109/TSP.2013.2288675.

[27] D. Zosso, "Variational mode decomposition," Matlab Central File Exchange. Accessed: June 22, 2020. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/44765-variational-mode-decomposition

[28] Z. Zhao, W. Chen, X. Wu, P. C. Chen, and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intell. Transport Syst.*, vol. 11, no. 2, pp. 68–75, 2017. doi: 10.1049/iet-its.2016.0208.

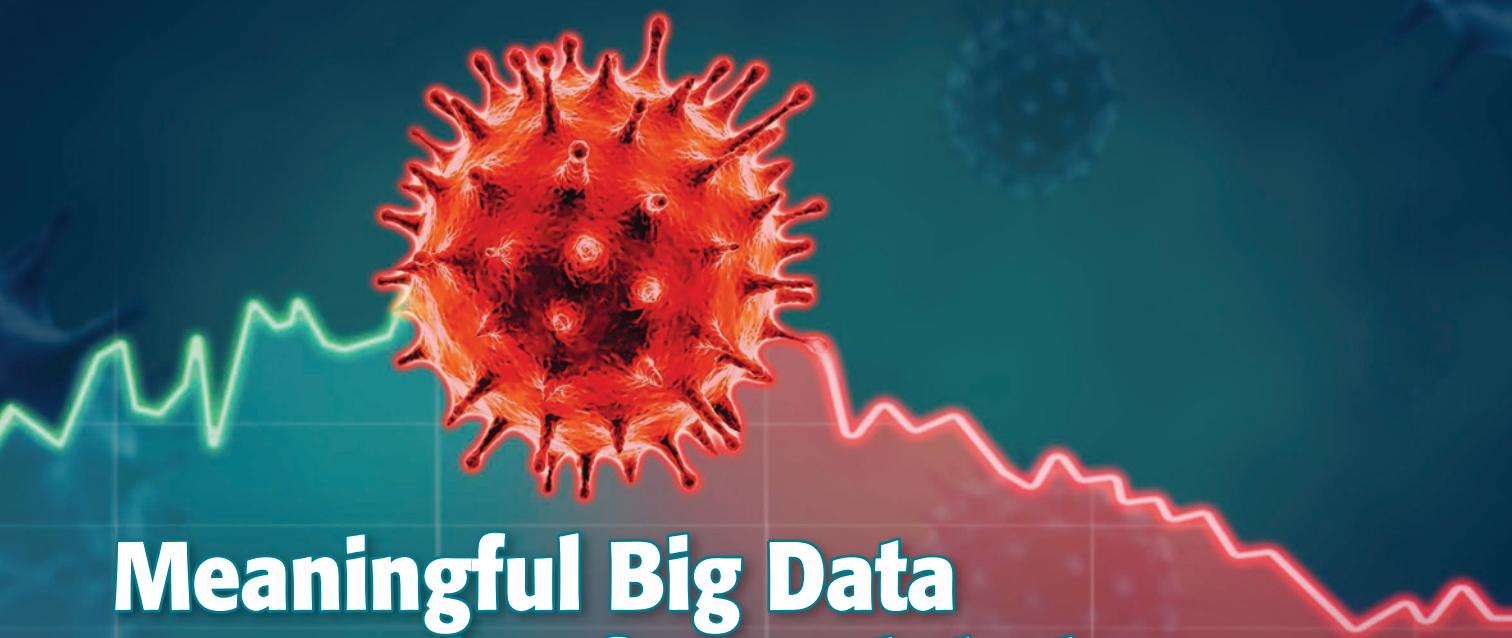
[29] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, 2017. doi: 10.1109/TSG.2017.2753802.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv:1412.6980.

[33] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Elect. Eng. III, Radio Commun. Eng.*, vol. 93, no. 26, pp. 429–441, 1946. doi: 10.1049/ji-3-2.1946.0074.



Meaningful Big Data Integration for a Global COVID-19 Strategy

Joao Pita Costa, Marko Grobelnik, Flavio Fuart, and Luka Stopar
Quintelligence & Jozef Stefan Institute, SLOVENIA

Gorka Epelde
Vicomtech & Biodonostia, SPAIN

Scott Fischaber
Analytics Engines, UK

Piotr Poliwoda
IBM, IRELAND

Debbie Rankin, Jonathan Wallace, Michaela Black, Raymond Bond, and Maurice Mulvenna
Ulster University, UK

Dale Weston
Public Health England, UK

Paul Carlin
Open University, UK

Roberto Bilbao
BIOEF, SPAIN

Gorana Nikolic, Xi Shi, and Bart De Moor
KU Leuven, BELGIUM

Minna Pikkarainen and Jarmo Pääkkönen
University of Oulu, FINLAND

Anthony Staines, Regina Connolly, and Paul Davis
Dublin City University, IRELAND

***Abstract*—**With the rapid spread of the COVID-19 pandemic, the novel Meaningful Integration of Data Analytics and Services (MIDAS) platform quickly demonstrates its value, relevance and transferability to this new global crisis. The MIDAS platform enables the connection of a large number of isolated heterogeneous data sources, and combines rich datasets including open and social data, ingesting and preparing these for the application of analytics, monitoring and research tools. These platforms will assist public health authorities in: (i) better understanding the disease and its impact; (ii) monitoring the different aspects of the evolution of the pandemic across a diverse range of groups; (iii) contributing to improved resilience against the impacts of this global crisis; and (iv) enhancing preparedness for future public health emergencies. The model of governance and ethical review, incorporated and defined within MIDAS, also addresses the complex privacy and ethical issues that the developing pandemic has highlighted, allowing oversight and scrutiny of more and richer data sources by users of the system.

Introduction

The COVID-19 outbreak was declared a Public Health Emergency of International Concern by the World Health Organization (WHO) on 30 January 2020 [35]. With its rapid expansion, health stakeholders are keen to find technologies to monitor and combat the spread and impact of the disease. Along with this, the world has seen the multiplication of surveillance efforts to monitor the epidemic by official global health agencies such as the WHO and the European Centre for Disease Control (ECDC). Businesses

Digital Object Identifier 10.1109/MCI.2020.3019898
Date of current version: 14 October 2020

Corresponding Author: Joao Pita Costa (joao.pitacosta@ijs.si).

The MIDAS intelligent system unleashes the potential of a range of analytics to explore the confidential and sensitive data owned by health authorities within a safe and secure environment.

and research institutions have rapidly refocused their monitoring platforms to assess the impact of this common threat. Examples include the WHO COVID-19 Dashboard by ArcGIS [2], the Coronaviruswatch platform by the UNESCO Research Centre for Artificial Intelligence (IRCAI) [34] and RavenPack's Coronavirus News Monitor [29], alongside many other global and local initiatives. Such platforms enable patterns and trends in disease behavior monitored throughout the population. They can also inform public health measures to reduce transmission and allow the impact of these to be assessed.

The novel MIDAS public health platform [21], presented in this paper and shown in Figure 1, goes a step beyond existing platforms, particularly in responding to the coronavirus pandemic, by providing its users in public health authorities with insightful information from a combination of sources including world news, social media and published science, alongside local public health data from the health institution itself and other relevant data sources. The MIDAS platform was co-created with academia, industry, and crucially, health professionals, policy-makers, public health authorities and citizens, to align innovative technology with concrete public health priorities and workflows [4]. It was developed to connect typically heterogeneous, isolated health data, and integrate it with additional social data sources, to enable the application of advanced data analytics techniques and visual analytics tools to support policy decision-making in public health institutes across Europe [7]. Further details, code repositories and demonstrators are openly available at www.midasproject.eu.

The MIDAS intelligent system unleashes the potential of a range of analytics to explore the confidential and sensitive data owned by health authorities within a safe and secure environment. The system achieves this through data visualization widgets, encapsulating the analytics processes, within user-customizable dashboards that summarize each user's selected output priorities, such as monitoring mental health or child obesity across the population within the COVID-19 confinement restrictions. The main contribution of this paper is the description of the MIDAS platform resources, how these were rapidly refocused to address COVID-19—related public health priorities, and how they can help researchers and public health policy-makers achieve a deeper and more complete understanding of the SARS-COV-2 disease.

The MIDAS integrated platform consists of software developed for data ingestion, processing, analytics and visualization. These services are deployed locally within health authority sites. The platform offers a bespoke co-created collection of software and services enabling secure access to and exploration of sensitive health data as well as open and social data brought in from external applications, via the MIDAS dashboard. A common user authentication service allows single sign-on across the services. Data downloads are prohibited to mitigate privacy concerns and risk. If a MIDAS user, such as a policy-maker, requires access to the raw and/or prepared data, they must make a data access request to the data gatekeepers to

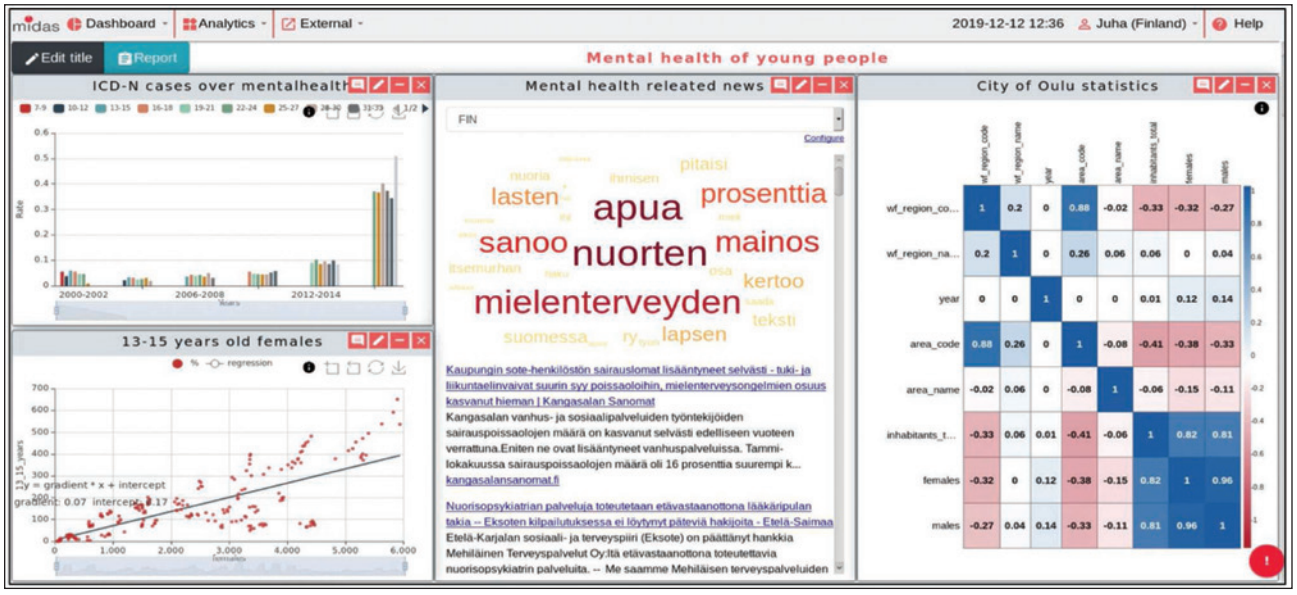


FIGURE 1 The MIDAS platform dashboard refocused to explore COVID-19 over proprietary data, worldwide news, social media campaigns and biomedical research.

secure a data access agreement (DAA) in compliance with relevant ethics and privacy legislation.

The core data platform is based on Apache HDFS, Apache Spark, and Apache Hive. In the system architecture diagram in Figure 2, the light grey boxes indicate user-facing web applications, whilst the dark grey boxes represent services and applications required for the platform that are not accessible to the end user. The MIDAS Platform runs across the health authority network (in blue) as well as externally (in orange). Data can be pulled into the platform from private or public data sources. The MIDAS platform has been successfully tested, validated [6] and evaluated by four pilot sites in public health institutes throughout Europe (Finland, Northern Ireland, Ireland and the Basque Region) realizing success across different priority policy areas. Each of these pilot study priorities is closely related to COVID-19, either in terms of the associated risk (e.g. diabetes, obesity and the ageing population), the impact of confinement restrictions (e.g. measures of mental health and childhood obesity), to a wide range of accessible social and spatial variables, such as measures of deprivation, social isolation, and access to and use of healthcare services [3]. From the outset MIDAS embedded a user-centered, co-creative, agile approach to its design and development, engaging with a wide-range of stakeholders, having the needs of each health policy pilot site driving the development in relation to both data analytics and visualizations for their individually selected health policy focus [5].

Making Data Meaningful Within a Local Context

The ability to enhance the usefulness of data located within health policy sites by integrating it with other global data sources provides significant potential value to the user. MIDAS provides tools such as a cross-filter analytics dashboard, an easy to use and understand interactive map visualization that updates its content automatically when the user selects different countries on the displayed map (as shown in Figure 3). This can be a useful tool for the initial exploration of many types of data within a specific topic, e.g. COVID-19. The form of visualizations is not uniform and can be customized based on the data type and research question (i.e. they can include line charts, bar charts, maps, tables, and other plots). The categorical variables used in the cross-filter are predefined. Users can select subgroups for analysis and the graphs will automatically update with data from the selected subgroups. An example of the cross-filter tool in the MIDAS platform is shown in Figure 3, created for the Irish pilot. The categorical variables are Gender, Region, and Age Group, and users can select subgroups either by clicking the buttons on the panel, selecting the region on the map, or the subgroups in the line chart or bar chart.

The application of the cross-filter tool can present results more intuitively on the map and can make cross-group comparisons easier through interaction with the user. It is not necessary to have the same components as the example of the Irish pilot shown in Figure 3, however all of the components can be replaced according to the categorical variables and data types. The flexibility of the tools developed for the MIDAS platform allows the user to apply the most suitable forms of visualization to their data.

To help users follow the results, emergency ICD-10 WHO codes U07.1 and U07.02 from [36] have been assigned to diagnose an identified presence of the virus. For example, the

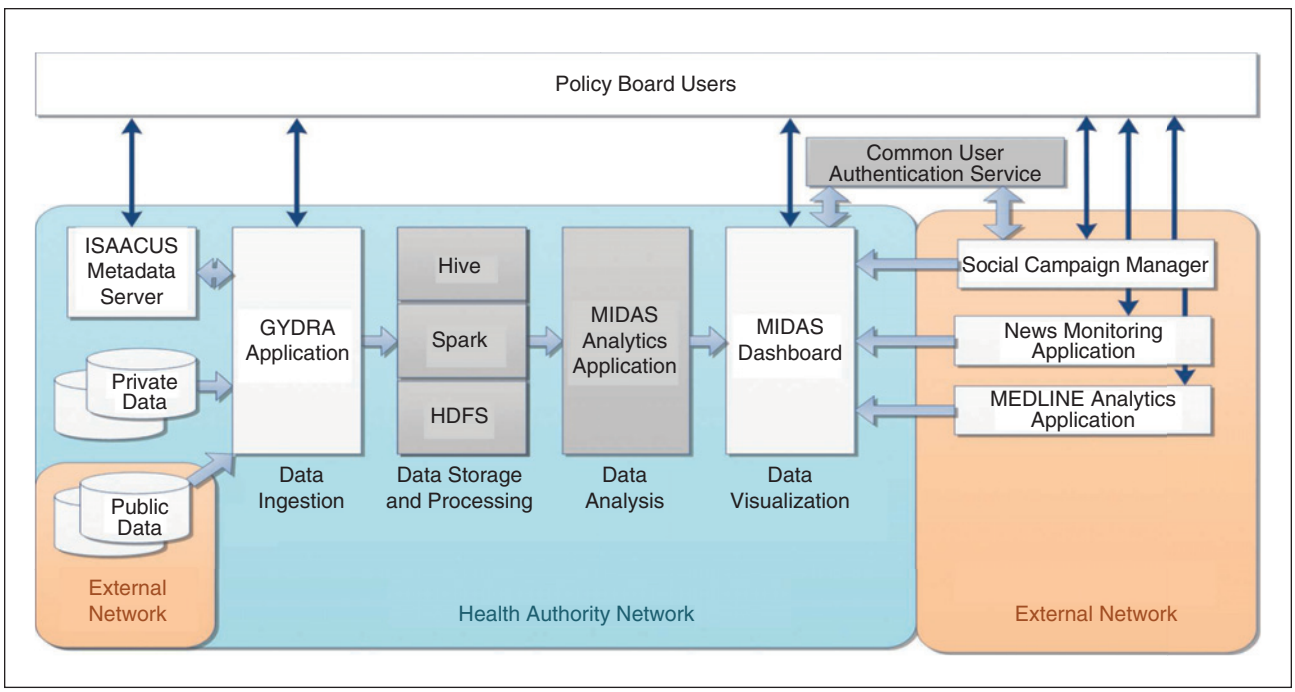


FIGURE 2 MIDAS platform architecture overview, designed to host Big Data to improve decision-making in public health.

Irish cross-filter tool, shown in Figure 3, allows the monitoring of diabetes cohorts and their Around the Clock prescription use by individual counties. With the MIDAS cross-filter tool it is possible to repurpose it to closely follow the COVID-19 codes and obtain a more accurate epidemiological overview over the regional county map.

Using the exploratory data analytics dashboards in the MIDAS platform, we have evidenced the known link [18] between diabetes and pneumonia outcomes. This insight from the available data is also highly useful in the response to COVID-19, considering that diabetes is a risk factor for the progression and poorer prognosis for COVID-19 patients [13], [16]. Using the MIDAS platform to understand the cohort of the population with diabetes in Ireland enables more targeted responses to the COVID-19 pandemic, for example, targeting local areas, which have high prevalence or higher historical hospitalizations with comorbidities, with specific messaging. This permits a more effective and rapid service delivery and may allow the identification of potential COVID-19 hot spots in advance.

The MIDAS pilot in the Basque Region, focusing on the topic of childhood obesity, is useful and important given the known associations of obesity in children with the emergence of comorbidities (e.g. diabetes and hypertension). Moreover, the Basque public health authorities are interested in monitoring the effect of the COVID-19 pandemic on childhood obesity using the MIDAS tools. Behavioral changes during lockdown in children and adolescents with obesity participating in a longitudinal observational study in Italy have been published recently [24]. No changes in vegetable intake were reported during lockdown, whilst in contrast, potato chip, red meat, and sugary drink intake increased significantly. Time spent in sports activities decreased, sleep-time increased and screen time increased. Taking into account the severity of the COVID-19 cases identified in people diagnosed with obesity, the MIDAS platform could be utilized to enhance our understanding of the adverse collateral effects and lasting impact of the COVID-19 pandemic lockdown on the adiposity level of adolescents. This includes the evolution of childhood obesity during the pandemic, and the evolution and adoption of policies based on the monitoring and visualization of a rapidly changing context.

The topic of mental health was studied in the context of the MIDAS pilot in Finland. This pilot ran a social media campaign in 2019 through the Twitter chatbot capabilities within the MIDAS platform (described in a later section of this paper). The existing data, ingested into the MIDAS platform previously and located within the policy sites, can be useful in this new COVID-19 framework, helping public health authorities to gain a better understanding of the health scenario over this difficult to assess topic, complemented by the news published around it. The additional ingestion of online participatory surveillance systems can contribute to a better understanding of the impact of the pandemic on mental health.

The COVID-19 situation has increased global concerns in relation to the mental wellbeing of children and young people. In Finland, the MIDAS pilot focused on the prevention of mental health problems in young people. As an example, within this context, the MIDAS platform could be used as a tool to support decision-making related to the organization of mental healthcare services and resources for young people. The COVID-19 pandemic has created a situation whereby long isolation periods at home and difficult economic situations are commonplace. This has raised concern amongst decision-makers in relation to child abuse and domestic

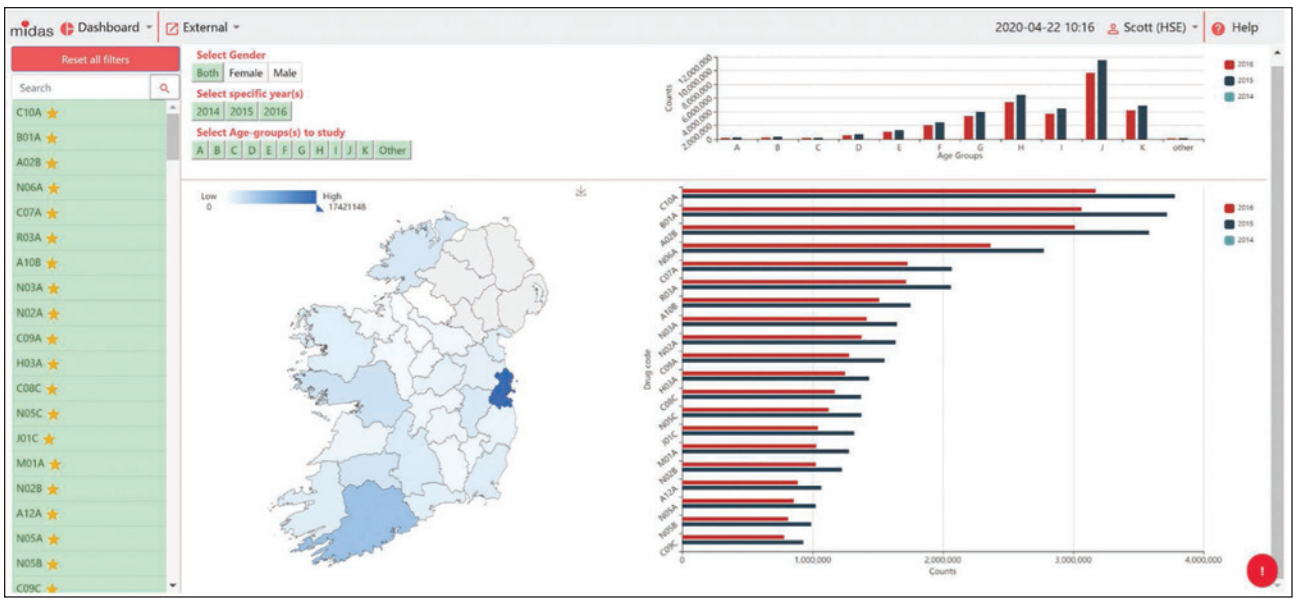


FIGURE 3 The cross-filtering at the MIDAS platform and its usage by the public health authorities in Ireland.

violence. This is an example of a complex and multi-level issue where a range of heterogeneous data is required at a policy-making level to enable better analysis and visualization of the situation. The hindering factor for the real use of the MIDAS platform in mental health cases in Finland is the current legislation often restricting the combined usage of the anonymized social and healthcare data collected on individuals in policy level decision making.

… the MIDAS platform could be utilized to enhance our understanding of the adverse collateral effects and lasting impact of the COVID-19 pandemic lockdown.

As an example of how the MIDAS platform could be used in the COVID-19 context, the lockdown effects on childhood obesity and mental health could be analyzed by updating the available longitudinal clinical with data covering the lockdown period, reusing existing data preparation techniques, or modifying them to allow the inclusion of new countries and health systems and reusing existing data analytics and visualization tools (on the period of interest). In the case of childhood obesity, these could focus on body mass index (BMI) and epidemiological analysis of new diabetes cases, covering person (gender, age-group), time (year) and location dimension (trust level and primary care unit level). Furthermore, the tool could be extended by plugging and ingesting new data sources such as physical activity captured by wearables, nutrition-related periodic questionnaires, or by alternatively adding grocery shopping aggregated data that, once integrated to comparable aggregation levels and described in metadata, could be analyzed side-by-side with current visualizations or further analyzed using statistical or machine learning technique.

Ingesting Useful Open Data Sources

The MIDAS platform includes heterogeneous datasets prepared and deployed in different pilot site locations, addressing their own specific challenges. These include city and government generated controlled datasets (i.e. health and social care data exports mainly at individual person level) and government open data (aggregated data) on air and water quality, national statistics (e.g. deprivation, education level or unemployment level per municipality), or city planning. These data sources are selected by each user to address various priority health policy questions across each pilot region.

The integration of different controlled public health data sources containing individual level data for the MIDAS pilot cases was completed by the data owners prior to loading these into the MIDAS platform, while the linking identifiers and datasets provided to the consortium were agreed to be provided on an anonymous basis. A different instance of the MIDAS platform was securely hosted at each policy site, to ensure the data owner retained control over the data being loaded. These heterogeneous datasets have been used to provide combined solutions in different sites, combining data at aggregated and individual level, by providing mappings at agreed location aggregation levels. Applying this process to the context of COVID-19, it is easy to map and analyze the relationship of clinical variables and open data indicators (i.e. COVID-19-related indicators such as number of cases, intensive care units or beds, and recovered people that have been published through different organizational open data agencies) at an aggregated location level (e.g. primary care unit or trust). The COVID-19 pandemic motivated the availability of diverse open datasets and indicators [1], presenting new opportunities (new sources for monitoring, modelling and forecasting the pandemic) and challenges (the need to correctly pre-process and integrate the data sources made available).

In the MIDAS platform the GYDRA Big data preparation tool (renamed from its initial in-memory processing version TAQIH) [30] has been developed for the preparation, ingestion and loading of the selected datasets. GYDRA has two main aspects: (i) an easy to use and interactive web-based interface (mimicking traditional data quality assessment and improvement flow) allowing non-technical users to use it; and (ii) data synchronization functionality to allow data owners and policy-makers to iteratively prepare and automatically deploy the prepared data to the analytics platform (relying on Apache Hive technology, and a defined metadata approach for the platform).

Within GYDRA's data preparation user interface items are placed from left to right following the usual iterative pipeline in exploratory data analysis. The "*General Stats and Features*" GYDRA sections provide global and detailed views of the data content, distribution and quality. The *"Missing Values"* section deals with the completeness of data. The *"Correlations"* section presents the correlations amongst variables, to help identify possible redundancies amongst variables or incoherent data. Finally, the *"Outliers"* section identifies outliers for each variable. Based on the insights identified during this analysis, a transformation pipeline can be configured to drop features and observations, handle missing values and outliers, or define operations to create new features from existing features or to change specific values. Within the MIDAS project each dataset from each pilot site has required a different preparation recipe. However, common tasks for each dataset include checking the number of columns per row (to avoid issues with separators being present in the content), merging data exported in chunks, format changing (e.g. for date fields loading), recoding of categorical values (after defining integration mappings), dropping features with few occurrences, dropping some meaningless outlying occurrences and creating new tables for specific analysis.

The metadata generated by the data synchronization functionality introduced above describes the data organization after

multi-resource data is ingested via the GYDRA tool and deployed to the analytics platform. This enables the automatic generation of generic data analysis and visualizations, as well as the development of specialized applications and machine learning models exploiting the secondary use of platform loaded data.

Figure 4 depicts the data ingestion, preparation and synchronization process of the GYDRA tool. Using the established data ingestion, preparation and synchronization methodology, it is straightforward to ingest, map and load COVID-19 related controlled and open datasets to the MIDAS platform. Additionally, the GYDRA tool relies on the interactive definition of dataset transformation pipelines which, once defined and refined, can be used to dynamically process and load partial and complete dataset updates. In this sense, the feature enables the ingestion of more dynamic data sources (dynamic in contrast to data export dumps provided at certain time periods by Health Care providers). In order to produce (re)usable COVID-19 data ingestions and mapping pipelines, analytics and visualizations within the MIDAS platform, it is necessary to work with the clinical and scientific community to achieve private and secure means to access data sources and agreed data models (as requested by [12]).

Moreover, the combination of the MIDAS developed GYDRA data preparation tool, alongside synthetic dataset generation strategies, can enable hospitals and healthcare providers, to: 1) refine and prepare their datasets (with the required metadata description), and; 2) share synthetically generated privacy-preserving datasets with the scientific community, that follow statistical patterns similar to the real data, and have proven to be reliable for training machine learning models [28]. These mechanisms would enable users to load a controlled dataset into the MIDAS platform and to develop in-house analytics, whilst simultaneously allowing the scientific community to develop AI models based on synthetic datasets that can later be fed back to the policy-makers through the MIDAS platform. This methodology provides a way to upscale and expedite the development of machine learning solutions through privacy preserving data sharing.

Extracting Insight from Published Biomedical Research

As the pandemic developed, MIDAS contributed to the many efforts to help biomedical researchers gain a better understanding of the disease (see examples in [14], [38] and [20]). To this end we have utilized the knowledge base MEDLINE [22] that serves the well-established biomedical search engine PubMed [23]. This open dataset stores structured information on more than 30 million records dating back to 1966. The comprehensive controlled vocabulary associated with MEDLINE—the MeSH Headings—delivers a functional system of indexing published biomedical science from journal articles and books. The MEDLINE articles are hand-annotated by humans with the established MeSH headings as health-related topics. These allow the user to explore a certain biomedical related topic (e.g. *"Biomarkers"* with the MeSH ID D015415), relying on curated information made available by the North American National Library of Medicine (NLM). The controlled vocabulary MeSH extends from 16 major health categories (covering topics such as anatomical terms, diseases, and drugs), each of which will be further

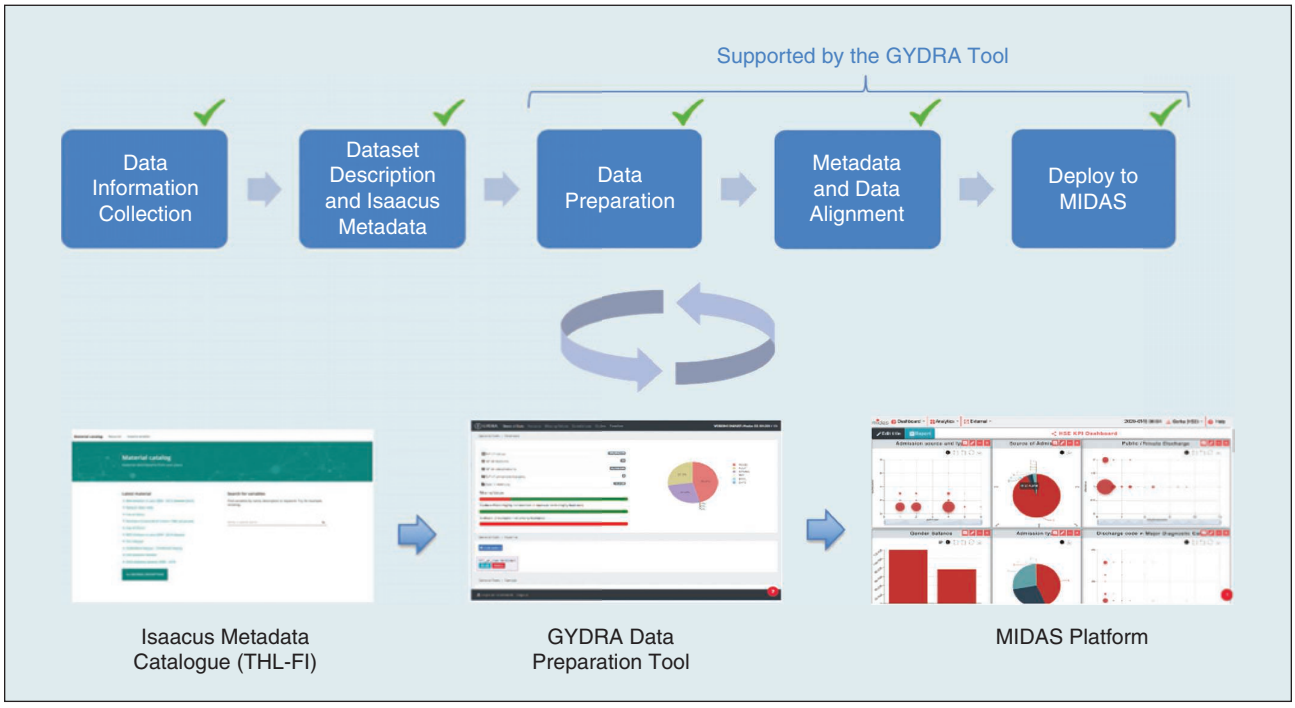


FIGURE 4 GYDRA tool-based data ingestion, preparation and synchronization with the MIDAS visualization and analytics platform.

distinguished from the most general to the most specific in up to 13 hierarchical depth levels. Although the recent introduction of the supplementary concept *"COVID-19"* on January 13, 2020, the MeSH heading "Coronavirus" was introduced in 1994 referring to the group of related viruses and prior known strains. These are located in the MeSH tree under the Coronaviridae family, introduced in 1999. MEDLINE includes 5976 research articles on coronavirus, relating to 4097 other health topics and 1706 substances. The articles that are hand-annotated with the MeSH class *"Coronavirus"* can help researchers better understand the new strain from the available scientific literature. With this in mind, the MIDAS platform offers an exploratory tool (see Figure 5) that allows the user to explore the published research through: (i) a query, based on keywords and operators, or an advanced query based on the syntax of the Lucene language; and (ii) a target pointer, which the user interacts by dragging it over the tag cloud to explore the results on the subtopics it relates to and to reprioritize the search results obtained. An example of such a precise syntax query (including two types of search categories—MeSH headings as health topics, or Chemicals)—*MeshHeadingList.desc:" Coronavirus" NOT ChemicalList.NameOfSubstance: "Viral Proteins"*—provides the user with, e.g. a subset of MEDLINE articles that are hand-annotated with the MeSH heading *"Coronavirus"* but are not labelled with the substance *"Viral Proteins."* The user can further explore the subset restricting it to labelled publications with the *"Biomarkers"* MeSH heading to explore new treatments in this specific context.

It is used to annotate scientific articles, news articles and reports relating to the COVID-19, allowing for the utilization of the MeSH headings as search topics …

The MIDAS platform also includes an exploratory dashboard that provides access to all MEDLINE records and enables users to explore these directly, and save samples based on queries. These are stored as JSON files in an elasticsearch based database, utilizing robust and well-established technology [11]. The external MIDAS MEDLINE explorer does not require the user to have expert technical knowledge and allows the average biomedical researcher to explore MEDLINE with further insight but little technical skills required. It also enables the user to rapidly build several data visualization modules (based on the Kibana open-source data visualization technology) that are easily configurable and based on templates, over the queried data (saved as a subset of records). It includes a variety of charts, tag clouds, heat maps and lists, as well as dashboards that integrate the created dynamic data visualization modules (see Figure 6). These enable MIDAS users to build and share dashboards that analyze published biomedical research on COVID-19 in relation to relevant topics for this study (similarly to diabetes, mental health etc.), and recover the biomedical articles relating to it. Moreover, this dashboard is served with a powerful API that allows the user to access and query the data from other systems. An extended instance of it enables easy ingestion of new articles, reports or news that can be annotated using a classifier, loaded, visualized and explored in the

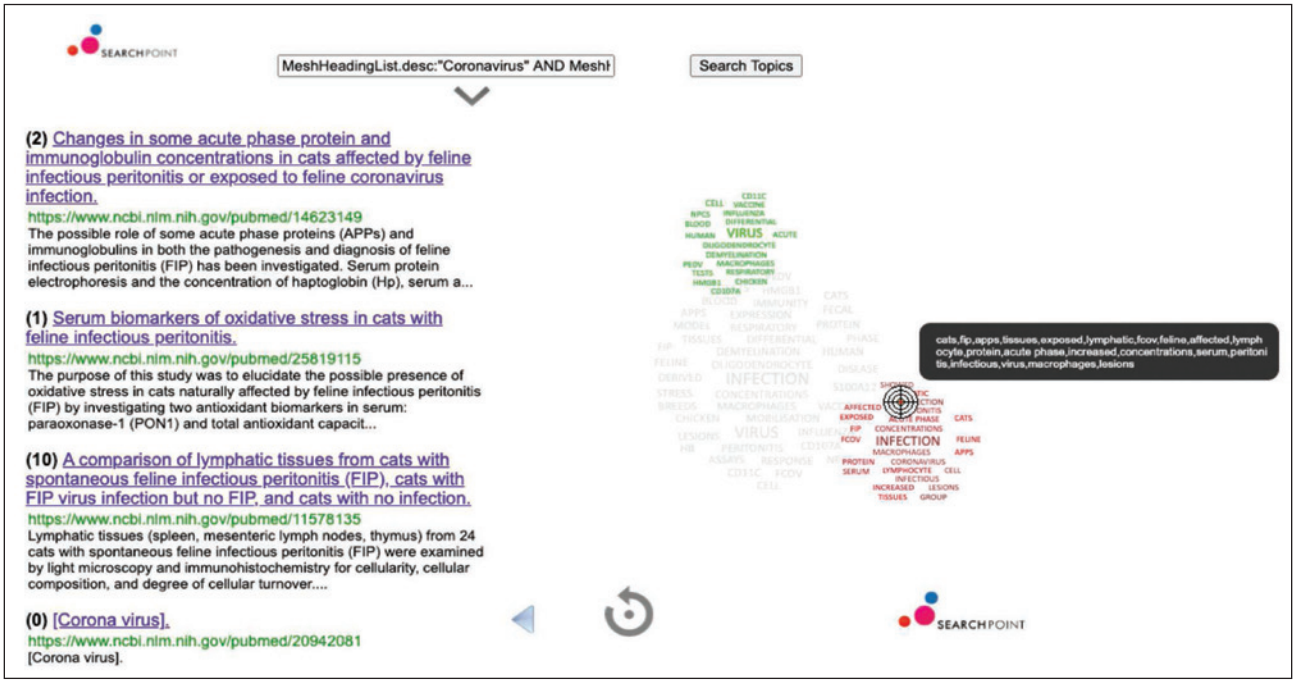


FIGURE 5 Exploring the biomarkers related to the Coronavirus in the published research using the MIDAS MEDLINE Explorer.

MEDLINE dashboard and explorer tools, and further mined through the available API.

One of the highly innovative technologies derived from the research carried out to build the MIDAS platform is the MeSH classifier [23]. It is an automated text classifier that has learned over human hand annotation of MeSH classes from more than 25 million biomedical articles (part of the corpus was used for evaluation) to perform the assignment of MeSH classes to any given snippet of text. It uses advanced text mining algorithms for this classification [19], and can classify any input text (including news, reports and health records) with this well-accepted health taxonomy. It is used to annotate scientific articles, news articles and reports relating to the COVID-19, allowing for the utilization of the MeSH headings as search topics over the corpus of these documents. It enhances the searchable information and allows for data visualization modules where the user can see the health topics (based on the most frequent MeSH classes) associated to the news over a query. This classifier was evaluated over: (i) scientific articles, part of the annotated MEDLINE dataset; and (ii) over news articles, hand annotated by some of the health experts in MIDAS on topics such as infectious diseases, diabetes, childhood obesity and mental health. The evaluation of the classifier resulted in an F1 measure of 0.43 in the MeSH tree depth level three for the classification of scientific articles, whilst F1 measures range between 0.55 to 0.85 for news articles in specific health domains (including diabetes, mental health and infectious diseases). The details will be published in [27]. The MeSH classifier offers a web portal and an API to enable a diversity of usages and integrations in other solutions. The web portal (accessible through the MIDAS COVID-19 toolset at the web portal *www.midasproject.eu/covid-19/*) provides the positioning of the MeSH categories that were assigned to the input text snippet, their similarity percentage and the MeSH tree branches to which the class belongs. This classifier allows us to generate useful metadata (based on the MeSH categories assigned to news articles, new research articles or medical reports) enabling its usage in the MEDLINE explorer and dashboard described previously. This explorer is served with an API that allows access to the structured Coronavirus dataset, and that can be enriched with other reports and annotated with the MeSH classifier. This allows researchers to leverage the existing knowledge generated in the current research.

Worldwide News Monitoring

The MIDAS news monitoring dashboard is fed by Newsfeed technology, collecting and analyzing more than 100 thousand news articles daily in real-time through the Event Registry technology, offering the MIDAS user insightful data visualizations to explore health-related news [17]. Since January 2020, the MIDAS news engine collected more than 13 million news articles on coronavirus-related topics across more than 60 languages. These included over 120 thousand articles on COVID-19 and diabetes, more than eight thousand articles on COVID-19 and retirement homes and elderly care, over 18 thousand articles on COVID-19 and obesity, more than 116 thousand articles on COVID-19 and mental health, and approximately 191 thousand articles on COVID-19 related to nursing. The news explorer within MIDAS allows the user to explore the overall sentiment of the news and the categories associated with it. From the total amount of collected coronavirus articles, approximately 36% have a positive sentiment and 0.69% relate to patient education whilst 0.71% relate to testing facilities. Recently, IRCAI released a worldwide news monitoring dashboard dedicated to COVID-19 based on the same news engine [30]. This general purpose health news monitoring dashboard exhibits the news on the epidemic outbreak in real-time and allows the reader to explore the information provided by country. However, the user cannot customize the news feed except using preset filters. MIDAS improves the usability of that by providing a news stream (see the visualization module at the center of Figure 1) where the user can personalize the search query and even include blogs. It allows further exploration of COVID-19 related news specific to topics on the user's own health policy priorities, such as home care or childhood obesity. It includes a tag cloud to have a first grasp over the main topics under discussion. Alongside this useful tool, the MIDAS news dashboard [25] allows the user to further explore the news based on data visualization modules including related concepts, entities and categories, or even the sentiment of the news article selection. This analysis is particularly important to avoid bias in the health news search [26], and to explore the several dimensions of misinformation caused by the *infodemia* [37], in conjunction with the pandemic. The search engine uses Wikipedia terms, to ensure the multilingual potential of the dashboard. These include the following COVID-19 related terms: *"Coronavirus"* (on the Coronavirus virus family, available in 73 languages), *"Coronavirus disease 2019"* (corresponding to the specific COVID-19 sort, available in 136 languages), and *"2019–20 Coronavirus pandemic"* (that writes on the pandemic itself, available in 132 languages). Moreover, we can backtrack the news articles about Coronavirus in Italy, discussing the triage of passengers commuting from Wuhan, China, in the timeline exploration visualization module, to January 20th this year, at the beginning of the European epidemic. We can further access these articles' entities to identify main actors and related topics. We can also explore the sentiment over these news articles and, in some cases, their impact on social media through the number of times a particular news article was shared. All these features are offered over comprehensive and well documented APIs.

The usefulness of the MeSH classifier, described in the latter section, is extended in the MIDAS platform through its integration with the news dashboard. With this integration, the user can use the MeSH heading terms together with keywords in a query, when exploring a certain news topic, in a similar fashion to the usage of the well-accepted PubMed workflow, providing data visualization modules that include those classes. A meaningful example is the visualization module *"Article Categories"* where a MIDAS user can see the distribution of news

articles subsequent to the query throughout the related MeSH classes. Figure 6 shows that 6.64% of the news on Coronavirus talks about topics related to the Mesh heading *Organisms/Viruses/RNA Viruses*. These new capabilities enhance the monitoring of health news over structured information, allowing a MIDAS user to have an understanding of media coverage in closer conjunction to the biomedical research itself.

Campaigning Through Social Media

A Twitter chatbot campaign led by MIDAS helped assess the global efforts of people during the pandemic. The aim of the social media campaign was to check-in with the global Twitter community during the "One World: Together at Home" initiative led by Global Citizen, asking questions concerning their hopes, technology usage and feelings towards a more connected world. The Global Citizen initiative was an event that aimed to support the WHO's COVID-19 Solidarity Response Fund, which supports and equips healthcare workers around the world. Hundreds of thousands of pieces of protective medical equipment, and 1.5 million diagnostic kits were provided to countries around the world through this fund. As part of the efforts this global initiative brought together change makers from over 150 countries and helped to raise funds for the cause. However, the globally televised event itself was not a fundraising telethon. It focused on entertainment, messages of solidarity and showing support for healthcare workers [31], [33]. The social media campaign was connected to a Twitter account managed by the IBM Corporate Social Responsibility team (@IBMOrg). Using the Social Campaign Manager's ability to spawn chatbots at will, the team was able to create a user-friendly conversation-led questionnaire asking the public a series of questions. The campaign included a number of multiple choice questions ranging from simple yes, no, maybe answers, to questions requesting the selection of one or more items from a list of options, or the answer to open questions where the respondent was free to write their response in free-form

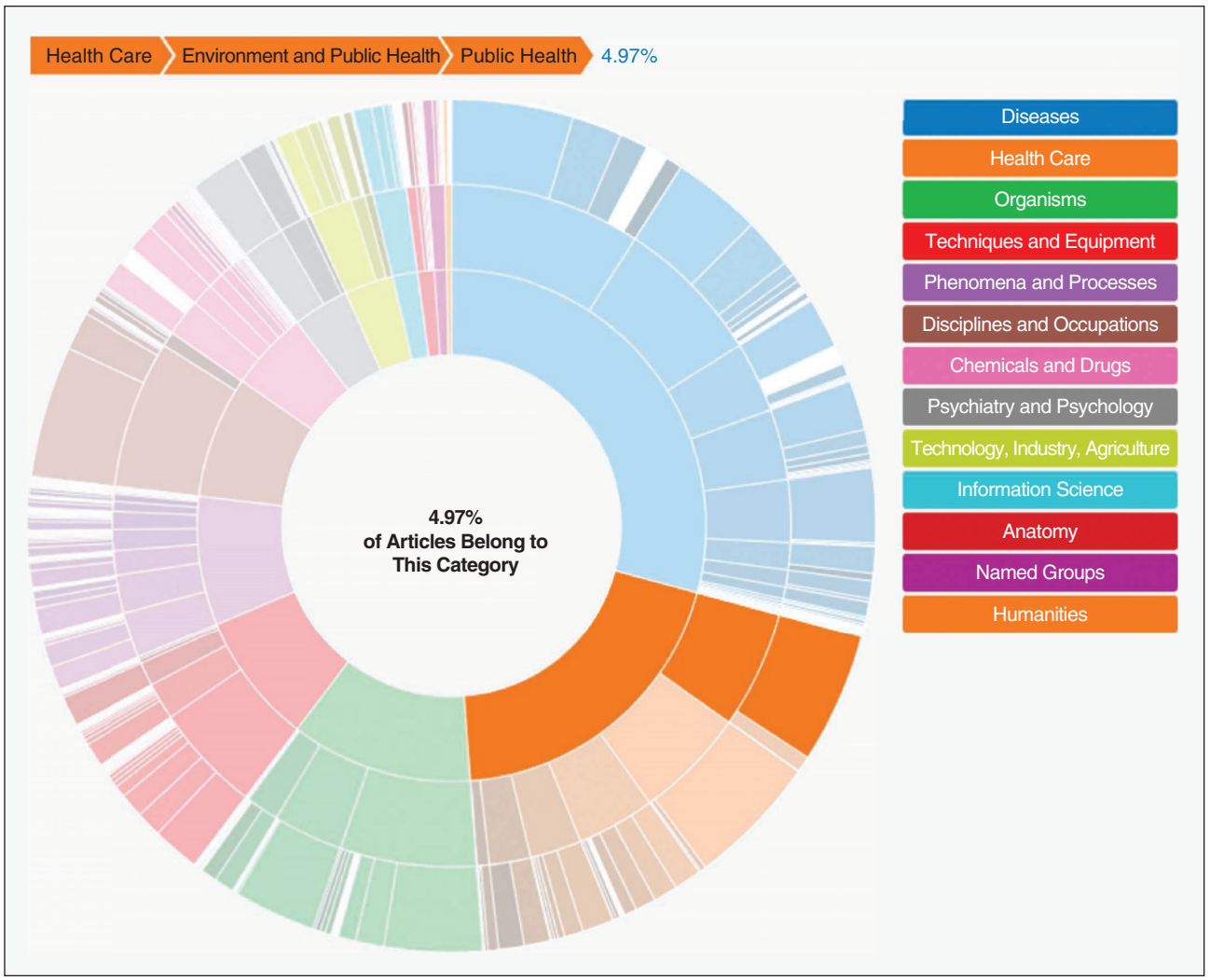


FIGURE 6 Analyzing the Coronavirus news articles through the percentage of health topics using the innovative integration between the MeSH classifier and the MIDAS news dashboard.

… a technological solution, such as the MIDAS platform, allows analysis of heterogeneous datasets, in an environment that allows relationships and policy to be explored.

text. The free form responses were analyzed using IBM's Watson Natural Language Understanding tool, which in turn provided the campaign owners with a general sense of the sentiment of the conversations they had with the chatbot, and emotions expressed in the responses. The aggregated view of these responses shows that the favorite stay at home activity of 38% of the public was the physical exercise, with video chatting at 25%, and hobbies at 20%. When asked how technology has shaped their life since the start of the COVID-19 pandemic a large majority (65%) of people said they rely on [technology] more than ever and a third of the respondents said they use it as much as before. Only 2% of the respondents said they use technology less than before.

Topics trending in the campaign responses were family and friends, health and fitness, resulting societal impacts of the disease, children and education. The scientific impact due to the global efforts worldwide was also mentioned. The system can identify the aggregated sentiment of free-form text responses given by all survey participants and provide the prediction of the mean probability of the emotions (within the categories: sadness, joy, fear, disgust or anger) in the free form responses using IBM's Watson NLU. Of all of the categories mentioned, only technology and computing were mentioned in a neutral context, whereas the majority of responses were given in a positive context which can be seen in green in the graph in Figure 6.

Ethics in the Time of COVID-19

"May you live in interesting times" is often quoted as an ancient Chinese curse, but dig deeper and this origin is erroneous. It is actually attributable to Joseph Chamberlain, a 19th century British politician. The parallels with the current COVID-19 crisis requires little imagination. The transitional world in which we currently live has unheard of restrictions of movements and freedoms normally available in democratic societies [15]. These restrictions are driven by modelling and the epidemiological evidence and, certainly to this point in May 2020, the public appears to have trusted the rationale and approach in large part [32]. Lockdown strategies are a matter of choosing short term loss over long term gain; these are the policy questions that are being dealt with and, as such, require the best evidence available. What is clear is that a technological solution, such as the MIDAS platform, allows analysis of heterogeneous datasets, in an environment that allows relationships and policy to be explored. A key output of the platform development was the realization that this environment should be apolitical, in the sense that policy should be based on science and the relationships of the data used in the system, robustly quality checked and analyzed, [5]. MIDAS therefore proposes a two-pronged approach to ethical and governance assurance: the public as partners and a system of robust ethical and scientific oversight from all parties involved in the MIDAS platform. Public engagement is core and needs to be meaningful. This requires a program of engagement, education and support for the public. Obviously this also requires nuance, resource and openness by science and government, as well as innovative techniques for engagement, such as the Chatbot discussed previously, and a platform such as 'engage' [9] used throughout the MIDAS platform development. In the time of SARS-CoV-2 (COVID-19) this may seem a luxury, but we need to plan now for future outbreaks, pandemics, or other public health emergencies. This public engagement and perhaps the use of opt in/opt out models of data use for public health is a discussion that needs to take place urgently. A measure of control is essential in managing public trust, expectation and compliance in the use of any system. MIDAS mitigates the risk by creating a system to manage this requirement: an Honest Broker Service model (HBS). This system creates an operational structure for review, scientific justification and oversight drawn from all interested parties, the public, government, academia and business. This is the ABCD model: Academia, Business, Client, and Direction. These parties set the bar in respect of the scientific/policy question at hand, allowing scrutiny of the hypothesis by parties not directly invested in specific work, within a framework that allows review for quality assurance and feasibility. This model creates a system that can be trusted, a regulatory framework much as the ones that exist for devices and pharmaceuticals that can allow science to drive decision making. Public messaging and education is integral to acceptance of the model, and of course the operational integrity of the system is dependent on user engagement through data use, a symbiotic relationship between the public and data investigators. A system that includes the public as contributor and gatekeeper, vouchsafed by independent review goes some way to safeguarding this trust.

Conclusion

A substantial part of technology adoption in public health and healthcare is the utility of the tools and the meaningfulness of their outcomes. As a result of being co-created with stakeholders [8], undergoing regular impact evaluations [10], and having usability formally evaluated by policy-makers [6], the MIDAS platform has proved its usefulness and has led to the development of components driven by stakeholder requests. Significant interest was established in the MIDAS platform in what it can offer to new regions, cities and organizations.

Moreover, the platform comprises a representative set of open, anonymized and synthetic data upon which the full range of available analytics and corresponding visualizations reside. This is valuable in an epidemic scenario, enriching the proprietary data of the public health authority, with existing

results in areas close to the disease (e.g. diabetes and old age), to the outcomes of related restriction measures (e.g. mental health and childhood obesity), and that can be refocused with low effort (e.g. child care to elderly care). The potential of (i) specific public health campaigns using social media; and (ii) worldwide news monitoring with a measure of impact in Facebook, can further help in understanding the spread of the disease and misinformation around it. In turn this will contribute to improvements in the public health campaigns that are an essential component for the success of disease control. Finally, the integration and utilization of open datasets, and the use of MEDLINE, in particular, greatly contribute to the understanding of the disease itself, when studying it side-by-side with the local data. A further ambition is to analyze and study how individuals' biological and psycho-emotional status with the actual data performs using adapted mental health and childhood obesity research questions for the COVID-19 pandemic. Results could influence the current pandemic response, alongside the development of health policy recommendations and preventive actions needed for prevention/control of future outbreaks or pandemics. The ethics and governance frameworks used in MIDAS, whilst operationally limited to the project and the HBS In Northern Ireland (with model development in the Basque region, and a similar model adopted within Finland), clearly articulate the demand and need for a system of oversight and review. Therefore, what is needed now is the adoption of the described model at scale, adequately resourced and built upon a funded and meaningful engagement piece. There will also need to be a shift to a position in which HBS becomes the Trusted system for review.

The MIDAS Open Source Foundation (OSF) will directly facilitate the long-term sustainability and growth of the MIDAS Platform and will provide an opportunity for health authorities, as well as regional, national and pan-European governments to embrace the platform to address strategic health policy development such as for COVID-19 or any future pandemic/global health crisis.

Acknowledgment

This project was funded by the European Union research fund 'Big Data Supporting Public Health Policies,' under GA No. 727721.

References

[1] T. Alamo et al., Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic. *Electronics*, vol. 9, no. 5, p. 827, 2020. doi: 10.3390/electronics9050827.

[2] "WHO COVID-19 Dashboard," ArcGIS. [Online]. Available: https://covid19.who.int/

[3] R. Armitage and L. B. Nellums, "COVID-19 and the consequences of isolating the elderly," *Lancet Public Health*, vol. 5, no. 5, p. e256, 2020. doi: 10.1016/S2468-2667(20)30061-X.

[4] M. Black et al., "Meaningful Integration of data, analytics and services of computer-based medical systems: The MIDAS touch," in *Proc. IEEE 32nd Int. Symp. Computer-Based Medical Systems (CBMS)*, 2019, pp. 104–105.

[5] P. Carlin, "D2.2 Good practice guide - Ethics and Governance Workpage at MIDAS H2020," Unpublished.

[6] B. Cleland et al., "Meaningful integration of data analytics and services in MIDAS project: Engaging users in the co-design of a health analytics platform," in *Proc. 32nd British Computer Society Human Computer Interaction Conf. (BCS HCI)*, 2018, pp. 1–4.

[7] B. Cleland et al., "Insights into antidepressant prescribing using open health data," *Big Data Res.*, vol. 12, pp. 41–48, 2018. doi: 10.1016/j.bdr.2018.02.002.

[8] B. Cleland et al., "Usability evaluation of a co-created big data analytics platform for health policy-making," in *Proc. Int. Conf. Human-Computer Interaction '19*, 2019, pp. 194–207.

[9] B. Cleland et al., "The 'engage' system: Using real-time digital technologies to support citizen-centred design in government," in *User Centric E-Government. Integrated Series in Information Systems*, S. Saeed, T. Ramayah, Z. Mahmood, Eds. Cham: Springer-Verlag.

[10] J. Connolly et al., "Impact evaluation of an emerging European health project: The MIDAS model," *Bus. Syst. Res., Int. J. Soc. Adv. Innov. Res. Econ.*, vol. 11, no. 1, pp. 142–150, 2020. doi: 10.2478/bsrj-2020-0010.

[11] "Elasticsearch portal," Elastic. [Online]. Available: https://www.elastic.co/

[12] C. J. Galvin et al., "Accelerating the global response against the exponentially growing COVID-19 outbreak through decent data sharing." *Diagnostic Microbiol. Infect. Dis.* p. 115070, 2020. doi: 10.1016/j.diagmicrobio.2020.115070.

[13] W. Guo et al., "Diabetes is a risk factor for the progression and prognosis of COVID-19" *Diabetes Metab. Res. Rev.*, p. e3319, 2020. doi: 10.1002/dmrr.3319.

[14] "COVID-19 Open Research Dataset Challenge (CORD-19)," Kaggle. [Online]. Available: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[15] M. Karwowski et al., "When in danger, turn right: Covid-19 threat promotes social conservatism and right-wing presidential candidates," 2020, PsyArXiv.

[16] J. B. Kornum et al., "Type 2 diabetes and pneumonia outcomes: a population-based cohort study," *Diabetes Care*, vol. 30, no. 9, pp. 2251–2257, 2007. doi: 10.2337/dc06-2417.

[17] G. Leban et al., "Event registry: Learning about world events from news," in *Proc. Int. Conf. World Wide Web*, 2014, pp. 107–111. doi: 10.1145/2567948.2577024.

[18] S. Madsbad, "COVID-19 infection in people with diabetes," Touch Endocrinology. [Online]. Available: www.touchendocrinology.com/insight/covid-19-infection-in-people-with-diabetes/

[19] C. Manning et al., *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008, pp. 269–273.

[20] "COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv," medRxiv. [Online]. Available: https://connect.medrxiv.org/relate/content/181

[21] "MIDAS project website," MIDAS. [Online]. Available: http://www.midasproject.eu/.

[22] "MEDLINE - Description of the database," National Library of Medicine (NLM). [Online]. Available: https://www.nlm.nih.gov/bsd/medline.html

[23] "PubMed search engine," NLM. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/

[24] A. Pietrobelli et al., "Effects of COVID-19 lockdown on lifestyle behaviors in children with obesity living in Verona, Italy: A longitudinal study," *Obesity*, vol. 8, no. 8, pp. 1382–1385, 2020. doi: 10.1002/oby.22861.

[25] J. Pita Costa et al., "Text mining open datasets to support public health," in *Proc. WITS Conf.*, 2017.

[26] J. Pita Costa et al., "Health News Bias and its impact in Public Health," in *Proc. Slovenian KDD Conf. (SIKDD'19)*, 2019.

[27] J. Pita Costa et al., "A new classifier designed to annotate health-related news with MeSH headings," *Artif. Intell. Med.*, submitted for publication.

[28] D. Rankin et al. "Reliability of supervised machine learning using synthetic data in healthcare: Model to preserve privacy for data sharing," *JMIR Med Inform.*, vol. 8, no. 7, p. e18910, 2020. doi: 10.2196/18910.

[29] "Coronavirus news monitor," RavenPack. [Online]. Available: https://coronavirus.ravenpack.com/

[30] R. Alvarez et al., "TAQIH, a tool for tabular data quality assessment and improvement in the context of health data," *Comput. Methods Programs Biomed.*, vol. 181, p. 104824.

[31] L. Snapes. "Lady Gaga, Billie Eilish and Paul Mccartney to play coronavirus benefit," *The Guardian*. [Online]. Available: https://www.theguardian.com/music/2020/apr/06/lady-gaga-billie-eilish-and-paul-mccartney-to-play-coronavirus-benefit

[32] J. Stone, "Coronavirus testing: Government accused of 'misleading the public' amid criticism over figures," *The Independent*, May 2, 2020. Accessed: May 26, 2020. [Online]. Available: https://www.independent.co.uk/news/uk/politics/coronavirus-testing-figures-uk-target-criticism-hancock-a9495621.html

[33] "One world: Together at home could be live aid for the coronavirus generation," The Verge, 2020. [Online]. Available: https://www.theverge.com/2020/4/7/21211716/one-world-together-at-home-benefit-concert-lady-gaga-covid-19-global-citizen

[34] "Coronavirus watch portal," UNESCO AI Research Inst. [Online]. Available: http://coronaviruswatch.ircai.org/

[35] "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020," World Health Organization. [Online]. Available: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020

[36] "Emergency use ICD codes for COVID-19 disease outbreak," World Health Organization. [Online]. Available: https://www.who.int/classifications/icd/covid19/en/

[37] J. Zarocostas. "How to fight an infodemic," *Lancet*, vol. 395, no. 10225, p. 676, 2020. doi: 10.1016/S0140-6736(20)30461-X.

[38] "Coronavirus Disease Research Community - COVID-19," Zenodo. [Online]. Available: https://zenodo.org/communities/covid-19/

Intelligent Optimization of Diversified Community Prevention of COVID-19 Using Traditional Chinese Medicine

©SHUTTERSTOCK/CHANUT IAMNOY

Abstract—Traditional Chinese medicine (TCM) has played an important role in the prevention and control of the novel coronavirus pneumonia (COVID-19), and community prevention has become the most essential part in reducing the risk of spread and protecting public health. However, most communities use a unified TCM prevention program for all residents, which violates the "treatment based on syndrome differentiation"

Yu-Jun Zheng and Si-Lan Yu
Hangzhou Normal University, CHINA

Jun-Chao Yang and Tie-Er Gan
Zhejiang Chinese Medical University, CHINA

Qin Song and Jun Yang
Hangzhou Normal University, CHINA

Mümtaz Karataş
National Defense University, TURKEY

Digital Object Identifier 10.1109/MCI.2020.3019899
Date of current version: 14 October 2020

Corresponding Author: Y. J. Zheng (email: yujun.zheng@computer.org).

principle of TCM and limits the effectiveness of prevention. In this paper, we propose an intelligent optimization method to develop diversified TCM prevention programs for community residents. First, we use a fuzzy clustering method to divide the population based on both modern medicine and TCM health characteristics; we then use an interactive optimization method, in which TCM experts develop different TCM prevention programs for different clusters, and a heuristic algorithm is used to optimize the programs under the resource constraints. We demonstrate the computational efficiency of the proposed method, and report the application results of the method in TCM-based prevention of COVID-19 in 12 communities in Zhejiang province, China, during the peak of the pandemic.

Using a "one-size-fits-all" prevention program for all residents is rarely effective, while developing a personalized program for each resident would be too expensive. We prefer to develop a set of diversified prevention programs, each for a group of residents.

I. Introduction

The ongoing outbreak of the novel coronavirus pneumonia (COVID-19), declared by the World Health Organization as a global public health emergency, has been reported in over twenty-two million cases in over 200 countries and territories as of August 21, 2020. Community prevention and control has become the most basic and essential part in reducing the risk of spread and protecting public health during the pandemic [1], [2]. Currently, community prevention is a significant challenge, not only because there is still no effective antiviral or vaccine, but also because of the pressing need to restart economy and restore social life [3], [4].

Although modern medicine offers accurate diagnosis and treatment methods for many diseases, it shows weakness in preventing emerging infectious diseases such as COVID-19 for which there is no vaccine, because epidemic prevention solutions based on modern medicine heavily rely on a clear understanding of the pathogenic mechanism and a number of large case-controlled studies [5], [6], and the misuse of antibiotics can cause severe side effects [7].

Traditional Chinese medicine (TCM) has been developed and used in prevention and treatment of various diseases for thousands of years in Chinese history. TCM is a comprehensive system of treatment of acute and chronic disorders as well as for the prevention of such disorders mainly based on herb medicine [8]. Unlike modern medicine that focuses on killing viruses, TCM pays attention to improving the inherent self-resistance and reducing the likelihood of disease onset by using a unique holistic approach to establish equilibrium in the whole and individual parts of the body [9]. The present principles on prevention of COVID-19 are to tonify body energy to protect the outside body, dispel wind, dissipate heat and dissipate dampness [10]. Facing an emerging infectious disease, TCM prescriptions (formulae) are made by combining existing crude herbs or minerals instead of developing new drugs. That is why TCM has achieved great success in response to recent epidemics such as SARS, H1N1, and Zika [11]–[15], and is playing a vital role in reducing the incidence rate and controlling the spread of COVID-19 [16]–[18].

However, we believe that there is much room for improving community prevention of the pandemic using TCM. For example, since the COVID-19 outbreak in China, many local public health administrations have issued TCM prevention programs for COVID-19, and some communities used a single program or prescription for all residents [19], [20]. Such a "one-size-fits-all" solution violates the "treatment according to three factors (time, place, and people)" and "treatment based on syndrome differentiation" principles of TCM and, therefore, limits the effectiveness of prevention.

According to requirements of local governments to improve community prevention of COVID-19, we propose an intelligent, diversified community prevention method for COVID-19 by combining TCM and modern computational intelligence methods. First, we use a fuzzy clustering method to classify the population based on both modern medicine and TCM health characteristics. According to the health characteristics, TCM experts develop a TCM prevention program for each cluster. The initial program for each cluster aims to maximize the prevention effect on residents in the cluster. Nevertheless, the demands of all initial programs often exceed the available resources. We then use an interactive optimization method to continually evolve the prevention programs until all programs are approved by the TCM experts while all resource demands are satisfied. The proposed method has been practiced in a number of communities in Zhejiang province and extended to many other regions in China. The main contribution of this work is the combination of traditional medicine with computational intelligence methods for diversified prevention of COVID-19; in particular, we propose an improved fuzzy clustering method for residence grouping, and employ bio-inspired optimization to fully utilize available resources to develop diversified programs to improve prevention effects.

II. Fuzzy Clustering of Community Residents

Using a "one-size-fits-all" prevention program for all community residents fails to consider interpersonal differences and is rarely effective. On the contrary, developing a personalized prevention program for each resident would be too expensive due to limited medical resources under the pandemic. We prefer to develop a set of diversified prevention programs, each for a group of residents with similar physical characteristics. The characteristics include both modern medicine characteristics and TCM health characteristics, as summarized in Table I. The

data sources include both modern healthcare records and TCM records of the residents. Those characteristics are used as input features for grouping, and the value of each feature is normalized, e.g., real values (such as height and weight) are normalized to [0,1], and exclusive labels (such as sex and illness) are represented by binary variables. In this study, we use a total of 1148 features for grouping. However, for most residents, many features can be absent.

A. An Improved Fuzzy c-Means Clustering Method

For such a high-dimension grouping problem with many missing values, classical hard clustering methods such as *K*-means clustering [22] are not very effective. In this study, we use an improved fuzzy *c*-means (FCM) clustering method [23]. In brief, FCM groups a set of data points by minimizing the overall fuzzy-membership-weighted distance of the data points from cluster centroids:

$$\min J(U,V)=\frac{1}{cn}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2} \tag{1}$$

where n is the number of data points, c is the number of clusters, u_{ij} is the membership degree of the jth data point to the ith cluster subject to $\Sigma_i u_{ij}=1$, d_{ij} is the distance between the jth data point and the ith cluster centroid, m is a control parameter with a default value of 2, $U=(u_{ij})_{c\times n}$ is the weight matrix, and $V=[\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_c]$ is the set of cluster centroid vectors.

Xu and Wu [24] extended the standard FCM to intuitionistic FCM based on new distance measures defined on intuitionistic fuzzy sets (IFS) [25] that capture more uncertainty information. Here, we use Pythagorean fuzzy sets (PFS) [26], which allow for a larger body of membership grades than IFS, to further improve clustering. Formally, let S be an arbitrary non-empty set, a PFS is a mathematical object of the form:

$$P=\{\langle x,\mu_P(x),\nu_P(x)\rangle \mid x\in S\} \tag{2}$$

where $\mu_P(x):S\to[0,1]$ and $\nu_P(x):S\to[0,1]$ are respectively the membership degree and non-membership degree of the element x to S in P. PFS extends IFS in that the membership degrees satisfy $\mu_P^2(x)+\nu_P^2(x)\le 1$. The hesitant degree of $x\in X$ is expressed as:

$$\pi_P(x)=\sqrt{1-\mu_P^2(x)-\nu_P^2(x)} \tag{3}$$

In our improved Pythagorean FCM, data points and cluster centroids are represented by Pythagorean fuzzy number (PFN) vectors, where the distance between two PFN $\beta_1=P(\mu_{\beta_1},\nu_{\beta_1})$ and $\beta_2=P(\mu_{\beta_2},\nu_{\beta_2})$ is calculated as [27]:

$$|\beta_1,\beta_2|=\sqrt{\frac{(\mu_{\beta_1}^2-\mu_{\beta_2}^2)^2+(\nu_{\beta_1}^2-\nu_{\beta_2}^2)^2+(\pi_{\beta_1}^2-\pi_{\beta_2}^2)^2}{2}} \tag{4}$$

and the distance d_{ij} in Eq. (1) between a D-dimensional data point $\mathbf{x}_j$ and a cluster centroid $\mathbf{v}_i$ is calculated as:

$$\|\mathbf{x}_j,\mathbf{v}_i\|=\sqrt{\frac{\sum_{d=1}^{D}|x_{jd},v_{id}|^2}{D}} \tag{5}$$

Algorithm 1 presents the Pythagorean FCM method, which minimizes the objective function $J(U, V)$ by iteratively updating the fuzzy membership weights according to the distances between data points and centroids (Line 11) and updating the centroids according to the membership and

TABLE 1 Physical characteristics for grouping community residents.

TYPE	CHARACTERISTICS	NUMBER OF INDICATORS
BASIC HEALTH METRICS	AGE, SEX, OCCUPATION, HEIGHT, WEIGHT, HEART RATE, BLOOD PRESSURE, VITAL CAPACITY, …	121
TCM CONSTITUTIONS	MILD, YANG DEFICIENCY, YIN DEFICIENCY, PHLEGM DAMPNESS, WET & HEAT, QI STAGNATION, QI DEFICIENCY, BLOOD STASIS, SPECIAL	9
TCM SYNDROMES [21]	SHIRE JINYIN, PIXU SHIYUN, XUEXU FENGZAO, SHIRE YUZU, SHIRE SHANGYIN, QIZHI XUEYU, QIXU BUZU, GANSHEN BUZU, …	50
PAST ILLNESSES	DISEASES AND THE CORRESPONDING INDICATORS	484
CURRENT ILLNESSES	DISEASES AND THE CORRESPONDING INDICATORS	484

Algorithm 1 Pythagorean fuzzy *c*-means clustering algorithm.

1 Initialize a $c\times n$ matrix U and a set $V^{(0)}$ of c cluster centroids;
2 Let $k=0$;
3 **while** $\|V^{(k+1)},V^{(k)}\|>\epsilon$ **do**
4 **for** $j=1$ *to* n **do**
5 **if** $\exists i':1\le i'\le c:\|\mathbf{x}_j,\mathbf{v}_{i'}\|=0$ **then**
6 **for** $i=1$ *to* c **do**
7 **if** $i=i'$ **then** $u_{ij}^{(k)}\leftarrow 1$;
8 **else** $u_{ij}^{(k)}\leftarrow 0$;
9 **else**
10 **for** $i=1$ *to* c **do**
11 $u_{ij}^{(k)}\leftarrow\dfrac{1}{\sum_{i'=1}^{c}\left(\dfrac{\|\mathbf{x}_j,\mathbf{v}_i\|}{\|\mathbf{x}_j,\mathbf{v}_{i'}\|}\right)^{\frac{2}{m-1}}}$;
12 **for** $i=1$ *to* c do
13 $\mathbf{v}_i^{(k)}\leftarrow\left\{\left\langle\beta_d,\dfrac{\sum_{j=1}^{n}u_{ij}^{(k)}\mu_{\beta_d}^2}{\sum_{j=1}^{n}u_{ij}^{(k)}},\dfrac{\sum_{j=1}^{n}u_{ij}^{(k)}\nu_{\beta_d}^2}{\sum_{j=1}^{n}u_{ij}^{(k)}}\right\rangle\mid 1\le d\le D\right\}$;
14 $k\leftarrow k+1$;
15 **return** (U, V);

non-membership degrees of data points to clusters (Line 13) to apply the derivative of $J(U, V)$ [24].

B. A Metaheuristic for Optimizing Clusters

The performance of FCM clustering heavily depends on the quality of initial cluster centroids [28], [29]. Instead of randomly setting initial cluster centroids, we use a metaheuristic algorithm, ecogeography-based optimization (EBO) [30], to optimize initial cluster centroids [29]. The algorithm starts by initializing a population of solutions, each representing a set $V^{(0)}$ of c initial cluster centroids. Let $J(\mathbf{x})$ denote the resulting $J(U, V)$ value obtained by the FCM method from the initial cluster centroids of $\mathbf{x}$; each solution $\mathbf{x}$ is assigned with an emigration rate $E_r(\mathbf{x})$ that is inversely proportional to $J(\mathbf{x})$ and an immigration rate $I_r(\mathbf{x})$ that is proportional to $J(\mathbf{x})$:

$$E_r(\mathbf{x}) = \frac{J_{\max} - J(\mathbf{x}) + \epsilon}{J_{\max} - J_{\min} + \epsilon} \tag{6}$$

$$I_r(\mathbf{x}) = \frac{J(\mathbf{x}) - J_{\min} + \epsilon}{J_{\max} - J_{\min} + \epsilon} \tag{7}$$

where $J_{\max}$ and $J_{\min}$ are the maximum and minimum $J(\cdot)$ values among the population, respectively, and ϵ is a small positive number to avoid division-by-zero. In this way, a better solution has a higher probability of emigrating features to other solutions, while a worse solution has a higher probability of immigrating features from other solutions [31].

The EBO algorithm then continually evolves the solutions using two migration operators: local migration and global migration. Local migration updates a solution $\mathbf{x}$ at each dimension d by migrating the corresponding dimension of a neighboring solution $\mathbf{x}^{\dagger}$ as follows:

$$x'_d = x_d + rand(0,1) \cdot (x^{\dagger}_d - x_d) \tag{8}$$

where $rand(0,1)$ produces a random number uniformly distributed in (0, 1), and $\mathbf{x}^{\dagger}$ is selected from all neighbors of $\mathbf{x}$ with a probability proportional to $E_r(\mathbf{x}^{\dagger})$.

Global migration updates a solution $\mathbf{x}$ at each dimension d by migrating the corresponding dimensions of both a neighboring solution $\mathbf{x}^{\dagger}$ and a non-neighboring solution $\mathbf{x}^{\ddagger}$ as follows:

$$x'_d = \begin{cases} x^{\dagger}_d + rand(0,1) \cdot (x^{\ddagger}_d - x_d), & f(\mathbf{x}^{\ddagger}) \le f(\mathbf{x}^{\dagger}) \\ x^{\ddagger}_d + rand(0,1) \cdot (x^{\dagger}_d - x_d), & f(\mathbf{x}^{\ddagger}) > f(\mathbf{x}^{\dagger}) \end{cases} \tag{9}$$

where $\mathbf{x}^{\dagger}$ is selected from all neighbors of $\mathbf{x}$ and $\mathbf{x}^{\ddagger}$ is selected from all non-neighboring solutions of $\mathbf{x}$; the selection probabilities are proportional to E_r.

EBO uses a parameter η as the probability of performing global migration and, therefore, $(1-\eta)$ as the probability of performing local migration. The value of η dynamically increases from a lower limit $\eta_{\min}$ to an upper limit $\eta_{\max}$ with generation g of the algorithm:

$$\eta = \eta_{\min} + \frac{g}{g_{\max}}(\eta_{\max} - \eta_{\min}) \tag{10}$$

In this study, we use a local random neighborhood structure [32], which randomly assigns k_N neighboring solutions to each solution in the population (where k_N is a parameter set to 3); if the current best solution has not been updated after a number $\hat{g}$ of consecutive generations, the neighborhood structure is randomly reset. Algorithm 2 presents the pseudocode of the EBO algorithm, where Line 4 invokes Algorithm 1 to evaluate the fitness of each solution (initial centroid setting).

III. Interactive Optimization of Prevention Programs

After clustering the community residents into c groups, we invite TCM experts to assess the health characteristics of each cluster (and examine typical residents if possible), and develop diversified prevention programs according to the characteristics. Note that the number p of prevention programs approximates, but does not necessarily equal, the number c of clusters. That is, the TCM experts typically develop a prevention program (including a TCM prescription and other supplementary means such as acupuncture and moxibustion) for a cluster; they

Algorithm 2 **The EBO algorithm for enhancing the fuzzy clustering method.**

```
1  Randomly initialize a population of solutions (initial set of
   cluster centroids);
2  while the stopping criterion is not satisfied do
3      foreach solution x in the population do
4          Use Algorithm 1 to produce the clustering results
           (U, V) from the initial cluster centroids of x;
5      Let x* be the best solution in the population;
6      foreach solution x in the population do
7          Compute E_r(x) and I_r(x) according to Eqs. (6) and
           (7);
8      foreach solution x in the population do
9          for d = 1 to n do
10             if rand(0,1) < I_r(x) then
11                 Select a neighboring x† with probability
                   proportional to E_r(x†);
12                 if rand(0,1) < η then
13                     Select a non-neighboring x‡ with prob-
                       ability proportional to E_r(x‡);
14                     Do global migration according to Eq. (9);
15                 else
16                     Do local migration according to Eq. (8);
17         if the migrated solution x' is better than x then
18             x ← x';
19     Update η according to Eq. (10);
20     if x* has not been updated for ĝ consecutive generations
       then
21         Randomly reset the neighborhood structure;
22 return the clustering result of the best known solution x*.
```

A TCM prescription can have dozens of ingredients, and a drug can have dozens of alternative drugs. Therefore, the solution space of the problem can be very large.

may also develop a prevention program for two similar clusters as they consider appropriate. For a cluster with high-risk residents (with suspected symptoms of COVID-19 or serious underlying illnesses), they can decide to develop one prevention program for each resident.

When developing initial programs, the TCM experts aim to maximize the prevention effect on residents of each cluster without considering the limits of medical resources including herbal medicines, patent medicines, medical devices, pharmacists, and other paramedical personnel. If the demands of the programs exceed the available resources, we use an intelligent optimization algorithm to optimize the programs subject to the resource constraints. However, any prevention program produced by computer algorithms must be checked and, if necessary, modified by the TCM experts before implementation. The above process continues until all prevention programs satisfy the resource constraints and are approved by the TCM experts. Fig. 1 illustrates the interactive optimization process.

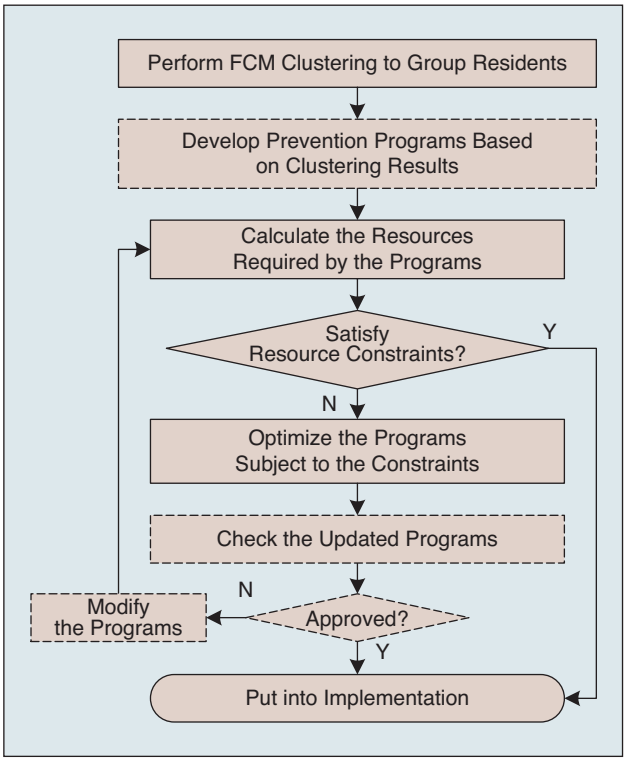


FIGURE 1 Flowchart of the interactive optimization of TCM prevention programs. Rectangles with dash borders are performed by TCM experts, while rectangles with solid borders are performed by computer.

A. Optimization Problem

The prevention program optimization problem is formulated as follows. TCM experts have developed N basic prevention programs, denoted by $\{P_1, P_2, \ldots, P_N\}$, for residents from M communities. The programs involve K types of drugs, K_1 types of other medical resources (such as TCM material and movable devices) that can be shared among the communities, and K_2 types of other medical resources (such as immovable devices and staff belonging to given communities) that cannot be shared among the communities. When there is no confusion, we use k as the index for the above three classes of resources.

The problem is to optimize the distribution of medical resources among the prevention programs for the M communities. In the local region, the total available quantity of each drug k is $\hat{q}_k^{\mathrm{D}}$ $(1 \le k \le K)$, total available quantity of each other sharable resource k is $\hat{q}_k^{\mathrm{G}}$ $(1 \le k \le K_1)$, and available quantity of each non-sharable resource k in community i is $\hat{q}_{ik}^{\mathrm{F}}$ $(1 \le k \le K_2; 1 \le i \le M)$.

Each prevention program P_j has the following attributes $(1 \le j \le N)$:

- The number n_j of residents using the program;
- The set Θ_j of communities that have residents using the program; for each community $i \in \Theta_j$, the number of residents using the program is n_{ij};
- The set Φ_j of drugs used by the program; for each drug $k \in \Phi_j$, the quantity used per prescription is q_{jk}^{D};
- The set Ψ_j of other sharable resources used by the program; for each resource $k \in \Psi_j$, the quantity used per prescription is q_{jk}^{G};
- The set Ω_j of other non-sharable resources used by the program; for each resource $k \in \Omega_j$, the quantity used per prescription is q_{jk}^{F};

In a TCM prescription, many ingredients have alternatives. We use $\Phi'_j \subset \Phi_j$ to denote the subset of drugs that have alternatives in P_j; for each drug $k \in \Phi'_j$, we use a list Λ_k to store its alternative drugs in decreasing order of priority, which are determined by TCM drug properties and effects to the disease (COVID-19 belongs to pulmonary disease in TCM). We consider two types of updates on a TCM prescription:

- Replacing an auxiliary drug $k \in \Phi'_j$ with an alternative drug $k' \in \Lambda_k$, for example, replacing coix seed with winter melon seed;
- Replacing a main drug $k \in \Phi'_j$ with an alternative drug $k' \in \Lambda_k$; however, according to compatibility of TCM, a main drug is related to one or more auxiliary drugs, and the corresponding auxiliary drugs should also be reapplied; an example is replacing "astragalus membranaceus (main) + cinnamon (auxiliary)" with "codonopsis pilosula (main) + yam (auxiliary)."

To prevent the updated prescriptions from deviating too much from the original prescriptions developed by TCM experts, for each prevention program, we allow at most one drug (except auxiliary drugs related to a main updated drug) to

be updated each time. For either of the two types of updates above, we use $P_j(x_j, x'_j)$ to denote the updated program, where x_j denotes the original drug in the prescription, and x'_j denotes the alternative drug in Λ_{x_j}. Therefore, for the set of original prevention programs $\{P_1, P_2, \ldots, P_N\}$, each solution to the problem can be represented by a $(2N)$-dimensional integer vector $\mathbf{x} = \{x_1, x'_1, x_2, x'_2, \ldots, x_N, x'_N\}$, which indicates that the x_j-th drug in P_j is to be replaced by its x'_j-th alternative $(1 \le j \le N)$; without loss of generality, $x_j = x'_j$ denotes that P_j is unchanged.

Based on the efficacy of the original and alternative drugs, we can determine the quantity of an alternative drug k' used to replace an original drug k in the prescription. Based on the change of the prescription, we can then determine the changes of other medical resources, such as the types and quantities of material and the working hours for processing the drugs. Consequently, we obtain the following attributes of the updated prevention program $P_j(x_j, x'_j)$:

- The set $\mathbf{\Phi}_j(x_j, x'_j)$ of drugs; for each drug $k \in \mathbf{\Phi}_j(x_j, x'_j)$, the quantity used per prescription is $q^{\mathrm{D}}_{jk}(x_j, x'_j)$;
- The set $\mathbf{\Psi}_j(x_j, x'_j)$ of other sharable resources used by the program; for each resource $k \in \mathbf{\Psi}_j(x_j, x'_j)$, the quantity used per prescription is $q^{\mathrm{G}}_{jk}(x_j, x'_j)$;
- The set $\mathbf{\Omega}_j(x_j, x'_j)$ of other non-sharable resources used by the program; for each resource $k \in \mathbf{\Omega}_j(x_j, x'_j)$, the quantity used per prescription is $q^{\mathrm{F}}_{jk}(x_j, x'_j)$.

The objective of the problem is to maximize the overall effects of the updated prevention programs, provided that the resources used by the programs do not exceed the available resources. The effect of each updated program $P_j(x_j, x'_j)$ is evaluated based on its deviation from the original program P_j; the larger the deviation, the smaller the effect is, as we should trust the ability of TCM experts who develop the original program. The deviation of $P_j(x_j, x'_j)$ from P_j is assessed in two aspects:

- The importance of drug x_j in the original P_j, which is measured by a weight w_{jx_j}; a larger priority indicates a larger deviation;
- The priority of drug x'_j in the alternative set Λ_{x_j}; a higher priority indicates a smaller deviation.

Here, we calculate the deviation as follows:

$$\Delta P_j(x_j, x'_j) = w_{jx_j} I(\Lambda_{x_j}, x'_j) \tag{11}$$

where $I(\Lambda_{x_j}, x'_j)$ is the index of x'_j in Λ_{x_j} (without loss of generality, we set $\Delta P_j(x_j, x_j) = 0$).

Moreover, we use a weight w_j to denote the susceptibility of residents covered by program P_j to the infectious disease, and use a weight w'_i to denote the importance of each community i (which is related to the openness and population density of the community). The objective of the problem is defined as:

$$\min f(\mathbf{x}) = \sum_{j=1}^{N} \sum_{i \in \Theta_j} w_j w'_i \Delta P_j(x_j, x'_j) \tag{12}$$

The constraints of the problem are the quantities of each drug, other sharable resource, and other non-sharable resource used by the programs cannot exceed available quantities:

$$\sum_{j=1}^{N} n_j q^{\mathrm{D}}_{jk}(x_j, x'_j) \le \hat{q}^{\mathrm{D}}_k, \quad 1 \le k \le K \tag{13}$$

$$\sum_{j=1}^{N} n_j q^{\mathrm{G}}_{jk}(x_j, x'_j) \le \hat{q}^{\mathrm{G}}_k, \quad 1 \le k \le K_1 \tag{14}$$

$$\sum_{j=1}^{N} n_{ij} q^{\mathrm{F}}_{jk}(x_j, x'_j) \le \hat{q}^{\mathrm{F}}_{jk}, \quad 1 \le i \le M; 1 \le k \le K_2 \tag{15}$$

It should be noted that, in Eqs. (13)–(15), we uniformly use the operator Σ for notational simplicity; however, it may not always necessarily be summation. Typically, for drugs and material, Σ denotes summation; for other resources such as devices and personnel, Σ can be other corresponding aggregation operators. For example, supposing that a decocting machine can process 50 doses of a prescription, the operator will add 1 per 50 doses, and will also add 1 if the number of remaining doses is less than 50.

B. Optimization Algorithm

A TCM prescription can have dozens of ingredients, and a drug can have dozens of alternative drugs. Therefore, when the number N of prevention programs is relatively large, the solution space of the problem can be very large, for which exact optimization algorithms are often inefficient.

We use a metaheuristic optimization algorithm, water wave optimization (WWO) [33], to efficiently solve the problem. The algorithm starts by initializing a population of N_P solutions. To evaluate the fitness of each solution $\mathbf{x}$, we employ three penalty functions $v_D(\mathbf{x})$, $v_G(\mathbf{x})$, and $v_F(\mathbf{x})$ to calculate the violations of constraints (13), (14), and (15) as follows:

$$v_D(\mathbf{x}) = \sum_{k=1}^{K} \max\left(0, \sum_{j=1}^{N} n_j n^{\mathrm{D}}_{jk}(x_j, x'_j) - \hat{n}^{\mathrm{D}}_k\right) \tag{16}$$

$$v_G(\mathbf{x}) = \sum_{k=1}^{K_1} \max\left(0, \sum_{j=1}^{N} n_j n^{\mathrm{G}}_{jk}(x_j, x'_j) - \hat{n}^{\mathrm{G}}_k\right) \tag{17}$$

$$v_F(\mathbf{x}) = \sum_{i=1}^{M} \sum_{k=1}^{K_2} \max\left(0, \sum_{j=1}^{N} n_{ij} n^{\mathrm{F}}_{jk}(x_j, x'_j) - \hat{n}^{\mathrm{F}}_{ik}\right) \tag{18}$$

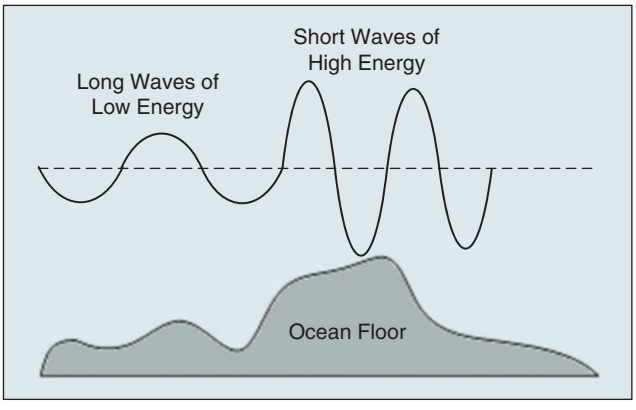


FIGURE 2 Wave lengths of high-fitness and low-fitness waves (solutions).

And the fitness $fit(\mathbf{x})$ is calculated as:

$$fit(\mathbf{x}) = 1/(f(\mathbf{x}) + v_D(\mathbf{x}) + v_G(\mathbf{x}) + v_F(\mathbf{x})) \tag{19}$$

We sort all N_P solutions in the population in decreasing order of the fitness value. Let $o(\mathbf{x})$ be the index of solution $\mathbf{x}$ in the sorted population; according to the principles of adapting WWO for combinatorial optimization [34], we calculate a wavelength $\lambda(\mathbf{x})$ for each $\mathbf{x}$ as an integer between 1 and N as follows:

$$\lambda(\mathbf{x}) = N - \left\lceil \frac{N_P - o(\mathbf{x})}{N_P - 1}(N-1) \right\rceil \tag{20}$$

where $\lceil \cdot \rceil$ denotes rounding up to the nearest integer.

The WWO iteratively evolves the solutions using three operators including propagation, refraction, and breaking. The propagation operator is based on two neighborhood structures. Given a solution $\mathbf{x}$ to the problem, its neighboring solutions can be obtained using one of the following two approaches:

- Randomly selecting a prescription P_j, changing x'_j to a random drug in Λ_{x_j}; this indicates modifying the alternative drug used in $P_j(x_j, x'_j)$;
- Randomly selecting a prescription P_j, changing x_j to another drug k in Φ_j, and then changing x'_j to a random drug in Λ_k; this indicates modifying both the drug to be replaced in P_j and the alternative drug used in $P_j(x_j, x'_j)$.

The propagation updates each solution $\mathbf{x}$ by performing $\lambda(\mathbf{x})$ steps of neighborhood search, i.e., propagates $\mathbf{x}$ to a $\lambda(\mathbf{x})$-step neighboring solution. In this way, a solution with higher fitness (and therefore smaller wavelength) exploits a smaller area, while a solution with lower fitness explores a larger area, as illustrated in Fig. 2 [33]. If the resulting $\lambda(\mathbf{x})$-step neighbor is better than $\mathbf{x}$, it replaces $\mathbf{x}$ in the population.

The refraction updates a stagnant solution $\mathbf{x}$ that has not been improved for $\hat{g}$ consecutive generations (where $\hat{g}$ is a control parameter) by making it learn from the current best solution $\mathbf{x}^*$. At each dimension j, the pair of components (x_j, x'_j) has a probability of 0.5 of being replaced by the corresponding components $(x^*_j, x^{*\prime}_j)$ of $\mathbf{x}^*$.

The breaking of a newly found best solution $\mathbf{x}^*$ generates at most N one-step neighboring solutions, each being obtained by trying to replace x'_j with a better alternative drug in Λ_{x_j} $(1 \le j \le N)$; if the best neighbor is better than $\mathbf{x}^*$, it replaces $\mathbf{x}^*$ in the population.

Algorithm 3 presents the pseudocode of the WWO algorithm for the prevention program optimization problem.

Algorithm 3 The WWO algorithm for the prevention program optimization problem.

```
Randomly initialize a population of solutions to the problem;
while the stopping criterion is not satisfied do
    Sort all solutions in increasing order of fitness;
    Let x* be the best solution in the population;
    foreach solution x in the population do
        Calculate λ(x) according to Eq. (20);
        //propagation
        Let x_λ = x;
        for k = 1 to λ(x) do
            Set x_λ to an immediate neighbor of x_λ;
        if fit(x_λ) > fit(x) then
            x ← x'_λ;
            if fit(x) > fit(x*) then
                //breaking
                x* ← x;
                for j = 1 to N do
                    if x'_j > 1 then
                        x'_j ← rand(1, x_j − 1);
                        if fit(x) > fit(x*) then
                            x* ← x;
        else
            if x has not been improved for ĝ generations then
                Refract x by learning from x*;
return the best known solution x*.
```

IV. Computational Results

During February and March, 2020, we applied the proposed method to TCM prevention of COVID-19 in two regions in Zhejiang Province, China:

- 39,720 residents in eight communities in Hangzhou city;
- 9,812 residents in four communities in Shaoxing city.

The following subsections report the results of resident clustering and prevention program optimization and implementation.

A. Results of Resident Clustering

Based on the analysis of TCM experts on local populations and TCM symptoms of COVID-19, the number C of clusters is set to 16. We compare the clustering results of K-means [22], DBSCAN [35], standard FCM, intuitionistic FCM (IFCM), Pythagorean FCM (PFCM), PFCM enhanced by EBO, and PFCM enhanced by five other popular metaheuristic algorithms including the genetic algorithm (GA) [36], differential evolution (DE) [37], comprehensive learning particle swarm optimization (CLPSO) [38], and hybrid biogeography-based optimization (HBBO) [39]. The experiments are conducted on a computer with Intel Xeon 3430 CPU and GTX 1080Ti GPU.

Each algorithm runs 30 times with different random seeds (K-means and four basic FCM methods use random initial cluster centroids, DBSCAN uses random order of data, and PFCM with different metaheuristics use randomly initial populations). As we do not know the ground truth class assignments, we use the Davies-Bouldin index [40], which is a function of the ratio of intra-cluster scatter to inter-cluster separation, as the performance metric (lower is better):

$$db = \frac{1}{c}\sum_{i=1}^{c} \max_{1 \le j \le c \wedge i \ne j} \left(\frac{s_i + s_j}{d_{ij}} \right) \tag{21}$$

where s_i is the average distance between each point and the centroid of cluster i, and d_{ij} is the distance between the centroids of clusters i and j.

Fig. 3(a) and (b) compare the performance of the ten clustering algorithms in Hangzhou and Shaoxing, respectively. On both instances, *K*-means exhibits the worst performance, because it aims to minimize the within-cluster sum-of-squares, but this criterion responds poorly to manifolds with irregular and uncertain shapes which often exist in population health data. FCM achieves significant performance advantage over *K*-means by incorporating fuzzy logic to effectively deal with uncertainties [24]. DBSCAN shows similar performance as FCM, and the standard deviation of DBSCAN over the 30 runs is smaller. IFCM achieves better results than the standard FCM, and PFCM achieves better results than IFCM, which shows that using extended fuzzy sets can improve the fuzzy clustering by capturing uncertainty information more effectively. Compared to the three basic FCM methods using random initial cluster centroids, PFCM enhanced by metaheuristic optimization to find optimal/sub-optimal initial centroids achieve significant performance advantages, because the quality of initial centroids heavily affects the clustering results. Among the five metaheuristic algorithms, the proposed PFCM-EBO exhibits the best performance, which demonstrates the efficiency of the EBO algorithm in optimizing initial centroids for clustering residents using health data.

B. Results of Prevention Program Optimization

The TCM experts develop 15 prevention programs for the 16 clusters of residents (two clusters are very similar and share one program). Among the 39,720 residents in Hangzhou, 4,625 residents agree to adopt the prevention programs, but medical resources required to implement the programs significantly exceed the available resource. Therefore, we use the proposed WWO algorithm to optimize the programs. Among the 15 updated programs, 13 programs are approved by the experts, and the remaining two programs are slightly modified by the experts. The updated programs do not violate the resource constraints and, therefore, are put into implementation. However, during one week of implementation, many resources have been consumed, and the available resources are not sufficient to implement the 15 updated programs. Therefore, we perform a second round of program optimization. Among the 15 programs updated in the second round, 12 programs are approved, and the remaining three programs are modified by the experts. However, the resources required by the programs again exceed the available resources. Therefore, we perform a third round of program optimization, the results of which are approved and put into implementation.

Among the 9,812 residents in Shaoxing, 1,227 agree to adopt the prevention programs. In Shaoxing, we perform two rounds of program optimization, the results of which are put into implementation successively.

In summary, we use the proposed algorithm to solve five instances of the prevention program optimization problem. For comparison, we also implement the following five popular metaheuristic optimization algorithms on these five instances:

- ❏ The GA for constrained optimization [41];
- ❏ The biogeography-based optimization (BBO) for constrained optimization [42];
- ❏ The DE algorithm for constrained optimization [43];
- ❏ The cuckoo search (CS) algorithm for integer programming [44];
- ❏ The grey wolf optimization (GWO) algorithm for integer programming [45].

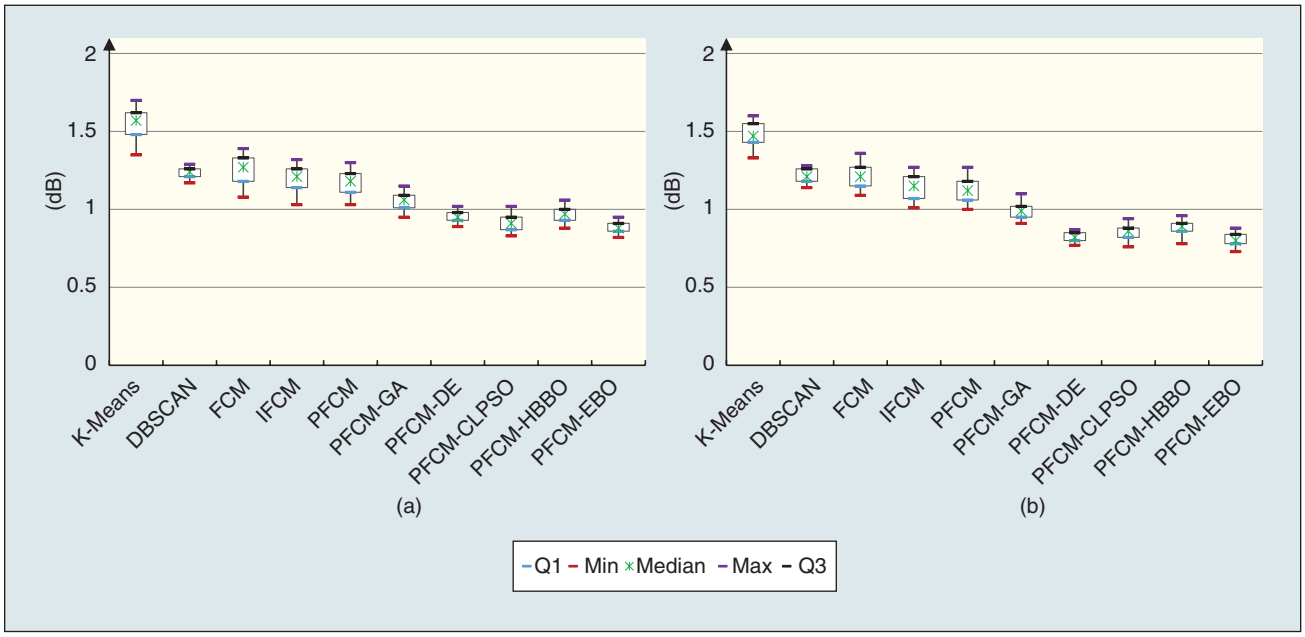


FIGURE 3 Comparison of the ten algorithms for clustering residents in (a) Hangzhou, and (b) Shaoxing. Each box plot shows the maximum, minimum, median, first quartile (Q1), and third quartile (Q3) of the resulting Davies-Bouldin index values over the 30 runs of an algorithm.

Each algorithm runs 30 times with different random seeds. Fig. 4 presents the resulting objective function values obtained by the six algorithms over the 30 runs. As the weights in the objective function (12) are normalized, the objective function value represents the average index of drugs selected from the alternative sets. On the three instances in Hangzhou, the median objective function values of the WWO algorithm are 1.83, 2.71, and 0.39, respectively,

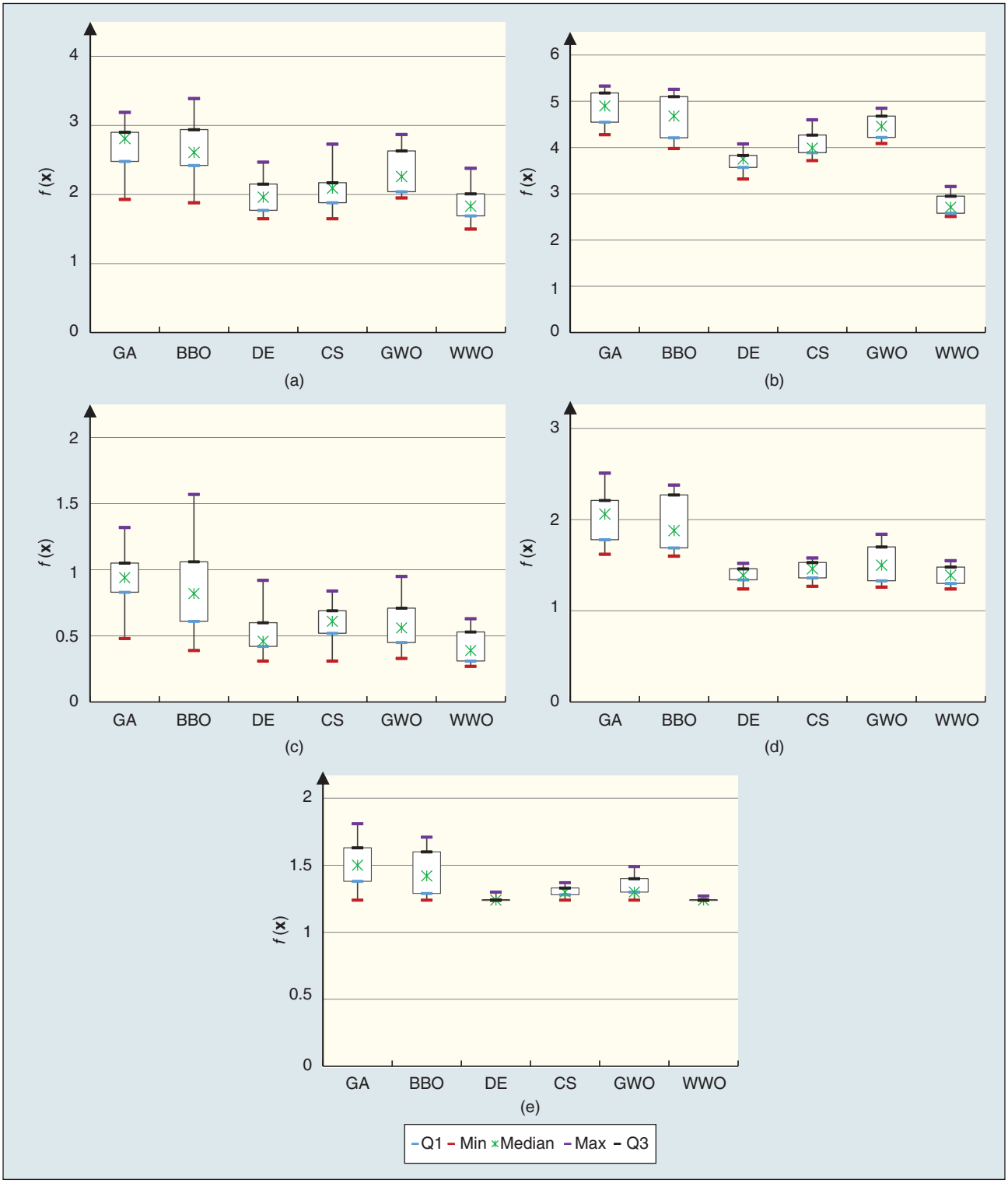


FIGURE 4 Comparison of the resulting objective function values obtained by the algorithms for prevention program optimization. Each box plot shows the maximum, minimum, median, first quartile (Q1), and third quartile (Q3) of objective function values over the 30 runs of an algorithm. (a) Hangzhou, 1st round. (b) Hangzhou, 2nd round. (c) Hangzhou, 3rd round. (d) Shaoxing, 1st round, and (e) Shaoxing, 2nd round.

TABLE 2 Age distribution, percentage with underlying diseases, and TCM constitution distribution of residents using diversified programs and residents using the unified program.

AGE		0–6	7–13	14–19	20–39	40–59	≥60	DISEASED		
HANGZHOU	DIVERSIFIED	10.88%	8.06%	7.83%	28.80%	33.82%	10.62%	5.27%		
	UNIFIED	11.03%	8.63%	8.01%	28.63%	32.60%	11.10%	6.30%		
SHAOXING	DIVERSIFIED	11.17%	7.50%	7.99%	26.89%	35.53%	10.92%	4.12%		
	UNIFIED	11.51%	7.64%	7.84%	25.79%	35.60%	11.60%	4.38%		
TCM CONSTITUTION		MILD	YANG DEFICIENCY	YIN DEFICIENCY	PHLEGM DAMPNESS	WET & HEAT	QI STAGNATION	QI DEFICIENCY	BLOOD STASIS	SPECIAL
HANGZHOU	DIVERSIFIED	35.26%	9.56%	5.17%	5.69%	10.59%	9.04%	11.35%	10.38%	6.88%
	UNIFIED	32.05%	9.88%	5.30%	5.98%	11.25%	8.86%	11.58%	10.66%	7.54%
SHAOXING	DIVERSIFIED	40.34%	9.21%	4.16%	5.13%	10.76%	8.07%	9.86%	10.19%	7.01%
	UNIFIED	39.53%	9.10%	4.48%	5.12%	11.28%	7.76%	9.34%	9.75%	7.28%

which are always the smallest among the six algorithms. These three instances have the same number N of programs and similar numbers of residents, but the constraints of the second-round instance is the most rigorous, while the constraints of the third-round instance is the least rigorous. The differences among the comparative algorithms are the largest on the second-round instance and the smallest on the third one. On the second-round instance, the performance advantages of WWO over the other algorithms are also the most significant. This demonstrates that the proposed WWO algorithm is efficient in solving complex instances of the problem.

On the two instances in Shaoxing, the median objective function values of WWO are 1.39 and 1.24, respectively, which are also the smallest among the six algorithms. In the second-round instance, the resources required by the basic prevention programs do not exceed the available resources too much; on this relatively simple instance, all algorithms achieve the same minimum objective function value of 1.24, which has been verified to be the exact optimal objective function value. However, only the median objective function values of DE and WWO are 1.24, and the maximum objective function value of WWO is less than that of DE. In summary, the results show that the proposed WWO algorithm exhibits the best performance on all instances.

C. Results of Prevention Program Implementation

We compare the actual prevention effects of the diversified TCM prevention programs developed using our method with those of the unified TCM prevention program released by Zhejiang Provincial Health Commission. We collect data of 36,138 residents from 71 communities in Hangzhou and 10,530 residents from 23 communities in Shaoxing who adopt the unified prevention program. Table II presents the distribution of basic characteristics in the residents using the different prevention policies. In either region, the two groups have similar age distribution, percentage with moderate and severe underlying diseases, and TCM constitution distribution (note that some particular people can have more than one TCM constitution). Therefore, the comparison of the effects of diversified and unified programs on the two different groups is justified, as a resident can take only one prevention program.

Table III presents the comparison results. In Hangzhou, during February and March, 2019, among 4,625 residents adopting diversified prevention programs, there is no reported case of COVID-19. During the same period, among 36,138 residents adopting the unified prevention program, there are six COVID-19 cases, including five imported cases and one local case. In Shaoxing, among 1,227 residents adopting our diversified prevention programs, there is also no reported case of COVID-19. During the same period, among 10,530 residents adopting the unified prevention program, there are two imported cases. According to the results in Table III, in terms of incidence rate, the effects of diversified TCM prevention programs are obviously better than those of the unified TCM prevention program in both regions.

Nevertheless, due to the low incidence rate of COVID-19 in China and the limited number of residents in this study, the comparison of incidence rates does not have sufficient statistical significance. Thus, we also conduct a questionnaire survey on the effects of prevention programs. There are two questions, the first is "TCM prevention program helps me improve health conditions," and the second is "TCM prevention program helps me

TABLE 3 Comparison of the effects of our diversified TCM prevention programs with those of the unified TCM prevention program.

		RESIDENTS	CASES	INCIDENCE RATE	LOCAL CASES	LOCAL INCIDENCE RATE
HANGZHOU	DIVERSIFIED	4,625	0	**0**	0	**0**
	UNIFIED	36,138	6	0.0166%	1	0.00277%
SHAOXING	DIVERSIFIED	1,227	0	**0**	0	**0**
	UNIFIED	10,530	2	0.0190%	0	0

prevent COVID-19." Each answer has seven choices: strongly agree, agree, weakly agree, neutral, weakly disagree, disagree, and strongly disagree.

There are a total of 7,358 residents, including 2,550 adopting diversified programs and 4,808 adopting the unified program, participating in the survey. Fig. 5 presents the survey results of the first question: among the participants adopting diversified programs, 73% give positive answers (18% strongly agree, 44% agree, and 11% weakly agree), which is significantly higher than the 59% among the participants adopting the unified program who give positive answers (16% strongly agree, 31% agree, and 12% weakly agree). Fig. 6 presents the results of the second question. Among the participants adopting diversified programs, 78% give positive answers (23% strongly agree, 41% agree, and 14% weakly agree), which is significantly higher than the 63% among the participants adopting the unified program who give positive answers (20% strongly agree, 29% agree, and 14% weakly agree). In summary, the survey results demonstrate that the residents adopting diversified programs are more satisfied with the effects of COVID-19 prevention and health condition improvement than those adopting the unified program.

V. Conclusion

In this study, we propose an intelligent optimization method to develop diversified TCM prevention programs for COVID-19. First, we use a fuzzy clustering method to divide the population based on combined modern medicine and TCM health characteristics. Based on the clustering results, TCM experts develop diversified prevention programs, which are then evolved by an interactive optimization method until all the resource constraints are satisfied. The method has been used for improving community prevention of COVID-19 in Zhejiang province, China, during the peak of the pandemic.

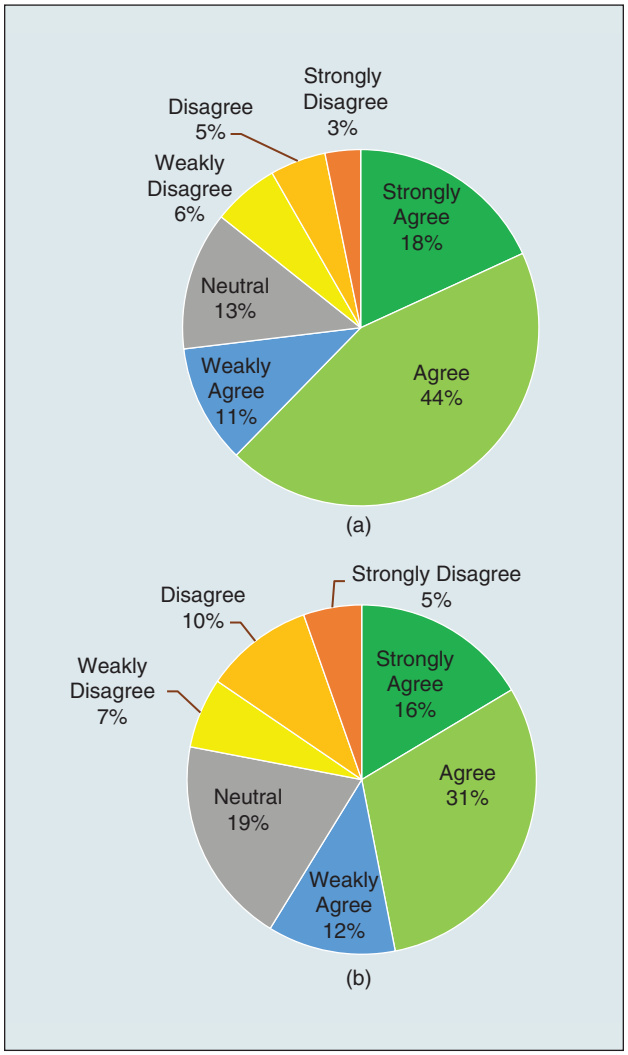


FIGURE 5 Distribution of different answers to the question "TCM prevention program helps to improve health conditions." (a) Participants adopting diversified prevention programs. (b) Participants adopting the unified prevention program.

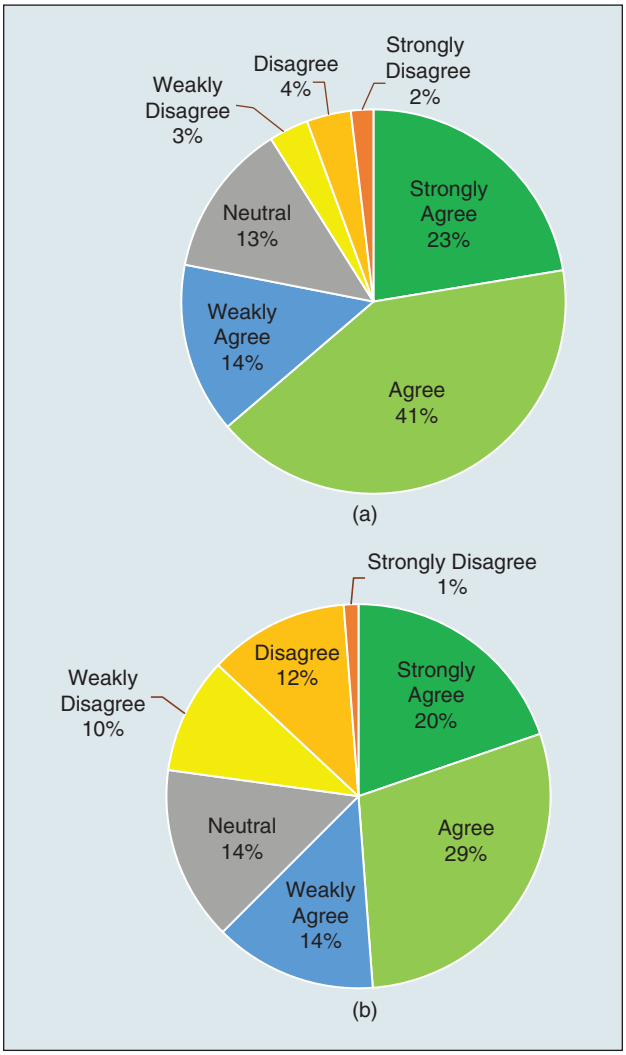


FIGURE 6 Distribution of different answers to the question "TCM prevention program helps to prevent COVID-19." (a) Participants adopting diversified prevention programs. (b) Participants adopting the unified prevention program.

The reported work is an emergency study aiming at COVID-19. We are continuously improving it in the following aspects: (1) using health big-data analytics to enhance the feature sets and incorporating multi-view learning to improve clustering; (2) incorporating more TCM knowledge to interactive optimization to reduce the efforts of TCM experts; (3) combining medical knowledge with machine learning to evaluate the effects of TCM prevention programs in a more accurate way. It is expected that the proposed method can be extended to the prevention and control of more epidemics in the future.

Acknowledgment

This work was supported by National Natural Science Foundation of China under Grant No. 61872123, Zhejiang Provincial Natural Science Foundation under Grant LR20F030002 and LQY20F030001, and Zhejiang Provincial Emergency Project for Prevention & Treatment of New Coronavirus Pneumonia under Grant 2020C03126.

References

[1] Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention," *JAMA*, vol. 323, no. 13, pp. 1239–1242, 2020. doi: 10.1001/jama.2020.2648.

[2] Y. Zhang, Q. Zhao, and B. Hu, "Community-based prevention and control of COVID-19: Experience from China," *Am. J. Infect. Control*, vol. 48, no. 6, pp. 716–717, 2020. doi: 10.1016/j.ajic.2020.03.012.

[3] M. McKee and D. Stuckler, "If the world fails to protect the economy, COVID-19 will damage health not just now but also in the future," *Nat. Med*, vol. 26, no. 5, pp. 640–642, 2020. doi: 10.1038/s41591-020-0863-y.

[4] Y.-J. Zheng, C.-X. Wu, E.-F. Chen, X.-Q. Lu, and M.-X. Zhang, "An optimization method for production resumption planning under COVID-19 epidemic," *Oper. Res. Trans.*, 2020, vol. 24, no. 3, pp. 43-56, 2020. doi: 10.15960/j.cnki.issn.1007-6093.2020.03.003.

[5] A. Loregian, B. Mercorelli, G. Nannetti, C. Compagnin, and G. Palù, "Antiviral strategies against influenza virus: Towards new therapeutic approaches," *Cell. Mol. Life Sci.*, vol. 72, pp. 3659–3683, 2014. doi: 10.1007/s00018-014-1615-2.

[6] F. Chen et al., "An urgent call for raising the scientific rigorousness of clinical trials on COVID-19," *Chin. J. Epidemiol.*, 2020, doi: 10.3760/cma.j.issn.0254-6450.2020.03.004.

[7] D. L. Smith, S. A. Levin, and R. Laxminarayan, "Strategic interactions in multi-institutional epidemics of antibiotic resistance," *Proc. Nat. Acad. Sci.*, vol. 102, no. 8, pp. 3153–3158, 2005. doi: 10.1073/pnas.0409523102.

[8] X. Wang, A. Zhang, H. Sun, and P. Wang, "Systems biology technologies enable personalized traditional Chinese medicine: A systematic review," *Am. J. Chin. Med.*, vol. 40, no. 6, pp. 1109–1122, 2012. doi: 10.1142/S0192415X12500826.

[9] A. Zhang, H. Sun, P. Wang, Y. Han, and X. Wang, "Future perspectives of personalized medicine in traditional Chinese medicine: A systems biology approach," *Compl. Ther. Med.*, vol. 20, no. 1, pp. 93–99, 2012. doi: 10.1016/j.ctim.2011.10.007.

[10] L. Wang, Y. Wang, D. Ye, and Q. Liu, "Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence," *Int. J. Antimicrob, Agents*, vol. 55, no. 6, p. 105,948, 2020. doi: 10.1016/j.ijantimicag.2020.105948.

[11] J. Liu, E. Manheimer, Y. Shi, and C. Gluud, "Chinese herbal medicine for severe acute respiratory syndrome: A systematic review and meta-analysis," *J. Alt. Compl. Med.*, vol. 10, no. 6, pp. 1041–1051, 2004. doi: 10.1089/acm.2004.10.1041.

[12] P. Leung, "The efficacy of Chinese medicine for SARS: A review of Chinese publications after the crisis," *Am. J. Chin. Med.*, vol. 35, no. 4, pp. 575–581, 2007. doi: 10.1142/S0192415X07005077.

[13] C. Ji, R. Zhang, J. Liu, and L. Wang, "Review of prevention and treatment on influenza A (H1N1) with traditional Chinese medicine," *China J. Chin. Mat. Med.*, vol. 35, no. 14, pp. 1900–1903, 2010. doi: 10.4268/cjcmm20101430.

[14] S. S. Chang, H. J. Huang, and C. Y. C. Chen, "Two birds with one stone? Possible dual-targeting H1N1 inhibitors from traditional Chinese medicine," *PLoS Comput. Biol.*, vol. 7, no. 12, p. e1002315, 2011. doi: 10.1371/journal.pcbi.1002315.

[15] D. Y. Lu, T. R. Lu, and H. Y. Wu, "Zika therapy by traditional Chinese medicine, a new proposal," *Adv. Pharmacol. Clin. Trial*, vol. 1, no. 1, p. 103, 2016. doi: 10.23880/APCT-16000103.

[16] H. Luo et al., "Can Chinese medicine be used for prevention of corona virus disease 2019 (COVID-19)? a review of historical classics, research evidence and current prevention programs," *Chin. J Integr. Med.*, vol. 26, no. 4, pp. 243–250, 2020. doi: 10.1007/s11655-020-3192-6.

[17] J.-L. Ren, A.-H. Zhang, and X.-J. Wang, "Traditional Chinese medicine for COVID-19 treatment," *Pharmacol. Res.*, vol. 155, p. 104,743, May 2020. doi: 10.1016/j.phrs.2020.104743.

[18] J. Xu and Y. Zhang, "Traditional Chinese medicine treatment of COVID-19," *Compl. Therap. Clin. Pract.*, vol. 39, p. 101165, May 2020. doi: 10.1016/j.ctcp.2020.101165.

[19] K. W. Chan, V. T. Wong, and S. C. W. Tang, "COVID-19: An update on the epidemiological, clinical, preventive and therapeutic evidence and guidelines of integrative Chinese-Western Medicine for the management of 2019 Novel Coronavirus Disease," *Ame. J. Chin. Med.*, vol. 48, no. 3, pp. 737–762, 2020. doi: 10.1142/S0192415X20500378.

[20] L. Ang, H. W. Lee, J. Y. Choi, J. Zhang, and M. S. Lee, "Herbal medicine and pattern identification for treating COVID-19: A rapid review of guidelines," *Integr. Med. Res.*, vol. 9, no. 2, p. 100,407, 2020. doi: 10.1016/j.imr.2020.100407.

[21] L. Wanghua and L. Hua, *Memorization of 50 types of syndromes in Traditional Chinese medicine*. Taiyuan, China: Shanxi Science and Technology Press, 2011.

[22] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 7, pp. 881–892, 2002. doi: 10.1109/TPAMI.2002.1017616.

[23] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[24] Z. Xu and J. Wu, "Intuitionistic fuzzy c-means clustering algorithms," *J. Syst. Eng. Electron.*, vol. 20, no. 4, pp. 580–590, 2010. doi: 10.3969/j.issn.1004-4132.2010.04.009.

[25] K. T. Atanassov, "New operations defined over the intuitionistic fuzzy sets," *Fuzzy Sets Syst.*, vol. 61, no. 2, pp. 137–142, 1994. doi: 10.1016/0165-0114(94)90229-1.

[26] R. Yager, "Pythagorean fuzzy subsets," in *Proc. Joint IFSA World Congress and NAFIPS Annual Meeting*, Edmonton, Canada, 2013, pp. 57–61.

[27] P. Ren, Z. Xu, and X. Gou, "Pythagorean fuzzy TODIM approach to multi-criteria decision making," *Appl. Soft Comput.*, vol. 42, pp. 246–259, May 2016. doi: 10.1016/j.asoc.2015.12.020.

[28] M. Gong, Y. Liang, J. Shi, W. Ma, and J. Ma, "Fuzzy c-means clustering with local information and kernel metric for image segmentation," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 573–584, 2013. doi: 10.1109/TIP.2012.2219547.

[29] M. Zhang, W. Jiang, X. Zhou, Y. Xue, and S. Chen, "A hybrid biogeography-based optimization and fuzzy c-means algorithm for image segmentation," *Soft Comput.*, vol. 23, no. 6, pp. 2033–2046, 2019. doi: 10.1007/s00500-017-2916-9.

[30] Y.-J. Zheng, H.-F. Ling, and J.-Y. Xue, "Ecogeography-based optimization: Enhancing biogeography-based optimization with ecogeographic barriers and differentiations," *Comput. Oper. Res.*, vol. 50, pp. 115–127, Oct. 2014. doi: 10.1016/j.cor.2014.04.013.

[31] D. Simon, "Biogeography-based optimization," *IEEE Trans. Evol. Comput.*, vol. 12, no. 6, pp. 702–713, 2008. doi: 10.1109/TEVC.2008.919004.

[32] Y.-J. Zheng, H.-F. Ling, X.-B. Wu, and J.-Y. Xue, "Localized biogeography-based optimization," *Soft Comput.*, vol. 18, no. 11, pp. 2323–2334, 2014. doi: 10.1007/s00500-013-1209-1.

[33] Y.-J. Zheng, "Water wave optimization: A new nature-inspired metaheuristic," *Comput. Oper. Res.*, vol. 55, no. 1, pp. 1–11, 2015. doi: 10.1016/j.cor.2014.10.008.

[34] Y.-J. Zheng, X.-Q. Lu, Y.-C. Du, Y. Xue, and W.-G. Sheng, "Water wave optimization for combinatorial optimization: Design strategies and applications," *Appl. Soft Comput.*, vol. 83, p. 105611, 2019. doi: 10.1016/j.asoc.2019.105611.

[35] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Proc. Int. Conf. Knowledge Discovery and Data Mining*, Portland, 1996, pp. 226–231.

[36] H. Mühlenbein and D. Schlierkamp-Voosen, "Predictive models for the breeder genetic algorithm I. continuous parameter optimization," *Evol. Comput.*, vol. 1, no. 1, pp. 25–49, Mar. 1993. doi: 10.1162/evco.1993.1.1.25.

[37] R. Storn and K. Price, "Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optim.*, vol. 11, no. 4, pp. 341–359, 1997.

[38] J. J. Liang, A. K. Qin, P. Suganthan, and S. Baskar, "Comprehensive learning particle swarm optimizer for global optimization of multimodal functions," *IEEE Trans. Evol. Comput.*, vol. 10, no. 3, pp. 281–295, 2006. doi: 10.1109/TEVC.2005.857610.

[39] H. Ma, D. Simon, M. Fei, X. Shu, and Z. Chen, "Hybrid biogeography-based evolutionary algorithms," *Eng. Appl. Artif. Intell.*, vol. 30, no. 1, pp. 213–224, 2014. doi: 10.1016/j.engappai.2014.01.011.

[40] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-1, no. 2, pp. 224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

[41] S. Koziel and Z. Michalewicz, "Evolutionary algorithms, homomorphous mappings, and constrained parameter optimization," *Evol. Comput.*, vol. 7, no. 1, pp. 19–44, 1999. doi: 10.1162/evco.1999.7.1.19.

[42] H. Ma and D. Simon, "Blended biogeography-based optimization for constrained optimization," *Eng. Appl. Artif. Intell.*, vol. 24, no. 3, pp. 517–525, 2011. doi: 10.1016/j.engappai.2010.08.005.

[43] F. Luchi and R. A. Krohlingb, "Differential evolution and nelder-mead for constrained non-linear integer optimization problems," *Proc Comput. Sci.*, vol. 55, pp. 668–677, 2015. doi: 10.1016/j.procs.2015.07.071.

[44] M. Abdel-Baset, Y. Zhou, and M. Ismail, "An improved cuckoo search algorithm for integer programming problems," *Int. J. Comput. Sci. Math.*, vol. 9, no. 1, pp. 66–81, 2018. doi: 10.1504/IJCSM.2018.090710.

[45] H. Xing et al., "An integer encoding grey wolf optimizer for virtual network function placement," *Appl. Softw. Comput.*, vol. 76, pp. 575–594, Mar. 2019. doi: 10.1016/j.asoc.2018.12.037.

FUZZ-IEEE 2021 LUXEMBOURG

HANDLING UNCERTAINTY IN INTERPRETABLE AI

July 11 – July 14, 2021

The 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2021), the world-leading event focusing on the theory and application of fuzzy set and systems, will be held in Luxembourg City, Luxembourg, Europe.

Located at the heart of Europe, Luxembourg provides a stimulating, picturesque and historic backdrop to the conference. With a vast set of cultural and leisure activities at your fingertips, the city and country also offer ample opportunities to recharge your batteries with friends, family and colleagues outside the conference. Linked to the rest of Europe by a first-class road network and excellent rail connections (e.g. TGV train from Paris in under 2.5 hours), Luxembourg is easily accessible from anywhere in Europe. Internationally, frequent air connections are available from Luxembourg airport via major hubs like London, Frankfurt, Munich, Zurich, Amsterdam or Paris, making Luxembourg reachable within one hour flight from across Europe and via a single connection from most countries worldwide.

FUZZ-IEEE 2021 will be hosted at the Alvisse Parc Hotel (http://www.parc-hotel.lu), set in quiet, green surroundings within minutes of Luxembourg City. Public transport is free in Luxembourg, offering convenient mobility. With ongoing uncertainty around COVID-19, the conference will keep a close eye on international developments and shape the program to maximise engagement and deliver the best experience possible – whether in-person, virtual, or both.

FUZZ-IEEE 2021 will represent a unique meeting point for scientists and engineers, from academia and industry, to interact and discuss the latest enhancements and innovations in the field. The topics of the conference will cover all aspects of theory and applications of fuzzy sets and systems and its hybridisations with other artificial and computational intelligence techniques. Under its 2021 theme, 'Handling Uncertainty in Interpretable AI', the conference will emphasise the handling of uncertainty, in particular in the context of interpretable and interactive AI, actively promoting engagement across disciplines. FUZZ-IEEE 2021 topics include, but are not limited to:

•Mathematical and theoretical foundations of fuzzy sets, fuzzy measures, and fuzzy integrals, •Human-centric aspect of handling uncertainty in data and decision making, •Interpretable and Interactive approaches to uncertainty in AI, •Fuzzy control, robotics, sensors, fuzzy hardware and architectures, •Fuzzy data analysis, fuzzy clustering, classification and pattern recognition, •Type-2 fuzzy sets, computing with words, and granular computing, •Fuzzy systems with big data and cloud computing, fuzzy analytics, and visualization, •Fuzzy systems design and optimization, •Fuzzy decision analysis, multi-criteria decision making and decision support, •Applications of fuzzy sets and systems, fuzzy measures and integrals, •Fuzzy and uncertain information processing, information extraction, and fusion, •Fuzzy web engineering, information retrieval, text mining, and social network analysis, •Fuzzy image, speech and signal processing, vision, and multimedia data analysis, •Fuzzy databases and informational retrieval, •Theory and applications of imprecise probabilities and possibilities, •Theory and applications at the interface of fuzzy and probabilistic approaches, •Interdisciplinary work on fuzzy sets and soft computing in social sciences, •Handling of uncertainty and imprecision in applications from security to finance, •Hardware and software for fuzzy systems and fusion, •Hybrid, e.g., neuro- and evolutionary-fuzzy systems

The conference will include oral presentations, workshops, tutorials, panels, special sessions, and keynote presentations. Full details will be available on the conference website http://attend.ieee.org/fuzzieee-2021.org - we look forward to welcoming you with a warm 'Moien!' or hello, as we say in Luxembourgish! :-)

Conference Committee:

General Co-Chairs: Christian Wagner (UK), Holger Voos (LU)
Program Co-Chairs: Hani Hagras (UK), Susana Vieira (PT)
Special Session Co-Chairs: Sansanee Auephanwiriyakul (TH), Francisco Herrera (ES) Workshop and Tutorial Co-Chairs: Gabriella Pasi (IT), Derek Anderson (US) Keynotes Co-Chairs: Jonathan Garibaldi (UK), Rosangela Ballini (BR)
Conflict of Interest Chair: Humberto Bustince (SP)
Competitions Chair: Anna Wilbik (NL) Panel Sessions Co-Chairs: Tim Havens (US), Mika Sato-Ilic (JP) Publications Cho-Chairs: Tufan Kumbasar (TR), Shaily Kabir (BD)
Finance Co-Chairs: Sabine Bösl (LU), Pablo Estevez (CL)
Student and Early Career Engagement Co-Chairs: Marie-Jeanne Lesot (FR), Keeley Crockett (UK) Local Arrangement Chairs: Magali Martin (LU), Yu-Youn Song (LU) Web and Social Media Chair: Direnc Pekaslan (UK)
(for full committee details, please review the website)

Important dates and deadlines:

Special Session, tutorial, and panel proposals: Oct. 18, 2020
Notification of special session, tutorial, and panel approval: Nov. 16, 2020
Paper submission deadline: Feb. 10, 2021
Notification for acceptance of papers: Mar. 22, 2021
Camera-ready paper deadline: Apr. 12, 2021
Early registration deadline: May 3, 2021
Conference: Jul. 11 – 14, 2021

IEEE Computational Intelligence Society

Digital Object Identifier 10.1109/MCI.2020.3019904

Marley Vellasco
Pontifícia Universidade Católica do Rio de Janeiro, BRAZIL

Conference Calendar

* Denotes a CIS-Sponsored Conference
Δ Denotes a CIS Technical Co-Sponsored Conference

Δ 7th International Conference on Behavioral and Social Computing (BESC)
November 5–7, 2020
Place: Bournemouth, UK-virtual
General Chair: John McAlaney
Website: http://besc-conf.org/2020/index.html

Δ 7th International Conference on Soft Computing and Machine Intelligence (ISCMI 2020)
November 14–15, 2020
Place: Stockholm, Sweden-virtual
General Chair: Suash Deb
Website: http://www.iscmi.us

*** 2020 IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2020)**
December 1–4, 2020
Place: Canberra, Australia-virtual
General Chair: Hussein Abbass
Website: http://www.ieeessci2020.org/

*** 2020 IEEE Smart World Conference (IEEE SWC 2020)**
December 8–10, 2020
Place: Melbourne, Australia-virtual
General Co-Chairs: Stephen Yau, Mazin Yousif, and Jinjun Chen
Website: http://www.swinflow.org/confs/2020/swc/index.htm

*** 2021 IEEE Congress on Evolutionary Computation (IEEE CEC 2021)**
June 28–July 1, 2021
Place: Kraków, Poland
General Co-Chairs: Jacek Mańdziuk and Hussein Abbass
Website: https://cec2021.mini.pw.edu.pl

*** 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2021)**
July 11–14, 2021
Place: Luxembourg, Luxembourg
General Co-Chairs: Christian Wagner and Holger Voos
Website: https://attend.ieee.org/fuzzieee-2021/

*** 2021 IEEE International Conference on Development and Learning (ICDL)**
August 23–26, 2021
Place: Beijing, China
General Co-Chairs: Dingsheng Luo and Angelo Cangelosi
Website: TBA

*** 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)**
October 13–15, 2021
Place: Melbourne, Australia
General Chair: Madhu Chetty
Website: TBA

*** 2021 IEEE Smart World Conference (IEEE SWC 2021)**
October 18–21, 2021
Place: Atlanta, USA
General Co-Chairs: Yi Pan, Rajshekhar Sunderraman, and Yanqing Zhang
Website: https://grid.cs.gsu.edu/~smartworld2021/index.php

*** 2021 IEEE Latin American Conference on Computational Intelligence (LA-CCI)**
November 2–4, 2021
Place: Temuco, Chile
General Co-Chairs: Millaray Curilem and Doris Saez
Website: http://la-cci.org/

*** 2021 IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2021)**
December 5–8, 2021
Place: Orlando, FL, USA
General Co-Chairs: Sanaz Mostaghim and Keeley Crockett
Website: TBA

*** 2022 IEEE World Congress on Computational Intelligence (IEEE WCCI 2022)**
July 18–23, 2022
Place: Padua, Italy
General Co-Chairs: Marco Gori and Alessandro Sperduti
Website: https://wcci2022.org

Digital Object Identifier 10.1109/MCI.2020.3019900
Date of current version: 14 October 2020

IEEE Collabratec™

Bright Minds. Bright Ideas.

Introducing IEEE Collabratec™

The premier networking and collaboration site for technology professionals around the world.

IEEE Collabratec is a new, integrated online community where IEEE members, researchers, authors, and technology professionals with similar fields of interest can **network** and **collaborate**, as well as **create** and manage content.

Featuring a suite of powerful online networking and collaboration tools, IEEE Collabratec allows you to connect according to geographic location, technical interests, or career pursuits.

You can also create and share a professional identity that showcases key accomplishments and participate in groups focused around mutual interests, actively learning from and contributing to knowledgeable communities. All in one place!

Network.
Collaborate.
Create.

Learn about IEEE Collabratec at
ieee-collabratec.ieee.org

IEEE

`

IEEE Transactions on Artificial Intelligence

Call for Papers

Scope

The IEEE Transactions on Artificial Intelligence (TAI) is a multidisciplinary journal publishing papers on theories and methodologies of Artificial Intelligence. Applications of Artificial Intelligence are also considered.

Topics covered by IEEE TAI include, but not limited to, Agent-based Systems, Augmented Intelligence, Autonomic Computing, Constraint Systems, Explainable AI, Knowledge-Based Systems, Learning Theories, Planning, Reasoning, Search, Natural Language Processing, and Applications. Technical papers addressing contemporary topics in AI such as Ethics and Social Implications are welcomed.

Invitation

The journal invites impactful Artificial Intelligence research, survey articles, and applications. Submit your manuscript at the IEEE TAI Manuscript Central website at https://mc.manuscriptcentral.com/tai-ieee. Potential authors should consult the Information to Authors Document at https://cis.ieee.org/publications/ieee-transactions-on-artificial-intelligence/information-for-authors-tai. Further questions can be directed to the Founding Editor-in-Chief at ieee.tai.eic@gmail.com

Founding Editor-in-Chief

Hussein Abbass, University of New South Wales, Canberra, Australia

Associate Editors

- Amal El Fallah Seghrouchni, Sorbonne University, France
- Catherine Huang, McAfee, USA
- Christian Wagner, University of Nottingham, UK
- Dongbin Zhao, University of Chinese Academy of Sciences, China
- Fiora Pirri, University of Rome, Italy
- Gary Yen, Oklahoma State University, USA
- Guilherme DeSouza, University of Missouri, USA
- Haibo He, University of Rhode Island, USA
- Hao Luo, Harbin Institute of Technology, China
- Johan Suykens, Katholieke Universiteit Leuven, Belgium
- Kay Chen Tan, City University of Hong Kong, Hong Kong
- Lirong Xia, Rensselaer Polytechnic Institute , USA
- Matthew Garratt, University of New South Wales, Australia
- Michael Wooldridge, University of Oxford, UK
- Pau-Choo Chung, National Cheng Kung University, Taiwan
- Peter Stuckey, Monash University, Australia
- Ran Cheng, Southern University of Science and Technology, China
- Sanaz Mostaghim, Otto von Guericke University Magdeburg, Germany
- Simon Yang, University of Guelph, Canada
- Supratik Mukhopadhyay, Louisiana State University, USA
- Weizhong Yan, General Electric Global Research Center, USA
- Yo-Ping Huang, National Taipei University of Technology, Taiwan
- Pablo Estevez , University of Chile, Chile
- Pascal Van Hentenryck, Georgia Tech, USA

Digital Object Identifier 10.1109/MCI.2020.3019902


Main Market Square & Cloth Hall - 13th c.


Collegium Maius Courtyard - 13th c.


Wawel Royal Castle - 13th c.


Wieliczka Salt Mine - underground lake

IEEE CEC 2021 is a world-class conference that brings together researchers and practitioners in the fields of evolutionary computation and computational intelligence from all around the globe. The conference covers all topics in evolutionary computation and encompasses plenary lectures, regular and special sessions, tutorials, competitions and poster presentations.

IEEE CEC 2021 will be held in Kraków, the second largest and one of the oldest cities in Poland. Kraków is regarded as one of Europe's most beautiful cities and its historic centre is listed on the UNESCO World Heritage List. Old and modern architecture with vibrant centers of social life are the town's strongest assets. Kraków is also a mecca for innovative business and is often referred to as the Polish Silicon Valley.

Call for Papers

Papers for IEEE CEC 2021 should be submitted electronically through the Congress website at
cec2021.mini.pw.edu.pl
and will be refereed by experts in the field and ranked based on the criteria of originality, significance, quality and clarity.

Call for Special Sessions

Special session proposals are invited to IEEE CEC 2021. A special session proposal should include the title, aims and scope of the proposed session (including a list of the main topics), and the names and short biographies of the organizers. A list of potential contributors will be very helpful. All proposals should be submitted to the Special Session Chairs:
Sung-Bae Cho (sbcho@yonsei.ac.kr) and
Jarosław Arabas (jaroslaw.arabas@pw.edu.pl)

Call for Workshops

Workshops will be held to provide participants with the opportunity to present and discuss novel research ideas on active and emerging topics in Evolutionary Computation. Prospective workshop organizers are invited to submit their proposals to the Workshop Chair:
Maria Ganzha (m.ganzha@mini.pw.edu.pl)

Call for Competitions

Competitions will be held as part of the Congress. Prospective competition organizers are invited to submit their proposals to the Competition Chairs:
Simon Lucas (simon.lucas@qmul.ac.uk) and
Julian Togelius (julian@togelius.com)

Call for Tutorials

IEEE CEC 2021 solicits proposal for tutorials covering specific topics in Evolutionary Computation. If you are interested in proposing a tutorial, would like to recommend someone who might be interested, or have questions about tutorials, please contact the Tutorial Chair:
Mengjie Zhang (mengjie.zhang@ecs.vuw.ac.nz)

Important Dates

19 November 2020	17 December 2020	31 January 2021	22 March 2021	7 April 2021
Special Session Proposal & Workshop Proposal Deadline	Competition Proposal & Tutorial Proposal Deadline	Paper Submission Deadline	Notification to Authors	Final Paper Submission & Early Registration Deadline

Sponsored by

IEEE Computational Intelligence Society
Jagiellonian University in Kraków
Politechnika Warszawska Warsaw University of Technology
AGH University of Science and Technology

Digital Object Identifier 10.1109/MCI.2020.3019903

Find Us at

www.cec2021.mini.pw.edu.pl
www.facebook.com/cec2021
www.twitter.com/CEC2021
www.linkedin.com/company/cec2021
cec2021@mini.pw.edu.pl